Multimodal Unified Bidirectional Cross-Modal Audio-Visual Saliency Prediction

By

Tadele Melesse

Submitted to the School of Information Technology and Engineering

In partial fulfillment of the requirements for the degree of

Master of Science in Artificial Intelligence

Supervised by:

Main Advisor: Dr. Natnael Argaw

Co-Advisor: Dr. Beakal Gizachew

College of Technology and Built Environment

Addis Ababa University

Addis Ababa, Ethiopia

June , 2025

APPROVAL

This is to certify that this thesis titled **"Multimodal Unified Bidirectional Cross-Modal audio-Visual Saliency Prediction"** is prepared by Tadele Melesse Sishagn and submitted in partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Artificial Intelligence at School of Information Technology & Engineering, College of Technology and Built Environment.

Name	**Signature**	**Date**
Dr. Natnael Argaw (Advisor)	_____	_____
Dr. Beakal Gizachew (Co-Advisor)	_____	_____
Dr. Adane Leta (Internal Examiner)	_____	_____
Dr. Worku Jifara (External Examiner)	_____	_____
Dr. Fantahun Bogale (Chairman)	_____	_____

ABSTRACT

Human attention in dynamic environments is inherently multimodal and is shaped by the interplay of auditory and visual cues. Although existing saliency prediction methods predominantly focus on visual semantics, they neglect audio as a critical modulator of gaze behavior. Recent audiovisual approaches attempt to address this gap but remain limited by temporal misalignment between modalities and inadequate retention of spatio-temporal information, which is key to resolving both the location and timing of salient events, ultimately yielding suboptimal performance. Inspired by recent breakthroughs in cross-attention transformers with convolutions for joint global-local representation learning and conditional denoising diffusion models for progressive refinement, we introduce a novel multimodal framework for bidirectional efficient audiovisual saliency prediction. It employs dual-stream encoders to process video and audio independently, coupled with separate efficient cross-modal attention pathways that model mutual modality influence: One pathway aligns visual features with audio features, while the other adjusts audio embeddings to visual semantics. Critically, these pathways converge into a unified latent space, ensuring coherent alignment of transient audiovisual events through iterative feature fusion. To preserve fine-grained details, residual connections propagate multiscale features across stages. For saliency generation, a conditional diffusion decoder iteratively denoises a noise-corrupted ground truth map, conditioned at each timestep on the fused audiovisual features through a hierarchical decoder that enforces spatio-temporal coherence via multiscale refinement. Extensive experiments demonstrate that our model outperforms state of the art methods, achieving individual improvements of up to 11.52% (CC), 20.04% (SIM), and 3.79% (NSS) across evaluation metrics over DiffSal on the AVAD dataset.

Acknowledgements

First and foremost, I would like to thank MERCYFULL GOD! I am deeply grateful to my advisors, Dr. Natnael Argaw and Dr. Beakal Gizachew, for their expert guidance, continuous support, and invaluable feedback throughout this work. Their advice and encouragement played a key role in shaping both my research and personal growth. I also sincerely thank Dr. Adane Leta and Dr. Worku Jifara for their thoughtful comments, which helped improve and refine this study.

A heartfelt thank you also goes out to my friends and classmates for their great collaboration. I am especially appreciation to Azmeraw Bekele for his constant advice and support throughout every stage of my work. Special thanks also to Haileleul, Habtamu, and Sintayehu for their steadfast encouragement, insightful perspectives, and generous collaboration. Their contributions, along with the support of my other classmates, fostered an inspiring and motivating environment that greatly enriched this journey.

This work is dedicated to my beloved mother, who lived her entire life for me. *Etateye* I miss your smile every day. I also wish to express my deepest Thanks to my family: specially to my brother Temesgen, my wife Sisay, and my daughter Betselot, for their endless love, support, and encouragement. Their strength and sacrifice have been the foundation of my perseverance throughout this journey.

Contents

Abbreviations

Symbol	Meaning
3D CNN	3D Convolutional Neural Network
3DResNet-18	3D Residual Network with 18 layers
ACM	Association for Computing Machinery
ACRNet	Audio-Conditioned Recurrent Network
AEBA	Audio Encoder Backbone Architecture
APTM	Audio Pretraining Model
ASP	Audio Saliency Prediction
ASD	Audio Saliency Detection
Att.	Attention Mechanism
AVAD	Audio-Visual Attention Dataset
AVIM	Audio Visual Interaction Module
AVSP	Audio-Visual Saliency Prediction
AViNet	Audio-Visual variant of ViNet
BE	Bidirectional Encoder
BECA	Bidirectional Efficient Cross Attention
BCE	Binary Cross Entropy
CC	Correlation Coefficient
CDM	Conditional Denoising Module
CNN	Convolutional Neural Network
CONV	Convolutional (layer)
Cov	Covariance
CPC	Consistency Aware Predictive Coding
DAVE	Deep Audio-Visual Embedding
DAVE2	Deep Audio-Visual Embedding 2
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
Diff-VSP	Diffusion-based Visual Saliency Prediction
DConvLSTM	Deformable Convolutional Long Short-Term Memory
DSAM	Deeply Supervised Attention Module
DWT	Discrete Wavelet Transform
ECCV	European Conference on Computer Vision
FC	Fully Connected (layer)
FFT	Fast Fourier Transform
FLOPs	Floating Point Operations

FPN	Feature Pyramid Network
GRU	Gated Recurrent Unit
GPU	Graphics Processing Unit
HCI	Human-Computer Interaction
HRI	Human-Robot Interaction
IEEE	Institute of Electrical and Electronics Engineers
IoU	Intersection over Union
JCR	Journal Citation Reports
KL-divergence / KL	Kullback–Leibler Divergence
LSTM	Long Short-Term Memory
LR	Learning Rate
MAE	Mean Absolute Error
MDPI	Multidisciplinary Digital Publishing Institute
MSE	Mean Squared Error
MLP	Multi-Layer Perceptron
MViT	Multiscale Vision Transformer
MViTv2	Multiscale Vision Transformer version 2
NSS	Normalized Scanpath Saliency
PCA	Principal Component Analysis
PLoS	Public Library of Science
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Q, K, V	Query, Key, and Value (attention mechanism components)
RQ	Research Question
ReLU	Rectified Linear Unit
RGB	Red Green Blue (color channels)
RNN	Recurrent Neural Network
S3D	Separable 3D Convolution
SIM	Similarity Metric
SOTA	State of the Art
SPred	Predicted saliency map
Sgt	Ground-truth saliency map
S0	Initial saliency map or clean data in diffusion model
STFT	Short-Time Fourier Transform
STSANet	Spatio-Temporal Self-Attention Network
STSP	Spatio-Temporal Saliency Prediction
STT	Spatio-Temporal Transformer
STEA	Spatio-Temporal Encoder Architecture
STC	Spatial-Temporal Compression
TPU	Tensor Processing Unit
U-Net	U-shaped Network (Encoder-Decoder Architecture)
VGG	Visual Geometry Group (Network Architecture)

VGGish	VGG-based Audio feature extractor (pretrained network)
ViNet	Name of a specific Visual saliency prediction Network
VPTM	Visual Pretraining Model
VFP	Video Fixation Prediction
VOC	Visual Object Classes (dataset)
VSOD	Video Salient Object Detection

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Background

Human perception is constantly bombarded with a stream of multimodal sensory input images, motion, sounds competing for limited attention[1]. The concept of saliency prediction refers to the computational modeling of this selective attention process: identifying which regions in visual or audiovisual stimuli are most likely to draw human focus[2]. Inspired by cognitive neuroscience, this task mirrors the way our brains filter sensory stimuli highlighting only the most relevant information for focused perception and decision making[3]. In various systems, saliency prediction plays a key role in human robot interaction, autonomous driving, video surveillance, and augmented reality[4, 5, 6].

Early saliency prediction models focused primarily on visual content either static [7, 8, 9] or frame based videos[10, 11, 12, 13, 14]processed using usual computer vision techniques or deep learning based networks. These unimodal models, while effective in predicting gaze locations or attention heatmaps, overlook the inherent multimodal nature of real-world perception[15]. In practice, human attention is not guided by visuals alone; auditory cues, contextual semantics, and temporal dynamics all play crucial roles[16].

This recognition has sparked interest in multimodal saliency prediction, particularly integrating auditory and visual information[17]. Audio-Visual Saliency Prediction (AVSP) models aim to simulate this human like attention mechanism by jointly analyzing visual scenes and accompanying sound cues.Unlike static image saliency, AVSP must consider how sound events align or interact with visual

motion over time[18]. For instance, a sudden loud noise can direct attention even when visual cues are subtle. However, early approaches to AVSP relied heavily on handcrafted fusion strategies, late fusion pipelines, or simple concatenation of features, failing to capture the nuanced interactions between modalities[19, 20].

Deep learning has driven significant advancements in numerous computer vision problems, including object detection[21], action recognition[22], and semantic segmentation[23] applied for various applications. It has aslo played a transformative role in advancing AVSP, enabling end-to-end learning of spatio-temporal and cross-modal dependencies. Architectures based on 3D Convolutional Neural Networks (3D-CNNs) like Dave[24] , Recurrent Neural Networks (RNNs) like Avinet[25], and more recently, Transformers like Stavis and Casp-Net[26, 27] and Diffusion models like Diffsal[28], have been proposed to better capture the joint evolution of audio and visual cues over time.

The availability of complex datasets such as AVAD[29], which pairs human eye fixation data with real-world audiovisual content, has provided a foundation for benchmarking AVSP systems. Meanwhile, the demand for robust, generalizable, and temporally efficient AVSP models continues to drive research in this field.

1.2 Motivation

A. Human Multimodal Perception

Humans naturally integrate information from multiple sensory modalities, such as vision and hearing. Cognitive neuroscience has shown that we selectively attend to salient audiovisual stimuli through tightly coupled cross-modal mechanisms, enabling robust decision-making in complex and dynamic environments [18]. Motivated by this, our work aims to emulate this human-like perceptual integration using efficient *bidirectional cross-attention mechanisms*. Unlike traditional approaches that rely solely on visual saliency, we enable dynamic intermodal inter-

action to better reflect real-world attention behavior.

B. Challenges in Temporal Mismatch

While prior works primarily focus on fusing visual and auditory features, they often fail to address the *temporal misalignment* between modalities—leading to degraded saliency prediction. Methods like DiffSal [28] and CASP-Net [27] incorporate cross-modal fusion but do not adequately handle asynchronous events. Our initiative model tackles this with a Bidirectional Efficient Cross Attention (BECA) mechanism combined with a temporal-guided noise encoder, enabling mutual refinement and temporal alignment across modalities. This results in more temporally coherent saliency localization, even in the presence of audio-visual lag or weak correlations.

C. Cross-Modal Advancement through Adaptive Fusion

Recent advances in cross-modal learning emphasize the importance of adaptive attention mechanisms. Instead of relying on fixed or naive fusion, incorporating a Gated Fusion Module that dynamically learns the importance of each modality based on context is very important[30]. This ensures robust attention modeling even when one modality is noisy or ambiguous, advancing the field beyond unidirectional or usual fusion pipelines.

C. Realistic Applications Need

Audio-visual saliency prediction (AVSP) has impactful applications in domains such as human robot interaction [4], autonomous driving[6], video surveillance and assistive robotics. These scenarios demand accurate and generalizable multimodal models. Even though our implementation focuses on model innovation, the architecture of is motivated by these practical needs, aiming to serve as a step toward deployable and perceptually aligned systems.

1.3 Problem Statement

Although deep learning has significantly advanced visual saliency prediction, audio-visual saliency prediction (AVSP) remains a relatively underdeveloped field, particularly in modeling human like attention mechanisms[10, 31]. Humans naturally process visual and auditory information in a temporally aligned and mutually reinforcing manner, but existing state-of-the-art models such as DiffSal [28] and CASP-Net [27] often fall short in simulating this process. These models typically rely on unidirectional or static fusion techniques, which limit their ability to represent the complex and dynamic relationships between modalities. In particular, they are unable to fully model how visual motion and audio events interact across time, leading to weak audio-visual correlation and limited saliency prediction fidelity in dynamic scenes.

A central challenge in AVSP lies in the temporal misalignment between audio and visual streams: while video content is processed as discrete frames, audio signals span continuous and overlapping intervals, making synchronized feature extraction and fusion difficult [18]. Furthermore, to the best of our knowledge existing state of the art methods[26, 27, 28] lack adaptive mechanisms for dynamically adjusting the importance of each modality based on the scene context. This often results in poor performance, especially in cases where salient cues are ambiguous or one modality provides stronger evidence than the other. These limitations point to the need for a unified framework that can jointly align, fuse, and reason over both modalities in a temporally coherent and context-aware manner. To address these gaps, our work introduces a novel architecture that combines bidirectional efficient cross attention (BECA) and a temporal guided noise encoder with conditional diffusion based iterative denoising, enabling robust and accurate saliency prediction under complex audio-visual scenarios.

1.4 Research Question

RQ1: How can cross fusion techniques improve multimodal integration for enhanced audio-visual saliency prediction?

RQ2: How can temporal differences between modalities be synchronized while retaining spatial information to improve performance?

RQ3: How does fused conditional diffusion with iterative denoising contribute to high-resolution, visually coherent saliency mapping?

1.5 Objective

1.5.1 General objective

To Develop an end to end multimodal audio visual saliency prediction that effectively integrates audio and visual information.

1.5.2 Specific objective

- To develop and implement cross fusion techniques, leveraging bidirectional efficient cross attention (BECA) to improve multimodal feature integration in saliency prediction.
- To design and optimize methods for synchronizing temporal differences between audio and visual modalities while preserving spatial coherence for enhanced saliency mapping.
- To explore the impact of fused conditional diffusion with iterative denoising on generating high-resolution, visually coherent saliency maps with improved prediction accuracy.
- To evaluate the proposed model and compare its robustness and prediction quality against the state of the art works.

1.6 Significance

The significance of this research arises directly from addressing core limitations identified in existing audio-visual saliency prediction models. While prior works often rely on unidirectional or static fusion approaches, this study introduces a Bidirectional Efficient Cross Attention (BECA) mechanism that enables dynamic, mutual refinement between audio and visual modalities across time. This leads to stronger audio-visual correlation, which has been previously difficult to model.

Additionally, by integrating a Temporal Guided Noise Encoder with a Conditional Diffusion-based Decoder, the model not only synchronizes asynchronous audio-visual streams but also preserves fine-grained spatial details during iterative denoising. This design advances saliency prediction performance under complex, dynamic scenes where traditional models (e.g., DiffSal, CASP-Net) have shown limited robustness.

Practically, these architectural innovations make the model more adaptable to real-world scenarios where audio-visual interactions are temporally imbalanced and context-dependent. The adaptive gated fusion ensures resilience in scenes with ambiguous or conflicting modality signals. These contributions collectively offer a novel technical pathway toward generalizable, interpretable, and robust audio-visual saliency prediction.

1.7 Scope

The scope of this research focuses on advancing multimodal audiovisual saliency prediction by identifying possible challenges and investigate publicly available datasets with diverse content, sufficient size, and reliable ground truth annotations by accompanied different preprocessing stratagies. Furthermore, our work also covers the development training mechanisms and transfomer and conditional diffusion based architectural innovations to address the challenges in multimodal

saliency prediction, focusing on the integration of audio and visual modalities. The primary focus is on optimizing factors such as temporal alignment, computational efficiency, and prediction accuracy as measured by metrics such as correlation cofficient. However, this multimodal approach is limited to audio and visual inputs, excluding other sensory modalities. Additionally, subjective factors such as personal preferences, cultural background, and specific application requirements are not considered. The study aims to enhance the technical algorithmic and computational aspects of multimodal saliency prediction by improving the integration of audio and visual data, optimizing computational efficiency, and increasing the accuracy of saliency predictions. Comprehensive experiments has been conducted to evaluate the proposed approaches, comparing their performance against existing models and assessing metrics such as temporal alignment accuracy, computational efficiency, and overall prediction accuracy.Moreover, the proposed research aims to lay a strong foundation for future advancements in multimodal saliency prediction with real time processing, potential applications in video surveillance and human-robot interaction.

1.8 Contribution

This research aims to tackle the challenges in multimodal saliency prediction, focusing on the seamless integration of audio and visual data through multimodal unified bidirectional latent space cross modal attention representation shortly MUBiC.

- **Theoretical Advancements:** Prior audio-visual saliency models often treat modalities as static, isolated inputs, leading to suboptimal alignment of transient events[18]. MUBiC pioneers a bidirectional cross-modal attention framework that dynamically aligns audio and visual cues through mutual conditioning. By formalizing saliency prediction as a joint spatio-

temporal-auditory optimization problem, we advance the theoretical understanding of how human attention shifts are governed by both modality specific saliency and cross-modal reinforcement. This resolves long standing challenges in temporal misalignment and fragmented feature fusion.

- **Architectural Enhancements:** This study proposes novel architectural modifications, which introduces a novel architecture that bridges multimodal fusion and saliency generation. The contributions of this work are as follows.
  - **Unified Multimodal Fusion through Multiscale Bidirectional Cross Attention:** To the best of our knowledge, this work is the first to introduce a multiscale bidirectional cross attention mechanism for audiovisual saliency prediction. Unlike methods that apply only visual to audio guidance[27, 28] or basic concatenation [24], our design allows both modalities to refine each other dynamically across multiple temporal levels, enabling robust cross modal feature interaction in complex scenes.
  - **Content Aware Gated Fusion for Modality Relevance:** While related models specifically Casp-Net [27] introduce context-aware mechanisms, they do not employ an adaptive gating strategy that dynamically balances both modalities. Our gated fusion unit adaptively modulates the contribution of visual and audio features based on their contextual relevance, filtering out modality specific noise. This enables robust fusion under conditions of semantic imbalance or background clutter.
  - **Noise Aware Temporal Decoding for Stable Predictions:**As compared to prior work SOTA Diffsal[28], which models temporal dynamics using convolution, our decoder explicitly integrates temporal

noise features with audiovisual fusion maps. This improves temporal coherence, suppresses flickering, and leads to stable saliency predictions across consecutive frames.

- **Hierarchical Diffusion-Based Saliency Refinement:**By introduce a hierarchical decoder that progressively refines noisy saliency maps across multiple stages using a diffusion-based generation strategy. Unlike shallow or single-pass decoders[26, 27], this structure enables fine-to-coarse spatial refinement, achieving sharp and accurate saliency localization aligned with human attention.
- **Significant performance:** gain, outperforming state-of-the-art methods by achieving improvements of up to 11.52% (CC), 20.04% (SIM) and 3.79% (NSS) across evaluation metrics on the AVAD dataset.

1.9 Thesis structure

The remainder of this chapter is structured into the development of the proposed technique. In Chapter 2, provides an overview of the foundational concepts as literature review and related research works, encompassing various aspects such as visual saliency prediction which includes spaital saliency prediction, and Spatio-temporal saliency prediction with the deep learning model approach, aduio saliency prediction, audio visual saliency prediction as related work having intensive comparative analysis and their research gaps. Chapter 3 details the proposed audiovisual saliency prediction model. In Chapter 4 outlines the experimental setup and evaluation of the existing works. Finally, Chapter 5 summarizes its our research work key contributions and offering valuable recommendations for future research endeavors.

Chapter 2

Literature review

This section explores the application of artificial intelligence in the prediction of audiovisual saliency. We dive into the specific network architectures employed and their integration mechanisms. The discussion progresses from unimodal approaches for visual and audio saliency detection, which analyze each modality independently, to a more comprehensive exploration of combined multimodal saliency prediction techniques. Finally, we present a comparative analysis of these models for acquiring an improved performance that mimics human attention directed in audiovisual content.

2.1 Search Strategy

To compile the relevant literature for this review, a comprehensive and systematic search strategy was applied, following the PRISMA 2020 guidelines[32]. Multiple academic databases were used, including IEEE Xplore, Springer Nature, ACM Digital Library, Elsevier, MDPI, Wiley, ScienceDirect, arXiv, and PLoS Biology. In addition, leading computer vision and AI conferences such as CVPR, ECCV, ICCV, and IROS were examined. The search was tailored using a combination of relevant keywords, including 'attention','gaze', 'saliency prediction', 'visual saliency', 'multimodal saliency prediction', 'audiovisual attention,'audiovisual saliency', and 'deep learning for saliency prediction'. This ensured the inclusion of foundational and state-of-the-art contributions within the field.

A total of 2,432 records were initially retrieved: 1,864 from journals and

568 from conference proceedings. After eliminating 231 duplicate records, 2,201 unique entries were screened based on title and abstract, resulting in the exclusion of 1,544 irrelevant articles. The remaining 657 articles were selected for full-text review. Out of these, 134 were excluded due to irrelevance or outdated publication years. From the 523 papers retained, 87 were removed due to full-text inaccessibility. The remaining 436 articles underwent detailed eligibility screening based on predefined inclusion and exclusion criteria. Ultimately, 102 peer-reviewed papers were selected for the final systematic review. The entire process is illustrated using the PRISMA framework in Figure 2.1.

Table 2.1: Databases and Conferences with Web Addresses

Database / Other conferences	**Web Address**
IEEE Xplore	https://ieeexplore.ieee.org/
Springer Nature	https://www.springer.com/
ACM	https://dl.acm.org/
arXiv	https://arxiv.org/
MDPI	https://www.mdpi.com/
PLoS Biology	https://journals.plos.org/plosbiology/
Wiley	https://onlinelibrary.wiley.com/
Elsevier	https://www.elsevier.com/
Science Direct	https://www.sciencedirect.com/
CVPR	https://cvpr2024.thecvf.com/
ECCV	https://eccv2022.ecva.net/
ICCV	https://iccv2023.thecvf.com/
IROS	https://www.iros2024.org/

Inclusion Criteria

- Only very key works prior to 2018.
- Peer-reviewed papers (2018–2024).
- Significant to research objectives.
- Deep learning-based models only.
- Eye-tracking datasets used.
- Central to saliency prediction.
- Use of audio, visual, or multimodal data.
- Standard evaluation metrics applied.

Exclusion Criteria

- Non-English articles.
- Not peer-reviewed or off-topic.
- Lacking relevant findings.
- Duplicate studies.
- Weak methodology or validation.
- Inaccessible or incomplete text.
- No empirical evaluation.

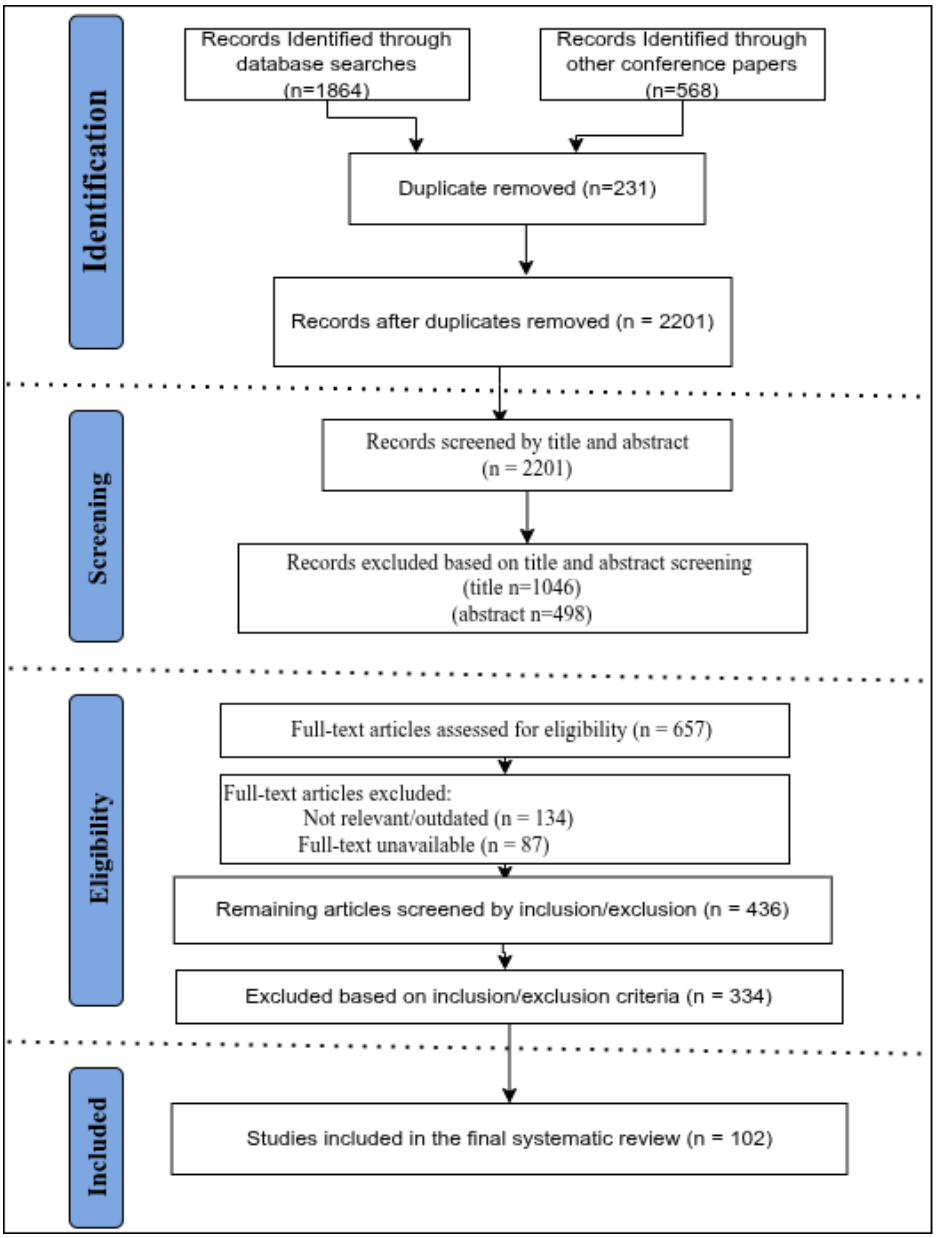


Figure 2.1: PRISMA Framework Selection Process

Quality Assessment: To ensure the reliability of this review, only peer-reviewed journal articles, conference papers, and high-quality reviews were included. Duplicates were carefully removed, and additional quality checks were applied based on citation count, publication venue, research methodology, and clarity of objectives and evaluation metrics. These criteria ensured the inclusion of only rigorous and relevant studies.

2.2 Background

2.2.1 Architectures for Saliency Prediction

Existing research papers adopted various deep learning architectural approaches in the area of spatiotemporal visual saliency prediction. Thus models include 3DCNN, RNN(LSTM), Transformers networks and Diffusion Models.

Convolutional Neural Network(CNN)

CNN [33] is the prominent and well known type of neural network for the application of saliency detection. It has mainly three components, which includes convolution layer, pooling layer, fully connected layer. Convolutional layer uses filters that slides across the input data(an image) and perform a convolution operation to extract features like edges, shapes and textures. The pooling layer down samples the data by summarizing local features and reducing computational cost and dimension. The third layer is fully connected layer also called classification layer, which is similar to simple neural network to combine features extracted from previous layer and make final prediction like object detection or classification. The generalized architecture of CNN is shown in figure 2.2. As stated in the

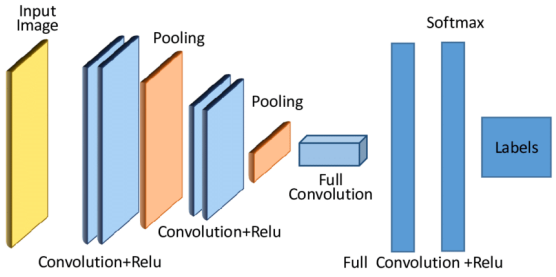


Figure 2.2: General Form of a CNN Architecture for Audio-Visual Feature Extraction [34]

above architecture the output feature map after a convolution with a kernels performed goes in to an activation function like Relu, softmax, and sigmoid. Then a down sampling mechanism applied to the the output feature maps. There are several types of pooling methods but the most commons are max-pooling and

average pooling. Max pools selects the highest value from a given N x N patches, whereas average pooling takes it average value. Finally, the pixel values of the feature map flatten in to a scalar vector and feed as input to the fully connected layer. Then as usual neural network, it computes a classification task.

Padding and stride are other important hyper-parameters for CNN. Padding means adding extra rows and columns of an input image at its boarder. It can be same(no) padding or valid(the output size is the same as the input) padding. The advantage of padding involves maintaining image size, prevent artifacts at image edges and provide context for edge pixels. On the other hand, Stride means the number of pixels which the filter moved across the image during convolution [33]. For a size of N X N input feature map(an image):

$$F = \lfloor ((N + 2P - f)/2S) + 1 \rfloor \tag{2.1}$$

Where F is the output feature map after floor function of performed in input pixel size N with padding size P, stride S and filter size fXf that results a size of FXF. 3D convolutional model architectures has been explored in slaiency estimation of visual scenes.

(i) **Architecture of CNN:** Various architectures have emerged, each with distinct features and performance metrics. LeNet [34], introduced in 1998, was designed to recognize digits from 32x32x1 grayscale images, without employing softmax in its concluding layer, which is unlike contemporary models. It includes 60,000 parameters, utilizing a convolutional-pooling framework progressing to fully connected layers prior to softmax, with Sigmoid and Tanh as activation functions. Modern methods generally favor ReLU for its superior efficiency and performance. AlexNet [35], created for the ImageNet challenge, was tasked with classifying images into 1000 categories, representing a notable enhancement over LeNet-5. In contrast

to the simpler layout of LeNet-5, AlexNet is characterized by a deeper architectural design: Conv → MaxPool → Conv → MaxPool → Conv → Conv → Conv → MaxPool ⇒ Flatten ⇒ FC ⇒ FC ⇒ Softmax. With 60 million parameters compared to LeNet-5's 60,000, AlexNet introduced ReLU activation and harnessed multiple GPUs for computation, adapting to the slower GPUs available at that time. Three years later, VGG-16 [36] was developed as a modified version of AlexNet. This model simplifies its architecture significantly, focusing on key blocks: CONV using 3x3 filters, stride 1, same padding, and MAX-POOL with 2x2 filters, stride 2. Despite the simplified architecture, the model is still considered large even by today's standards, featuring approximately 138 million parameters, primarily located in the fully connected layers. It requires around 96MB of memory per image just for forward propagation, with initial layers accounting for most of the memory usage. The number of filters increases steadily from 64 to 128, then to 256, and ultimately doubles to 512. Pooling operations consistently perform dimensionality reduction throughout the network. An alternative version, VGG-19, offers greater depth, but VGG-16 remains more popular due to its similar performance with fewer layers.

GoogleNet [37], developed in 2014 by Google, emphasizes efficiency by reducing parameters, memory, and computation. It employs a lightweight stem for aggressive downsampling, avoiding large spatial feature maps like those in VGG or AlexNet. The Inception module introduces parallel branches to eliminate kernel size hyperparameters and uses 1x1 bottleneck layers to reduce channel dimensions. GoogleNet ends with global average pooling and a single linear layer for class scores, unlike VGG's fully connected layers. Before the discovery of batch normalization, training networks with more than 10 layers was difficult. ResNets [38], introduced

in 2015, effectively address this issue. They consist of residual blocks, each usually with two convolutional layers. Between blocks, skip connections add or concatenate input and output, facilitating gradient flow and reducing error. Blocks can be identity or Convolutional (Conv) with batch normalization applied after convolution in Conv blocks. ResNets also use global average pooling with a linear layer at the end. Popular ResNet models include ResNet18, ResNet32, ResNet50, ResNet101, and ResNet152.

Table 2.2: Comparison of CNN Architectures

Architecture name	Authors	Year	Memory(MB)	Parameters(million)	Error rate(ImageNet)
LeNet	LeCum et al.[34]	1998	6	6	30
AlexNet	Krizhevsky et al. [35]	2012	1.9	61	16.4
VGG-16	Simonyan et al. [36]	2014	48.6	138	7.3
GoogleNet	Szegedy et al. [37]	2014	27	6.8	6.6
ResNet	He et al. [38]	2015	240	60.2	3.6

(ii) **3D CNN for video:** 2D CNN performs tasks of spatial data which lacks capturing temporal dynamics within video sequences. So each inputs are a single frames and processed each frame independent of temporal information. As a result, 2D CNNs limits their effectiveness in tasks related to video analysis.In contrast, 3D CNN are purposefully designed to overcome these shortcomings by integrating temporal dimensions into the convolutional framework. By employing 3D filters that extend across both spatial and temporal dimensions, 3D CNNs can effectively learn features that encapsulate motion and changes over time.Its architecture is the same to that of 2D except formulating and designed for 3D shapes like height, width and depth(number of frames in videos) i.e it takes (H(height),W(width),D(depth),Cin(channel), number of filter), kernel size and pooling size having(f(filter size),f,f) and (fp(pooling kernel),fp,fp) respectively. To extract the spatial feature it utilizes 3D filters convolve over input volume which consists of multiple frames stacked together. This en-

ables the network to capture the spatial features like edge, texture and shape, and the temporal features like motion and the change over time of the consecutive frames[39, 40, 41].

$$y(i,j,k) = \sum_{h=0}^{H-1} \sum_{d=0}^{D-1} \sum_{w=0}^{W-1} x(i+h, j+w, k+d).w(h,w,d) \tag{2.2}$$

where: y(i,j,k) is the output value at position (i,j,k). H,W,D are the height, width, and depth (or time) dimensions of the kernel, respectively. x is the input volume and w is the 3D filter (kernel).

Hara et al. [42] developed 3D Residual Networks (3D ResNets) for action recognition, adapting the conventional residual network framework to three dimensions to enable coherent processing of video frames. By integrating residual connections, their model effectively addresses the vanishing gradient issue, facilitating the training of deeper networks. This methodology has resulted in notable enhancements in action recognition tasks, achieving leading performance on benchmark datasets such as Sports-1M and Kinetics-400. However, despite these advancements, the significant computational demands of 3D convolutions present a challenge, necessitating considerable memory and processing resources, which can restrict real-time applications. Building on this foundation, Zhu et al[43] proposed an entropy-based method to enhance the efficiency of 3D CNNs for video recognition. Their approach focuses on maximizing spatio-temporal entropy within the network, directing the learning process toward the most informative segments of the video data. This entropy-driven strategy balances computational efficiency with recognition accuracy by dynamically adjusting the network's focus based on entropy measures. The model demonstrates reduced computational costs while maintaining high

accuracy, achieving competitive performance on datasets like Kinetics and ActivityNet. However, accurately estimating and maximizing entropy introduces computational complexity and requires sophisticated optimization techniques. Additionally, the dynamic adjustments based on entropy measures can lead to instability during training, necessitating careful hyperparameter tuning to ensure robust performance. Together, these studies highlight the evolving landscape of 3D CNNs, showcasing innovative solutions to enhance their capability and efficiency in video analysis.

Recurrent Neural Network(RNN)

Neural network models face several challenges, particularly when dealing with sequential data [44, 45, 46]. First, input and output lengths can vary between different training examples, and while padding can address this issue, it is not always ideal due to the introduction of unnecessary information and computational overhead. Second, traditional neural networks do not share the features learned in different positions in the text, limiting their ability to generalize. Convolutional Neural Networks (CNNs), though effective for spatial features, struggle with temporal information [47]. Even 3D CNNs, which can capture temporal features, are insufficient for long-range sequential data. Sequential datasets like video analysis require specialized training approaches to provide the necessary machine learning work.

Recurrent Neural Networks (RNNs) are designed to handle such sequential data by incorporating feedback connections within the network, enabling them to maintain information over time [48, 49]. However, traditional RNNs only uses information that is earlier in the sequence to make prediction. So that it struggle with long-term dependencies due to the vanishing gradient problem and also face issues with data parallelism [50, 51].

To solve the issue of vanishing gradient problem a special type of RNN called

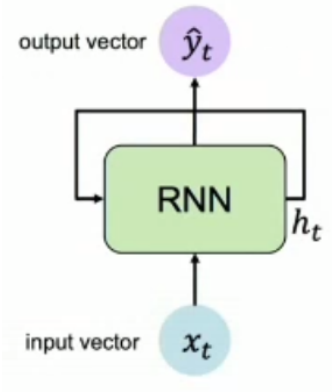


Figure 2.3: Simplified RNN Architecture [45]

Output Vector

$$\hat{y}_t = W_{hy}^T h_t \tag{2.3}$$

Update Hidden State

$$h_t = \tanh(W_{hh}^T h_{t-1} + W_{xh}^T x_t) \tag{2.4}$$

Long Short-Term Memory (LSTM) is introduced [52]. LSTMs, excel in modeling long-term dependencies in sequential data by using memory cells that can selectively remember or forget information over multiple time steps. This makes LSTMs highly effective for tasks such as natural language processing, speech recognition, and time series analysis. The cell comprises several components: the forget gate, the input gate, the candidate cell state, the cell state, and the output gate [52, 53, 54, 55]. These components are mathematically defined as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget gate)} \tag{2.5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input gate)} \tag{2.6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate cell state)} \tag{2.7}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad \text{(Updated cell state)} \tag{2.8}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output gate)} \tag{2.9}$$

$$h_t = o_t \cdot \tanh(C_t) \quad \text{(Hidden state)} \tag{2.10}$$

The forget gate decides which information from the previous cell state should be discarded. The input gate determines the relevance of new information from the input. The candidate cell state represents new potential information to add to the cell state. The updated cell state is a combination of the previous state and

the new candidate state, controlled by the forget and input gates. The output gate selects which parts of the cell state should be used as the output. Finally, the hidden state is computed by applying the output gate to the cell state after passing it through a hyperbolic tangent function.

Despite LSTM strengths in effectively mitigates the vanishing gradient problem, allowing for long-term memory retention and controlled information flow, thereby enhancing the modeling of sequential data. However, it still face data parallelism challenges.

Transformers

CNNs are better suited for spatial data, but their fixed filter size limits their ability to capture temporal information in sequential data. On the other way, RNN have traditionally been the go-to models for sequential data, but their sequential processing nature limits their ability to leverage modern GPUs for parallel computation. This results in slow training times, especially for long sequences. Additionally, the vanishing and exploding gradient problems can hinder RNNs' ability to capture long-term dependencies. This led to the development of Transformers for more efficient handling of such challenges. The research titled "Attention all you need" introduced transformer [56]. The backbone of this architectural model is attention. Attention mechanism computes which part of the input should desired focus. It derived from query(Q), key(K) and Value(V) vectors. It enable models to prioritize relevant input elements by assigning weights to them. They function within encoder-decoder frameworks, allowing the model to focus on specific parts of the input sequence during decoding.

Primarily, it has encoder and decoder blocks, which accepts the concatenated vector from input embedding and positional encoding. The input sequence is converted into fixed-dimensional vectors, capturing the semantic meaning of each word or token. These embeddings are learned during training. Since Transformers

lack an inherent understanding of word order, positional encodings are added to provide information about the position of each token in the input sequence. The self attention mechanism of the transformer can be formulated as:

$A(q, k, v) =$ attention based representation of a word

$Example =$ I visit Addis tommorow

$$A(q, k, v) = \sum_i \left(\frac{e^{q \cdot k_i^\top}}{\sum_j e^{q \cdot k_j^\top}} \right) \cdot v_i \tag{2.11}$$

Each word in the example has corresponding $q, k,$ and v vectors.

For example, let's calculate the attention for the word "Addis":

$q_{\text{Addis}} =$ query vector for "Addis"

$k_{\text{I}}, k_{\text{visit}}, k_{\text{Addis}}, k_{\text{tomorrow}} =$ key vectors for each word

$v_{\text{I}}, v_{\text{visit}}, v_{\text{Addis}}, v_{\text{tomorrow}} =$ value vectors for each word

Attention weight for "Addis":

$$\alpha_{\text{Addis},i} = \frac{e^{q_{\text{Addis}} \cdot k_i^\top}}{\sum_j e^{q_{\text{Addis}} \cdot k_j^\top}} \quad \text{for each } i \text{ in the sentence} \tag{2.12}$$

The attention-based representation of "Addis" is then:

$$A_{\text{Addis}} = \sum_i \alpha_{\text{Addis},i} \cdot v_i \tag{2.13}$$

Generally , the equation looks like [56]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \right) V \tag{2.14}$$

Multi-head attention employs multiple attention heads, each of which is independently trained and focuses on different aspects of the input sequence. These heads can learn to capture different patterns, dependencies, and relationships within the data. For instance, one head might focus on long-range dependencies, while another might concentrate on local context. By combining the outputs

of these multiple heads, the model can obtain a richer and more comprehensive representation of the input sequence. This allows the model to capture a wider range of relationships and dependencies within the data.

The output from the attention mechanism is passed through a feed forward neural network consisting of two linear transformations with a ReLU activation in between. A residual connection is added to the network's output, combining it with the input to enhance gradient flow during backpropagation. Finally, layer normalization is applied to the combined output, stabilizing and accelerating the training process.

Vision Transformer(VIT):The traditional transformer has been developed for NLP tasks [56]. The inspiration of this model also led to the development of an encoder-only transformer model for computer vision application called a vision transformer (VIT) [57] as shown in the figure 2.4.

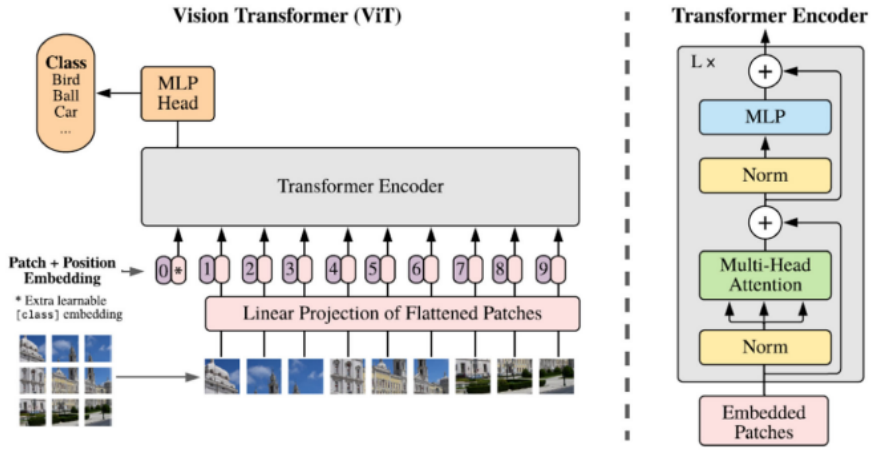


Figure 2.4: Vision Transformer Architecture [58]

It works by breakdown the input image or frame into smaller patches. each patch is typically a square regions of image. In the paper [57] an image is worth P of 16 by 16 pixel value patches. For an image having H and W with a channel C, the total number of patches N can be calculated as:

$$N = \frac{H \times W}{P \times P} \tag{2.15}$$

The stride and the patch size P is the same and it ensures there will be no overlap between patches. These patches should be flatten converted from 2d to 1D. The linear projection layer lowers the its dimension. Lowering the dimension benefits for small memory and computation expense. Then a positional encoding is added to image patch that indicates the patch location in the image. It helps to maintain the spatial order of the patches by adding unique positional information to each patch embedding.

The next stage is the a multi-head self attention mechanism used, allowing the model to weigh the importance of different patches relative to each other and capture complex dependencies between them. The Q , K and V represents the path that is querying the importance of other patch, the patch being queried and the actual content of the patch respectively. Then the feed forward network received the output of the each patch to capture more complex non linear relationships with in the patches. A linear layer is then added for further computational tasks like classification or object detection.

Diffusion Model

Diffusion models have recently emerged as powerful generative frameworks in deep learning, notable for their strong generalization and high-quality outputs in image, video, and multimodal generation tasks [28]. The core idea is to model data generation as a Markov process, where clean data (such as a ground-truth saliency map) is gradually corrupted by noise in a forward process, and a neural network is trained to reverse this process, denoising the data step by step. Formally, let $\mathbf{x}_0$ denote the original data and $\mathbf{x}_T$ the fully noised version. The forward process adds Gaussian noise at each timestep:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \qquad [58] \tag{2.16}$$

where β_t is a variance schedule. The reverse process, parameterized by a neural network ϵ_θ, attempts to reconstruct $\mathbf{x}_0$ from $\mathbf{x}_T$ by iteratively removing noise, optimizing a mean squared error loss between predicted and true noise. In conditional diffusion models, auxiliary information (such as audio or video features) is used to guide the denoising process, enabling controllable generation.

In the context of saliency prediction, diffusion models offer a conditional generative approach in which the model is conditioned on both audio and video cues to generate saliency maps. Traditional audio saliency prediction approaches have typically relied on hand-crafted features or direct neural fusion of audio and visual cues, often requiring customized architectures and task-specific loss functions. The introduction of diffusion models marks a shift toward a more principled, generative approach: By treating audio features as conditional information, diffusion-based models can more robustly model the temporal and semantic relationships between sound and visual attention. This generative formulation allows the model to capture complex dependencies and interactions, outperforming previous architectures that were limited by their reliance on specific network designs or heuristic loss functions [28].

2.3 Modality Type

2.3.1 Visual saliency prediction

Our vision constantly receives a flood of information, but visual saliency acts as a filter, grabbing our attention towards specific areas in an image or video [10]. Visual saliency prediction, attempts to replicate this mechanism by identifying the most salient regions – those most likely to capture our initial gaze. These predictions can be categorized as image saliency, focusing on contrasting within a spatial features, or video saliency, which additionally considers how the scene changes over time, which can be seen as spatio-temporal feature detection [31].

By understanding both types of saliency, we can develop computer vision models that better predict human attention, impacting areas like image compression and human-computer interaction [59].

A. Spatial Saliency Detection

The mechanism of identifying the the most attractive or significant region from static images or a video frame. As the name spatial indicates space, spatial saliency involves analyzing the spatial distribution of features from the region within the image that draw attention [9].

This domain often focuses solely on spatial features, neglecting the temporal dimension of videos. In other words, it analyzes individual frames without considering how elements move or change over time. Furthermore, it typically operates in two dimensions, limiting its ability to handle complex 3D information [7, 8].

These limitations prevent its application to video analysis, and researchers are increasingly delving into spatio-temporal saliency detection approaches. This method incorporates spatial and temporal information, allowing it to account for movement and changes within a video sequence. This broader approach can address the drawbacks mentioned above and provides a more holistic understanding of visual saliency in videos.

B. Spatio-temporal Visual Saliency Prediction

Spatio-temporal Saliency prediction (STSP) maps predictions based on both spatial features within a frame and dynamic temporal information across frames in video sequences. Although the broader concept of spatiotemporal analysis finds applications in diverse areas such as geospatial analysis and financial data [12], this section focuses solely on its use for video saliency prediction.

In the field of deep learning-based video saliency detection, most research can

be categorized by the model architecture used into bistream [60, 61, 62] and single stream architectures [39, 63]. Bi-stream architectures for STSP utilize two sub-branches of network design one for the input of colors of an which are spatial distribution features and the other sub-branches for the input of temporal information which is motion saliency clue. In contrast, the single-stream architecture integrates spatial and temporal information by handling spatial features from individual frames and temporal features capturing motion across frames in a unified pathway. The two types of models also differ from each other in various aspects. Some of these includes network design, processing path, information flow, fusion level and interpretability. The following table shows the generalized expression of these two architectures.

Table 2.3: Bistream Vs Single Stream Ways Design

Feature	Bi-stream [13, 64, 65, 66, 67, 68, 69, 70]	Single Steam [39, 63]
Processing Path	Separate	Combined
Information Flow	Later Fusion	Internal Fusion
Learning Focus	Separate	Unified
Network Design	Complex	Flexible
Interpretability	Easier	Less Easy
Generalizability	More	less
Complexity	High	Low
Efficiency	Low	High
Relationships	Better	less

Bi-stream designs capture complex interactions but reduce computing efficiency and increase complexity. Single-stream architectures combine temporal and spatial data, allowing for faster training and simpler design, though they may miss intricate spatiotemporal linkages. The choice between them depends on the task's focus: Bi-stream is suitable for detailed linkage recording if resources are available, while Single Stream is better for simplicity and efficiency.

Studies on video saliency focuses on two main tasks: video salient object detection (VSOD) and video fixation prediction (VFP). VSOD targets identifying and segmenting prominent objects in videos using manual annotations and

cross-entropy loss, often employing bi-stream architectures. In contrast, VFP predicts human eye fixations using fixation data as labels, optimizing with KL divergence and NSS, and typically using single-stream networks with fusion techniques. These tasks differ in their goals, training labels, loss functions, and network designs [18].

3D CNN based VSP: Various research works have been conducted for visual saliency prediction using an eye fixation data [71, 72, 73, 74, 75, 76, 77]. To the best of our knowledge, 3D convolution-based Vsp models [77] are generally leading the SOTA performance interms of its accuracy and efficiency.

Wang et al. [74] developed Spatio-Temporal Self-Attention Network (STSANet), which integrates self-attention mechanisms into a 3D convolutional framework to address the limitations of traditional 3D convolutions that struggle with capturing long-range dependencies across temporal frames. The proposed STSA module enables the model to learn long-range interactions between features at different time steps, allowing it to focus on relevant visual information over time, which is crucial for accurately predicting salient regions in videos. Additionally, the Attention Multi-Scale Fusion (AMSF) module fuses spatio-temporal features from various levels of the 3D backbone, enhancing the model's ability to capture contextual information and narrow semantic gaps during feature fusion, resulting in more coherent saliency maps.

On the other hand, Bellitto et al. [73] have introduced a 3D encoder-decoder network architecture for the VSP task. A key feature of their approach is the innovative decoder design, which incorporates two novel concepts. To address the issue of domain shift, they assign an unsupervised binary classifier to each side output of the encoder. This classifier is primarily aimed at facilitating adversarial training, which helps to reduce the disparity between features learned from the source and target domains. Additionally, the decoder dynamically learns and integrates multiple domain-specific priors at each layer, enhancing the net-

work's ability to adapt to specific domains. This strategy has the potential to significantly boost quantitative performance metrics.

Recently, Hu et al. [77] presented TinyHD, a novel approach to video saliency prediction by employing a lightweight, heterogeneous multi-decoder architecture that significantly enhances computational efficiency while maintaining high accuracy. Unlike the others, TinyHD utilizes multiple simple decoders, each designed to capture different aspects of salient regions in videos. This strategy allows the model to effectively combine the strengths of various decoding methods, leading to improved saliency predictions without the computational overhead typically associated with more complex architectures. The hierarchical multi-map knowledge distillation employed by TinyHD is particularly noteworthy. This technique allows the model to learn from a teacher network that provides multiple hierarchical predictions, enabling it to capture a broader range of features and patterns. By leveraging this knowledge distillation approach, TinyHD can generate high-quality saliency maps while keeping the model size to a mere 16 MB, making it suitable for deployment on low-power devices. This is a significant advantage in practical applications where computational resources are limited. In contrast STSANet, which si for long-range dependencies across temporal frames, TinyHD focuses on reducing model complexity through its multi-decoder strategy. While STSANet excels in modeling intricate temporal interactions, TinyHD's design prioritizes efficiency and the ability to produce multiple saliency maps simultaneously and aims for a balance between performance and computational cost.Furthermore, TinyHD's performance on benchmarks such as DHF1K, UCF-Sports, and Hollywood2 demonstrates its effectiveness in generating accurate saliency predictions made better than state-of-the-art methods.

RNN based VSP: consecutive advancements in VSP have utilized RNNs to highly competitively capturing the long range temporal dependencies in video sequences [78, 79, 80, 81, 82, 83]. Linardos et al. [80] enhanced traditional LSTM

frameworks by placing the LSTM in the middle stage of an encoder-decoder CNN architecture. The encoder extracts spatial features from input frames, and these features are fed into the LSTM to capture temporal dependencies. The LSTM's output, which contains spatiotemporal information, is then passed to the decoder to generate the final fixation predictions. The key innovation in their approach is the use of a recurrent mechanism where the LSTM's output serves as intra-attention, significantly improving the network's ability to sense temporal changes and enhancing overall performance. After a while Chen et al.[79] advanced the LSTM-based VFP by using three consecutive frames as input instead of one. This approach enhances the network's ability to capture temporal dynamics over multiple frames. They addressed the issue of spatial misalignment of features from consecutive frames by incorporating deformable convolutions. This technique dynamically adjusts the spatial positions of convolutional kernels, aligning features before they are input to the LSTM. This alignment reduces confusion in the learning process, leading to more accurate fixation predictions. By considering multiple frames at a time, Chen et al. improved the network's temporal sensing capability and overall prediction accuracy.

Droste et al. [81] adopted a recurrent neural network (RNN), an early prototype of LSTM, to sense temporal information. They placed the RNN between the encoder and decoder, maintaining a network structure similar to [79]. The encoder extracts features, the RNN processes temporal information, and the decoder generates the predictions. This choice highlights the flexibility of the network structure and showcases the potential of simpler recurrent models in VFP tasks. By using an RNN, they demonstrated that alternative recurrent architectures could still effectively process temporal information and achieve robust VFP results.

These studies collectively advance the state-of-the-art in LSTM-based VSP through different approaches to enhance temporal information processing. [80]

improved temporal sensitivity using intra-attention mechanisms within the LSTM. [79] addressed spatial misalignment issues and enhanced temporal sensing by using multi-frame inputs and deformable convolutions. [81] demonstrated the potential of simpler RNN models as viable alternatives to LSTMs in VFP tasks. The progression from single-frame to multi-frame inputs and the exploration of different recurrent architectures highlight the dynamic nature of this research field and its ongoing evolution towards more accurate and efficient VFP models. These advancements reflect a deeper understanding of the challenges and innovative solutions in LSTM-based VFP, driving the development of more robust and precise models..

Transformer based VSP: The first pure-transformer framework for video saliency prediction was presented by Ma et al. [84]. Instead of only concentrating on earlier frames, the writers of the work have projected future frames' visual saliency in an educational manner. Because employing solely VFP methods may cause focusing regions to lag behind the scenes that are happening, it may become difficult to retain important things in the center of recorded films or to follow objects at a fast speed. In order to investigate temporal and spatial semantic information from input videos for video saliency forecasting, they have therefore suggested a video saliency forecasting transformer. Additionally, the time embedding layer uses a cross-attention decoder to remove the time dimension of the decoder feature.

Additionally, self-attention mechanism was presented by Wang et al. in [85] to get spatiotemporal correlations between saliency zones and features. The main distinction between this study and [70] is that fixation predictions were made by combining CNN, Transformer, and LSTM together. More specifically, a CNN-based DConvLSTM is employed as the decoder for dynamic information learning, and a CNN-based multi-scale feature-fusion network seeks to efficiently extract features in multi-category space. In both the temporal and space domains, the

Transformer encoder learns the global link between pixels and human visual attention.

2.3.2 Audio saliency prediction

The audio saliency prediction (ASP) task aims to identify significant changes in audio signals that can capture human attention. Unlike visual saliency, ASP is considered relatively straightforward due to the less informative nature of audio signals compared to visual cues. ASP focuses solely on audio signals, also known as audio saliency detection or salient event detection. Various non-deep learning-based approaches have been developed for ASP, emphasizing the concept of salient regions exhibiting high contrast with their surroundings.

In previous research, such as the work by Kayser et al. [86], audio signal changes were analyzed using multiple filters to identify salient fragments based on temporal variations in intensity and frequency. Similarly, Schauerte et al. [87] assessed audio saliency by calculating the KL-divergence between spectral histograms, introducing a non-local method as opposed to traditional local approaches. Tsuchida et al. [88] presented a non-local signal feature representation technique using 2D spectral histograms, leveraging principal component analysis (PCA) for feature extraction and contrast computation to determine saliency.

Other techniques, like those by Zlatintsi et al. [89], transformed audio amplitude and frequency into a 3D feature space utilizing Teager energy, emphasizing abrupt increases in loudness as key indicators of saliency. Merve et al. [90] expanded the methods for detecting audio saliency by introducing new features such as envelope, bandwidth, rate, and pitch over various time scales.

In general, Audio Saliency Detection (ASD) methods often rely on manually designed techniques, such as contrast computation for feature extraction and saliency identification. While these techniques are consistent with human attention processes, they tend to be computationally demanding and may lack

precision. The field remains largely influenced by traditional approaches, with only limited exploration of deep learning techniques, highlighting the need for further research and development.

2.4 Related Works

Previously, we discussed the unimodal aspects of saliency prediction, i.e., either video or audio[3]. It is known that only visual stimuli cannot entirely capture our attention. Most research has focused on video saliency. Treating only spatial and temporal features while ignoring auditory features as a source of information for capturing attention is not accurate [91]. A multimodal integration of audio cues with visual information significantly enhances the effectiveness of saliency prediction. When auditory and visual elements are combined, they create a richer and more engaging experience that can more accurately direct attention [18].

Audio-visual saliency prediction(AVSP) is in its infancy, with limited existing work available. Most early findings were based on handcrafted approaches, attempting to establish cross-modal connections between the two modalities. Most studies used canonical correlation analysis (CCA) to locate the moving target [2, 19, 20, 92, 93]. However, an end-to-end deep learning scheme is still under explored. To the best of our knowledge, there are Seven SOTA deep learning papers in this area [14, 24, 25, 26, 27, 28, 94]. Since the focus of this review is on end-to-end deep learning-based works.

Tavakoli et al [24] proposed an end to end trainable deep model for dynamic AVSP by prepared their own AVE dataset called DAVE.They proved audio cue is an important source for the task which outperforms having only spatio-temporal information. They adopted a learning based paradigm which follows encoder decoder neural architecture. Also formulated the saliency segmented video $V =$

I, A as the probability S:

$$S = P(S \mid I, A) = f(I, A) \tag{2.17}$$

where $f(.,.)$ is the neural network, and I, A are frame sequence and an audio signal respectively. These is a bi-stream network based on 3D ResNet18, one for video frames and the other for audio mel-spectrogram. Also after the audio signal transformed into log mel spectrogram, then it converted into a sequence of successive overlapping frames. To encoding the video and audio by taking the weights from a pretrained kinetic dataset developed for action recognition. Then, the two streams process 16 frames of each at a time. The encoded features are then concatenated for decoding by upsampling and fed to 2D CNN to get the final saliency map.

They developed two baseline models called the audio model and the video model, each with the same single-stream architecture, to compare with the AVSP model. They proved that the multimodal model provides better performance. Furthermore, when compared to earlier spatio-temporal models, it achieved the best saliency prediction score. As a result, at that time, it was the state-of-the-art finding for dynamic saliency prediction. However, the model is very simple and lacks diverse contextual annotation. Their work did not capture any attention mechanisms for the information, especially for temporal features, where attention mechanisms typically perform better. Additionally, the use of a 3D CNN can be inferior for temporal data, as it may not effectively capture long-range dependencies compared to more advanced models like transformers or LSTMs.

Similarly, Tsiami et al. [26] introduced a STAViS to solve saliency prediction in videos by incorporating audio cues. STAVis and [24] concurrently developed, which are the first deep learning saliency approach that employs both audio and video data. DAVE [24] tried to concatenate independent modalities for the two

modalities, where as STAViS provided a single like multimodal network which combines the two information at multiple stage and used a sound source localization and attention mechanism.

It consists of five modules for five computing tasks. These includes 1) Spatio-Temporal visual network, like [24] utilized 3D ResNet. Here, in this there are 4 blocks, which provides X output having different spatial and temporal scale. Also Deeply super vised attention module(DSAM) is applied to each feature channel map X^m and attention map M^m to enhance the most salient region.

$$\tilde{X}^m = (1 + M^m) \cdot X^m, m = 1, 2, 3, 4. \tag{2.18}$$

2) Audio representation network computes the audio features from its sound waveform by employed 1D CNN rather than as [24] its time frequency representation. Then cropped to 16 frames for balanced to visual frames. 3) After produced an audio representation, Sound source localization module performed to detect a cross-modal semantic concepts in videos. It clarifies the audio-visual correspondence to obtain their salience. 4) Module four is audiovisual saliency estimation to combined and fused auditory and visual saliency map. The auditory saliency map found from sound source localization. 5) Saliency losses like cross entropy, and linear correlation coefficient to compare the saliency map created in module four and the activation function with the ground truth. Their work improves the above all five metrics in AVAD dataset and counter dataset from all the 8 closest works of that time. Allover, Their work showed a strong interaction between visual and audio and retaining intact in spatial information.

In contrast, another fully convolutional encoder-decoder model called ViNet and its audio-visual variant AViNet has been developed by Jain et al. [25]. Their work significantly demonstrated that visual saliency can be effectively predicted using solely visual information, surpassing the performance of previous models

such as those by [24] and [26]. They showed that the earlier works did not extract visual data as efficiently, which led to inferior performance. Furthermore, Their study challenges the notion that audio cues are essential for saliency prediction. By incorporating audio features into ViNet, the resulting model, AViNet, achieved nearly the same results as the visual-only ViNet.

Their ViNet efficiently predicts saliency for the last frame in a sequence and outperforms SOTA models of the time across nine (3 visual only and six audio-visual) datasets. It is capable of real-time processing at 60 frames per second, making it suitable for practical applications. Despite these strengths, AViNet becomes agnostic to audio input after training, indicating that current audio integration methods do not significantly enhance performance. This indicated that their model either lacks the necessity of audio information or has not yet found an optimal method for incorporating audio features effectively.

previously studied[14, 24, 26] to show audio and visual frame correspondence for AVSP. However could not address what happened if there is inconsistency between audio and visual frames which is mismatch in timing between audio and visual elements. Those studied methods [14, 24, 26] have demonstrated the importance of audio-visual frame correspondence and effectively leverage the correlation between audio and visual modalities to improve AVSP. However, they do not adequately address the issue of temporal inconsistency between audio and visual frames, which refers to the mismatch in timing between audio and visual elements. This inconsistency can significantly impact the performance of AVSP systems, leading to inaccurate saliency maps. Xiong et al. [27] introduced Consistence aware audio visual saliency prediction (CASP-Net) to address this challenge by incorporating mechanisms to handle temporal inconsistencies. This models is composed of an encoder for two input streams, an audio visual interaction module , a consistency aware predictive coding module and the saliency decoder module.

Its two stream encoder indicates the video decoder which encodes the spatio temporal information by using S3D and The audio encoder processes audio data after converted into 2D spectrogram using VGGish network. Then the audio visual interaction module(AVIM) integrates the visual and auditory information by performing atrous Spatial pyramid pooling. The third is the consistency aware predictive coding(CPC) to evaluate the coherency of visual and audio features. The last successive decoder strategies is developed for saliency map estimation. Eventhough, Their work achieved a better performances when trained in various (six) datasets compared to previous works, it is still face challenge in addressing effectively the latency and delay that makes a temporal difference. Furthermore, their model is insufficient to capture semantic correlations and its iterative verification required higher computational demand.

More recently, another different approach has been explored using generative modeling, which is DiffSal [28] reframed audiovisual saliency prediction as a conditional task using diffusion models. By incorporating a Saliency-UNet architecture that modulates multimodal attention, DiffSal achieves a unified framework capable of handling both unimodal and multimodal inputs. This generative approach enables the model to produce highly accurate and coherent saliency maps, outperforming previous methods on challenging datasets. Althouhg having various challenges as discussed in the next analysis.

2.4.1 Comparative Analysis and Synthesis

The area of multimodal audiovisual saliency prediction (AVSP) remains relatively underexplored compared to models that predict saliency using only visual or auditory signals. Early research focused predominantly on visual saliency prediction from static images or videos, with only a few recent attempts to blend auditory and visual inputs [39, 44]. Limited attention to AVSP is due to the complexities of merging auditory and visual data, each with unique temporal and semantic

challenges. Although deep learning advancements have improved the prediction of unimodal saliency, merging auditory and visual streams is more complex [18]. Incorporating auditory signals is essential for human attention, but their fusion in deep learning remains limited to a few studies [14, 24, 25, 26, 27, 28].

. From these studies, the Dave method[24] simply concatenates the information from the two modalities, and it lacks mechanisms to handle the complex interaction between the two streams, leading to random temporal alignment and semantic coherence. On the other hand, Stavis[26] used DSAM and a sound source localization module. This model uses attention mechanisms to highlight important spatial and temporal features in both visual and audio streams and shows a better improvement in capturing audiovisual correlations and maintaining spatial information across frames, yet it still struggles with issues of temporal misalignment and global spatial interaction. However, another study Avinet[14] disproves the two previous works as they did not adequately use the audio signals. Because only visual stream of their architecture produced the same result. Later, CASP-Net[27], tried to better address temporal inconsistencies between audio and visual streams also to retian spatial features. This model is particularly effective in handling cross-modal timing mismatches by actively checking for alignment during training. Despite its improved accuracy, CASP-Net comes with a high computational cost.

Table 2.4: Transposed Summary of End-to-End Deep Learning Based AVSP Papers (STEA: Spatio-Temporal Encoder Architecture, AEBA: Audio Encoder Backbone Architecture, Att.: Attention Mechanism, Comx: Dataset Complexity, VPTM: Visual Pre-Training, APTM: Audio Pre-Training, Imp: Implementation Complexity)

Property	**Tavakoli et al. [24] (2019)**	**Tsiami et al. [26] (2020)**	**Jain et al. [25] (2021)**	**Xiong et al. [27] (2023)**
STEA	3D ResNet18	3D ResNet50	S3D8	S3D
AEBA	3D ResNet18	1D CNN	3D ResNet18	VGGish
VPTM	Kinetics	Kinetics	Kinetics	Kinetics
APTM	Kinetics	SoundNet	SoundNet	AudioSet
Dataset	AVE	DIEM Coutrot1 Coutrot2 AVAD SumMe ETMD	DHF1K Hollywood-2 UCF-sports DIEM Coutrot1 Coutrot2 AVAD SumMe ETMD	DIEM Coutrot1 Coutrot2 AVAD SumMe ETMD
Imp	easy	moderate	easy	easy
Att.	✗	✓	✗	✓
Comx	✗	✓	✓	✓

The synthesis of current studies in AVSP encompass aspects such as performance accuracy, efficiency, scalability, computational requirements, computational methods, temporal consistency, and semantic correlation. It highlights the critical importance of integrating audio and visual cues to enhance saliency prediction, especially in complex environments shaped by both sensory inputs. Models like STAViS and CASP-Net, which incorporate attention mechanisms, show significant improvements in maintaining spatial and temporal coherence. These mechanisms dynamically adjust focus based on each modality's importance, addressing inconsistencies and enhancing semantic correlation between audio and visual features (e.g., matching spoken words with lip movements). Cross-modal attention further improves performance by selectively focusing on relevant features, reducing noise, and maintaining spatial and temporal precision, although this increases implementation complexity. Furthermore, model performance varies with dataset complexity: While models such as ViNet[14] excel on simpler datasets, their advantage diminishes in complex environments such as DIEM or AVAD, where temporal and semantic mismatches are common. This

suggests that future research should focus on developing models robust enough to handle complex audiovisual environments, multiple sound sources, and asynchronous audio-visual streams.

Table 2.5: Performance Comparison Across Six Datasets

Dataset	**AVAD**[29]					**DIEM**[95]					**SumMe**[96]				
Methods	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS
STAViS[26]	0.9196	0.4578	**0.5936**	0.6086	3.18	0.8838	0.4824	**0.6741**	0.5795	2.26	0.8883	0.3373	**0.6562**	0.4220	2.04
AviNet[14]	0.9050	0.4460	0.5600	0.5800	3.17	0.8690	0.4270	0.6220	0.5220	2.02	0.8680	0.2960	0.6090	0.3790	1.79
STANet[94]	0.8730	0.3340	0.5800	0.4380	2.02	0.8610	0.3910	0.6580	0.4690	1.72	0.8540	0.2940	0.6270	0.3680	1.65
CASP-Net[27]	0.9320	0.5280	-	0.6850	3.77	0.9040	0.5360	-	0.6490	2.58	0.9040	0.3770	-	0.4860	2.52
DiffSal[28]	**0.9350**	**0.5710**	**0.6200**	**0.7380**	**4.22**	**0.9060**	**0.5430**	**0.6250**	**0.6600**	**2.65**	**0.9210**	**0.4470**	**0.5000**	**0.7290**	**3.14**

Dataset	**ETMD**[97]					**Coutrot1**[98]					**Coutrot2**[99]				
Methods	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS
STAViS[26]	0.9316	0.4251	**0.7317**	0.5690	2.94	0.8686	0.3935	**0.5847**	0.4722	2.11	0.9581	0.5111	**0.7106**	0.7349	5.28
AviNet[14]	0.9150	0.3290	0.6750	0.4770	2.36	0.8500	0.3610	0.5420	0.4250	1.92	0.9260	0.3220	0.5940	0.4480	3.16
STANet[94]	0.9080	0.3180	0.6820	0.4480	2.18	0.8290	0.3060	0.5420	0.3390	1.37	0.8500	0.2470	0.5970	0.2730	1.48
CASP-Net[27]	0.9390	0.4760	-	0.6160	3.31	0.8870	0.4530	-	0.5600	2.66	0.9630	0.5730	-	0.7660	6.11
DiffSal[28]	**0.9430**	**0.5060**	**0.5720**	**0.8330**	**3.66**	**0.9010**	**0.5150**	**0.5150**	**0.6380**	**3.20**	**0.9640**	**0.6250**	**0.6250**	**0.8350**	**6.61**

The table 2.5 shows the evaluated AVSP methods using standard metrics such as AUC-Judd, SIM, s-AUC, CC, and NSS, across multiple datasets like AVAD, DIEM, and Coutrot. The analysis shows that attention-based models like STAViS and CASP-Net outperform simpler concatenation-based models like DAVE in terms of spatial and temporal consistency. However, the higher computational demands of these models may make them less efficient. Notably, DiffSal consistently achieves top scores across nearly all benchmarks, surpassing prior models in both accuracy and robustness, especially in CC and NSS, which are critical for measuring saliency quality. Its generative diffusion-based framework effectively captures multimodal dependencies, enabling refined, coherent saliency maps. Despite its larger parameter size compared to some lightweight baselines, DiffSal offers a strong trade-off between performance and complexity, making it one of the most reliable AVSP solutions in the table. Although these diffusion-based functions work better compared to the others , its metrics values like the CC, NSS and SIM are still under weighted. While Diffsal achieves audiovisual synergy in some cases, it shows a dominance of the visual modality, with limited contribution from audio.

In this work, we consider several key computational properties that are desirable for effective multimodal audio-visual saliency prediction as visualized in the Table 2.6.

- **Spatial Retention:** The ability of the model to preserve fine-grained spatial information throughout the network, which is essential for accurately localizing salient regions without distortion or resolution loss during processing.
- **Semantic Relevance Modeling:** Adapting feature integration by aligning semantic content from audio and visual modalities, ensuring that only semantically meaningful interactions influence the saliency output.
- **Temporal Synchronization:** Addressing temporal misalignment between audio and visual streams by synchronizing features across time, enabling temporally consistent saliency predictions in dynamic video sequences.
- **Bidirectional Cross Attention:** Enhancing feature interaction by allowing both audio-to-visual and visual-to-audio attentional pathways, which creates a more balanced and context-aware multimodal representation.
- **Generative Framework:** Employing a generative mechanism (e.g., diffusion models) to iteratively refine the saliency map, resulting in smoother, higher resolution and more perceptually coherent predictions.
- **Weak AV Correlation:** Maintaining robustness when audio and visual signals are loosely related or asynchronous, by preventing misleading contributions from poorly aligned features.
- **Audio Impact:** utilizing the saliency influence of audio in different contexts to avoid over or under representation, especially in environments with ambiguous sound cues.

- **Benchmark Superiority:** Achieving competitive or superior results across standard AVSP datasets and metrics, serving as an empirical validation of architectural and methodological contributions their previous SOTA comparison.

- **Adaptive Gated Fusion:** A dynamic fusion strategy that uses learned gating mechanisms to balance modality contributions at each feature scale, allowing the model to prioritize more informative cues in context.

Table 2.6: Final Comparison of Related Works in Audio-Visual Saliency Prediction Based on Core Architectural Properties.

Method	Spatial Retention	Semantic Relevance	Temporal Synchronization	Bidirectional Cross Attention	Generative Framework	Weak AV Correlation	Audio Impact	Benchmark Superiority	Adaptive Gated Fusion
DAVE [24]	X	X	◐	X	X	X	◐	◐	X
STAViS [26]	✓	◐	◐	X	X	X	◐	✓	X
AViNet [14]	X	X	◐	X	X	✓	X	✓	X
STANet [94]	X	X	✓	X	X	✓	X	◐	X
CASP-Net [27]	✓	◐	✓	X	X	✓	◐	✓	X
DiffSal [28]	✓	✓	✓	X	✓	✓	◐	✓	X

Legend: ✓ = Fully addressed, ◐ = Partially addressed, **X** = Not addressed.

Despite notable advances in Audio Visual Saliency Prediction (AVSP), current models such as CASP-Net[27] and DiffSal[28] still fall short in several key areas. Most struggle to generalize across complex scenes, particularly under weak or asynchronous audio-visual correlation. Additionally, prior works tend to rely on shallow fusion techniques with static concatenation or one-way attention, which fail to model bidirectional dependencies or dynamically adjust the influence of each modality. These limitations are evident in the comparative analysis shown in Table 2.6, where most methods lack bidirectional attention, adaptive fusion, and generative refinement capabilities.

To address these gaps, our work builds up on these insights by proposing a unified multimodal architecture shortly MUBiC that combines several innovative components. A multiscale Bidirectional Efficient Cross-Attention mechanism enables dynamic alignment between audio and visual streams in both directions,

enhancing semantic fusion across time. An adaptive gated fusion module selectively modulates modality influence based on context, mitigating dominant or misleading cues. Finally, a conditional diffusion-based decoder iteratively denoises saliency maps with high spatial fidelity, reinforcing temporal coherence as depicted in chapter 3. Unlike prior approaches that either overfit to strong modality alignment or rely on shallow fusion, we demonstrates robustness, and temporal coherence setting a new standard for generalizable and interpretable audio-visual saliency prediction as shown in chapter 4.

Chapter 3

Methodology

The primary aim of this work is to present a unified multimodal architectural method that fuses visual and auditory features to improve audio-visual saliency prediction. The proposed method enables robust cross-modal integration and addresses temporal misalignment between audio and visual signals. Unlike existing 3D convolution-based approaches, MUBiC introduces bidirectional inter-modal efficient cross attention to facilitate reciprocal feature exchange. A temporal-guided noise encoder is integrated into a conditional diffusion framework, allowing iterative refinement of spatiotemporal saliency maps. This chapter outlines the methodology, including the research process, architectural components, training and inference strategies, and evaluation metrics.

3.1 Research Methodology

To guide our research, we adopted the Design Science Research Process (DSRP) [100], a widely used framework in information systems and computer science for developing innovative and practical artifacts. DSRP's structured, iterative approach aligns well with our objective: designing a novel audio-visual saliency prediction model that addresses temporal alignment and fusion challenges. The key steps followed are:

Problem Identification: We conducted a systematic literature review focusing on deep learning-based audio-visual saliency models. This allowed us to identify core limitations such as weak cross-modal alignment, static fusion strategies, and limited handling of temporal correlation, which are central to the

problem outlined in Section 1.3.

Objective Formulation: Based on the identified gap, we formulated research questions and specific objectives section 1.5 to guide the design and evaluation of our solution.

Design and Development: We developed a dual-stream encoder architecture integrated with Bidirectional Efficient Cross Attention (BECA) and a gated fusion module. To address temporal misalignment, we incorporated a noise-aware encoder within a conditional diffusion framework, enabling progressive refinement of saliency predictions. Section 3.2.3 provides the detailed model design.

Experimentation: We conducted extensive ablation studies and training strategy tests to evaluate performance across various architectural variants and regularization techniques, as detailed in Chapter 4.

Evaluation: Our model was evaluated using quantitative saliency metrics (CC, SIM, NSS) and human evaluation on AVAD. Comparisons with SOTA models validate its effectiveness under both constrained and extended training setups.

3.2 Design and development

This phase follows standard deep learning training procedures, customized for our model. The process involves five main stages: Data Acquisition, Preprocessing, Model Building, Evaluation, and Deployment, though deployment is not addressed in this thesis due to scope limitations. Subsection 3.2.1 covers data acquisition, including source identification and annotation. Subsection 3.2.2 details preprocessing: visual data are extracted from video frames, resized, and normalized while audio data are converted to waveforms or spectrograms, normalized, and resampled. The modeling subsection 3.2.3 is encoded using Bidirectional Efficient Cross Attention and fused into a unified representation. Also the model then predicts saliency maps by decoding noise corrupted ground truth, conditioned on the fused features.

3.2.1 Data Acquisition

A. Identifying relevant data sources

It is known that data plays a crucial role, particularly in the field of AI. Separately there are plenty of visual saliency datasets like DHF1k [101]. Currently, our work requires datasets that incorporate both audio and visual elements. We explored the most common publicly available eye fixation datasets collected in the audio-visual environment.

The **AVAD** [29] is the dataset consists of both visual and audio, which conducted an experiment using a Tobii T120 eye tracker with 30 videos sourced from YouTube and other platforms. They prepared 45 video clips from the test video which contains musical instruments, ball play or kicking, speaking faces, and conversations, each lasting between 5 to 10 seconds. This experiment involved 26 participants. The soundtracks and visual scenes were watched synchronously with an eye tracker positioned 60 cm away. The dataset was heavily skewed towards playing and conversations due to tracker issues.

Importantly, AVAD has been used as a benchmark in multiple state-of-the-art AVSP models, including STAViS, CASP-Net and DiffSal [26, 27, 28].This wide adoption underscores the relevance and representativeness of AVAD for modeling and evaluating cross-modal saliency in dynamic environments.

B. Saliency map annotation

The annotation of the data implies labeling specific parts of the dataset with relevant information. In our case, it involves tagging video frames with saliency fixation maps generated from eye-tracking procedures. These maps indicate the most visually important regions of each frame and serve as ground truth for training models to predict saliency in unseen data.

We obtained these annotations from the AVAD dataset, which provides 30

frames per second. This ground truth is crucial for the later stages of our architecture, which follows a supervised model-based approach. The annotated saliency maps guide the training process, ensuring the model learns to focus on the most salient regions within the visual data.

3.2.2 Data preprocessing

Acquiring data alone is not sufficient for our model; the data must be thoroughly prepared for effective training. Since our data is multimodal, we need to preprocess the audio and visual data separately to ensure they are properly formatted and aligned for the model.

A. Audio data preprocessing

Audio signals, represented as continuous waveforms, are digitized using sampling (e.g., 16 kHz) into formats like WAV or MP3. A waveform offers time-domain representation shown in figure 3.1, but to capture both time and frequency information, we apply the Short-Time Fourier Transform (STFT), producing spectrograms as seen in figure 3.2 and mel-spectrograms represented in figure 3.3. Spectrograms use linear frequency scaling, while mel-spectrograms use a perceptual mel scale aligned with human hearing. These are widely used in speech and music analysis.

After converting raw audio into mel-spectrograms, we normalize values to [0, 1], resample to fit model input dimensions, segment into time aligned frames. These steps prepare audio features for synchronization with visual data.

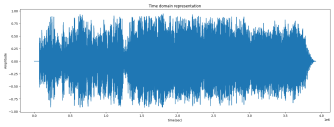


Figure 3.1: Audio Waveform Representation.

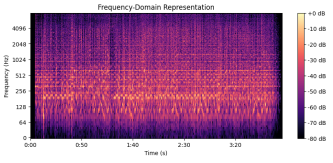


Figure 3.2: Frequency-Time Domain Representation(Spectrogram)

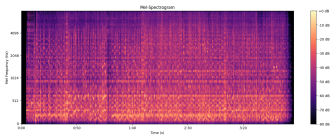


Figure 3.3: Mel-Spectrogram Representation.

B. Visul data preprocessing

Video data are first decomposed into sequential frames. For example, the AVAD dataset provides 30 frames per second (fps), and a 10-second clip yields 300 frames. We extract frames using the formula [102]:

$$\text{Total Frames} = T \times F \tag{3.1}$$

where T is duration and F is frame rate (e.g., 30 fps). A fixed sample duration of 16 frames is used for model input. Frame sampling is controlled via step size, determining the interval between frames and successive samples.

Each frame is then normalized, and resized while preserving aspect ratio. Cropping is applied as needed to focus on regions of interest. The following figures 3.4, and 3.5 show the extracted frame with its corresponding saliency maps.

Figure 3.4: Video Frame

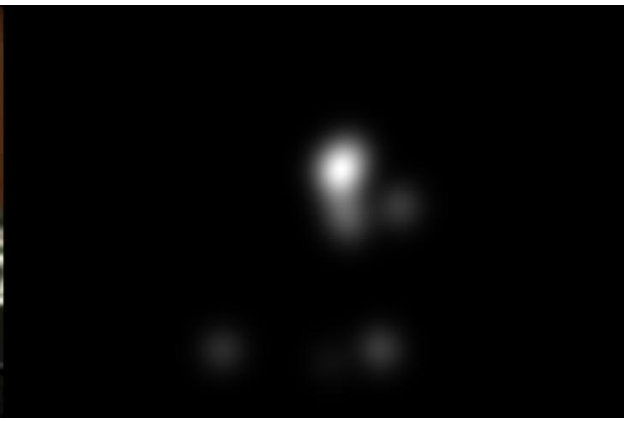


Figure 3.5: Eye Tracking Map

C. Alignment of audio and visual data

Aligning video frames with the corresponding audio segments is vital for feature extraction and model training. This alignment ensures that every video frame is paired with the appropriate audio segment, which is indispensable for our task, where both visual and auditory cues influence the results. To capture audio

samples for each video frame, one should generate ‘starts‘ and ‘ends‘ arrays which will hold the beginning and ending indices of the audio sample for each frame.

$$\text{start} = \max\left(0, \left(\frac{\text{videoframe} - 1}{\text{video\_fps}}\right) \times \text{Fs} - \frac{2}{\text{n\_samples}}\right) \tag{3.2}$$

$$\text{end} = \min\left(\text{total\_audio\_samples}, \left(\frac{\text{videoframe} - 1}{\text{video\_fps}}\right) \times \text{Fs} + \frac{\text{n\_samples}}{2}\right) \tag{3.3}$$

Where, videoframe=index of the current video frame, video_fps= frame rate per second, and Fs= audio sampling frequency in Hz. The variable n_samples represents the number of audio samples corresponding to a single video frame, which can be calculated as:

$$\text{n\_samples} = \frac{\text{Fs}}{\text{video\_fps}} \tag{3.4}$$

We derived this equation based on the concept in the audio and visual synchronization work[102]. Upon determining the start and end arrays for all frames, the audio segment corresponding to each video frame may subsequently be extracted.

3.2.3 Modeling

The proposed architecture consists of four core modules: (1) an Audio-Visual Encoder, which extracts hierarchical features using MViTv2 for video frames and VGGish with temporal refinement for audio signals; (2) a Cross-Modal Fusion Module composed of four stacked Multimodal Interaction (MMI) blocks that apply bidirectional efficient cross-attention and gated fusion to integrate features across modalities and scales; (3) a Diffusion Noise Encoder, where the ground-truth saliency map is corrupted and encoded via a 4-stage 3D ResNet guided by timestep embeddings; and (4) a Diffusion Decoder that refines noisy feature maps through four Conditional Denoising Modules (CDMs), each containing 3D temporal convolutions, cross-modal attention, and spatial upsampling, to generate high-resolution saliency predictions.

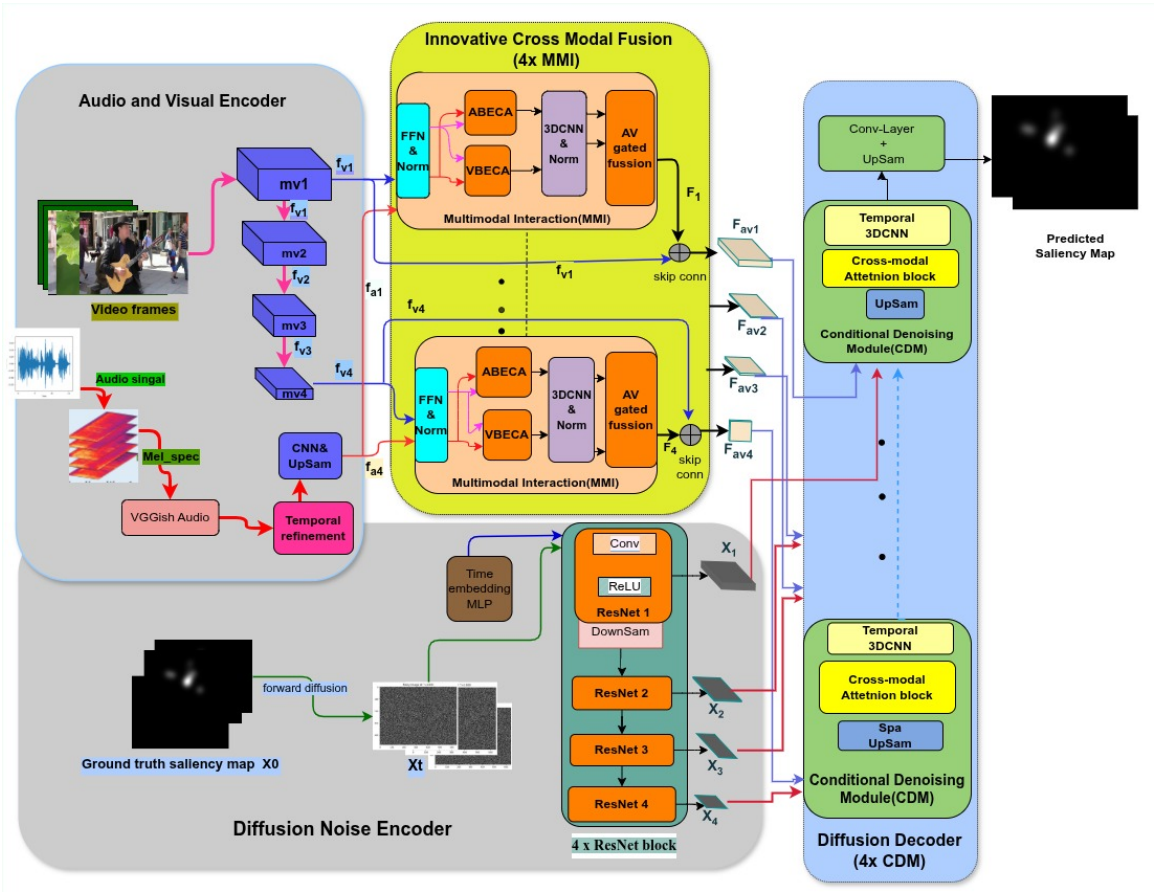


Figure 3.6: Proposed Audio-Visual Saliency Prediction Model

3.2.3.1 Audio and Visual encoder

1. **Audio encoder:** This architectural stream of the audio encoder is designed to derive an audio feature map. The 1D audio waveform must be transformed into a 2D log-mel spectrogram via the Short-Time Fourier Transform (STFT). Once the mel spectrogram is generated, it is segmented into portions of Sa slices. Each segmented portion has dimensions of Ha × Wa × 1 with a hop size of 21 ms. To extract the per-frame audio feature $f_{a,i}$, where $i \in \{1, \cdots, Ta\}$, a fully convolutional pre-trained 2D VGGish network trained on AudioSet is utilized, producing a feature map with dimensions $R_{ha} \times w_a \times C_a$. To ensure better inter-frame consistency, we introduce an additional temporal enhancement module.

   Temporal Refinement: designed to improve the consistency of audio fea-

tures across frames in multimodal settings. This module ensures the temporal coherence of audio signals, which is essential for capturing audiovisual interactions.

The audio features are rearranged into a spatio-temporal format denoted by $f_a \in \mathbb{R}^{T_a \times h_a \times w_a \times C_a}$, where T_a, h_a, w_a, and C_a represent the temporal length, height, width, and channel dimensions, respectively. A learnable positional embedding e_{pos} is added to preserve frame ordering and enhance temporal understanding.

Next, multi-head self-attention is applied along the temporal axis to capture dependencies between frames, inspired by the Transformer framework [56]. This is followed by a feedforward network (FFN) to model nonlinear patterns. Residual connections [38] are employed after both attention and FFN to facilitate stable training [57]:

$$\text{Attention Output} = \text{MultiHeadAttention}(\bar{F}_a) \tag{3.5}$$

$$x_1 = \text{LayerNorm}(f_a + \text{Attention Output}) \tag{3.6}$$

$$x_2 = \text{LayerNorm}(x_1 + \text{FFN}(x_1)) \tag{3.7}$$

$$f_a = x_2 \tag{3.8}$$

After refinement, the final representation becomes $f_a \in \mathbb{R}^{T_{\text{ai}} \times D_a}$, where T_{ai} is aligned with the visual stream, and D_a is the feature dimensionality. These enhanced features are passed to the fusion and decoder modules and also serve as skip connections.

2. **Visual encoder:** To compute a spatiotemporal feature map, we used the four consecutive stages of MViTv2 [103] represented as $mv1$, $mv2$, $mv3$, and $mv4$, which produce feature maps $fv1$, $fv2$, $fv3$, and $fv4$, respectively. It

is chosen as our visual encoder backbone due to its exceptional video recognition capabilities. MViTv2's multiscale architecture captures features at varying resolutions, offering both fine-grained and global representations. Its pooling attention mechanism aggregates information more efficiently than standard local attention, improving performance and managing compute. The use of decomposed relative positional embeddings [103] allows the Transformer blocks to better capture spatial relationships. Residual pooling connections mitigate the effect of stride in attention computation, improving stability and accuracy. MViTv2 leverages both CNN and Transformer design principles and achieves 86.1% top-1 accuracy on Kinetics-400, making it a robust and efficient backbone for video based tasks.

Input Layer: Takes a sequence of video frames with dimensions $T \times H \times W \times 3$, where T is the number of frames. Frames are sampled (e.g., 16), normalized, and resized.

Patch Embedding: Feature maps are flattened into patches. A linear transformation reduces patch token dimensionality to prepare them for attention via query (Q), key (K), and value (V) generation.

Multi-Head Pooling Attention (MHPA): Reduces computational cost by pooling token sequences before attention. This prioritizes salient features and uses relative position encoding [103] and residual pooling connections.

Decomposed Relative Position Embedding: Improves spatiotemporal modeling by injecting relative position information directly into attention scores (not the token embedding, unlike ViT [57]):

$$\begin{aligned} \text{Attn}(Q, K, V) &= \text{Softmax}\left(\frac{QK^\top + E_{\text{rel}}}{\sqrt{d}}\right) V, \\ \text{where} \quad E_{\text{rel}}^{ij} &= Q_i \cdot R_{p(i),p(j)} \end{aligned} \tag{3.9}$$

To reduce complexity, the relative embedding is decomposed into height, width, and time:

$$R_{p(i),p(j)} = R_h(h(i), h(j)) + R_w(w(i), w(j)) + R_t(t(i), t(j)) \tag{3.10}$$

Residual Pooling Connection: Enhances gradient flow by adding the pooled query vector to the attention output:

$$Z := \text{Attn}(Q, K, V) + Q \tag{3.11}$$

We apply temporal attention across frames to model long-range dependencies. A global average pooling operation summarizes the spatial dimension, producing a compact feature representation. Consider:

$$I = [I_1, \cdots, I_{T_v}], \quad I_j \in \mathbb{R}^{H_v \times W_v \times 3} \tag{3.12}$$

This is processed through a video backbone with 4 stages, outputting hierarchical features (see Table 3.1).

Stages	**Operators**	**Output Sizes**
Data Layer	Stride $\tau \times 1 \times 1$	$D \times T \times H \times W$
Cube_1	$c_T \times c_H \times c_W, D\text{Stride } s_T \times 4 \times 4$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
Scale_2	$\text{MHPA}(D)\text{MLP}(4D) \times N_2$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
Scale_3	$\text{MHPA}(2D)\text{MLP}(8D) \times N_3$	$2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$
Scale_4	$\text{MHPA}(4D)\text{MLP}(16D) \times N_4$	$4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
Scale_5	$\text{MHPA}(8D)\text{MLP}(32D) \times N_5$	$8D \times \frac{T}{s_T} \times \frac{H}{32} \times \frac{W}{32}$

Table 3.1: Multiscale Vision Transformer Stages and Output Sizes

The output features from each stage are:

$$\begin{aligned} \{fv_i\}_i^N &\in \mathbb{R}^{T_{vi} \times h_{vi} \times w_{vi} \times C_{vi}}, \\ (h_{vi}, w_{vi}) &= \frac{(H_v, W_v)}{2^{i+1}}, \quad N = 4 \end{aligned} \tag{3.13}$$

Flattening along spatial dimensions yields:

$$f_v \in \mathbb{R}^{T_{vi} \times D_v}, \quad D_v = h_{vN} \times w_{vN} \times C_{vN} \tag{3.14}$$

Where T_{vi} is the number of frames, and h_{vN}, w_{vN}, C_{vN} are the final stage height, width, and channel count. This visual embedding is passed to the FFN and used for multimodal interaction.

3.2.3.2 Diffusion Noisy Encoder:

The noisy saliency map X_t is generated via a forward diffusion process, where Gaussian noise is gradually added to the original ground-truth saliency map X_0 over T time steps. This process is modeled as a Markov chain, formally defined by Ho et al. [58] as:

$$q(X_t|X_0) = \mathcal{N}\left(X_t \mid \sqrt{\bar{\alpha}_t}X_0, (1 - \bar{\alpha}_t)\mathbf{I}\right), \tag{3.15}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_s = 1 - \beta_s$. The noise schedule β_s is set using a cosine schedule, as proposed in [104], defined by:

$$\alpha_t = \cos^2\left(\frac{\pi/2 \cdot (t/T + \gamma)}{1 + \gamma}\right), \tag{3.16}$$

where $\gamma = 0.008$ controls the steepness of the schedule. This schedule enables smooth noise addition across time steps, ensuring that $\bar{\alpha}_t \to 0$ as $t \to T$, resulting in a pure Gaussian noise sample $X_T \sim \mathcal{N}(0, \mathbf{I})$.

A single diffusion step at time t involves corrupting the original saliency map X_0 into a noisy version X_t, expressed as:

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{3.17}$$

where ϵ is standard Gaussian noise. Typical configurations set the total diffusion steps to $T = 1000$, with $t \in \{0, 1, ..., T\}$, ranging from the clean image (X_0) to fully corrupted noise (X_T).

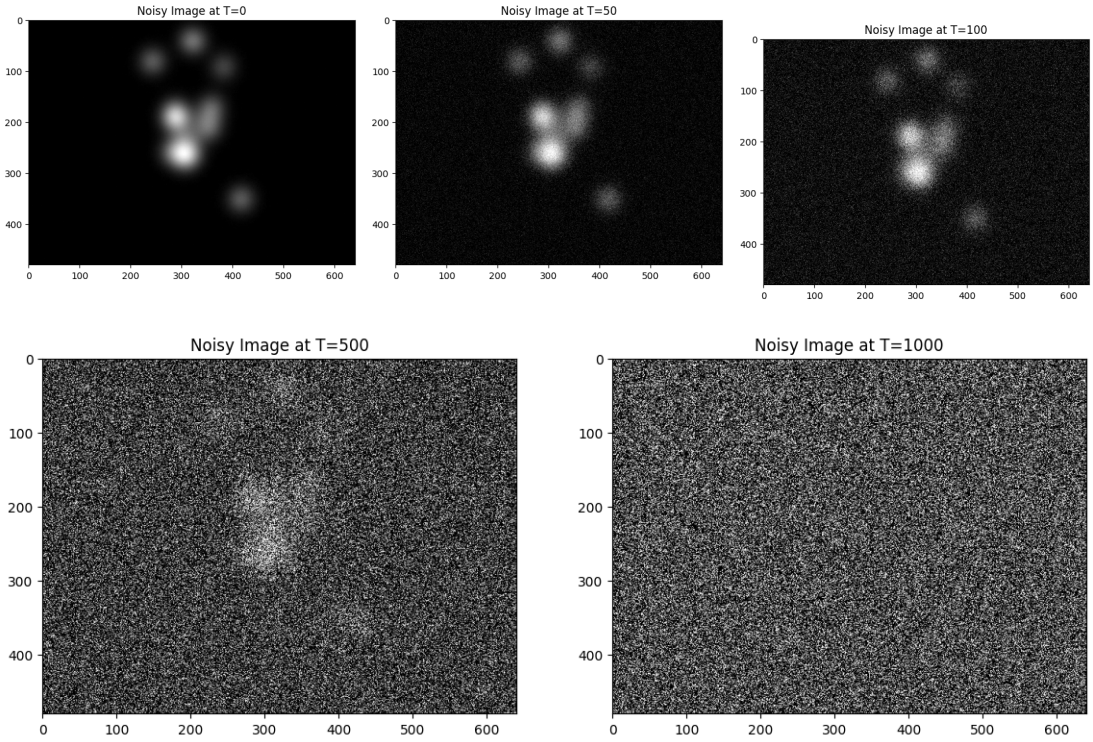


Figure 3.7: Noised GT Images With Variable Time Steps

Figure 3.7 illustrates this degradation process through five visualizations of the ground truth saliency map at selected time steps. At $T = 0$, the original saliency map is fully intact. By $T = 50$, a small amount of noise (5.7%) is noticeable, obscuring fine details. At $T = 100$, noise (15.4%) becomes prominent. At $T = 500$, roughly 70% of the image is corrupted, and by $T = 1000$, the input is indistinguishable from pure Gaussian noise. This visual degradation demonstrates the difficulty of the reverse process recovering structure from noise,

which is learned through the decoder.

The encoder then transforms the noisy saliency input X_t into hierarchical latent features. To accomplish this, we use four 2D ResNet blocks [38], each with increasing channel depth. The base convolution operation is given by:

$$\mathbf{X}_{\text{conv}} = \sum_{i=-1}^{1} \sum_{j=-1}^{1} W[i,j] \cdot \mathbf{X}_{\text{in}}[x+i, y+j] + b, \tag{3.18}$$

using a 3×3 kernel and 2×2 stride for downsampling. Batch Normalization [105] is applied to stabilize training:

$$\mathbf{X}_{\text{bn}} = \gamma_{\text{BN}} \odot \frac{\mathbf{f} - \mu}{\sigma + \epsilon} + \beta_{\text{BN}}. \tag{3.19}$$

To modulate the encoder behavior with respect to the diffusion timestep t, we apply time-step conditioning [106]. A learned embedding of the timestep is passed through a multilayer perceptron (MLP) to produce scale and shift parameters:

$$\gamma_t = \text{MLP}(\text{Embed}(t)), \tag{3.20}$$

$$\beta_t = \text{MLP}(\text{Embed}(t)), \tag{3.21}$$

$$\mathbf{X}_{\text{cond}} = \gamma_t \odot \mathbf{X}_{\text{in}} + \beta_t. \tag{3.22}$$

Each ResNet block also contains a residual connection to retain spatial information and refine it through learning residual corrections. This is computed as:

$$\mathbf{X}_{\text{out}} = \mathbf{X}_{\text{in}} + \mathcal{F}(\mathbf{X}_{\text{in}}), \tag{3.23}$$

where $\mathcal{F}$ represents a subnetwork composed of Conv2D, BatchNorm, and ReLU operations. The encoder contains four stages with increasing channel sizes. The input saliency map $S_t \in \mathbb{R}^{112 \times 192}$ is downsampled progressively: Stage 1: Output $\mathbf{X}_s^1 \in \mathbb{R}^{56 \times 96 \times 64}$, Stage 2: Output $\mathbf{X}_s^2 \in \mathbb{R}^{28 \times 48 \times 128}$, Stage 3: Output $\mathbf{X}_s^3 \in$

$\mathbb{R}^{14\times24\times256}$ and Stage 4: Output $\mathbf{X}_s^4 \in \mathbb{R}^{7\times12\times512}$ having kernel size of 3x3, and stride of 2x2 for each stage.

The hierarchical features $\{\mathbf{X}_s^i\}_{i=1}^4$ are then passed to the decoder. They are used in conjunction with aligned visual and audio features through multimodal attention and diffusion-based denoising to reconstruct a clean saliency map.

3.2.3.3 Innovative Cross-Modal Fusion Module

To capture rich audio-visual interactions, we introduce a novel Innovative Cross-Modal Fusion module composed of four Multimodal Interaction Blocks (MMI). Each MMI integrates bidirectional attention (ABECA and VBECA), 3D spatiotemporal convolution, and gated fusion to enable fine-grained cross-modal reasoning. This design ensures synchronized and adaptive fusion across modalities, directly addressing challenges in weak audio-visual correlation and temporal misalignment.

A. Mutlimodal interaction(MMI)

1. **Matching Temporal Dimensions**

   To enable synchronized interaction between the audio and visual modalities, their temporal dimensions must align. We achieve this by upsampling the audio feature matrix F_a using transposed convolution, also known as deconvolution [27]:

$$F_a' = \text{ConvTranspose}\Big(F_{a\_out}, \quad \text{stride} = T_a T_v,\ \text{kernel\_size} = k\Big) \tag{3.24}$$

   - k denotes the kernel size of the transposed convolution layer.

   This operation ensures that the upsampled audio feature F_a' temporally matches the visual stream F_v, enabling effective multimodal attention.

2. **Separate Feed Forward Network (FFN)**

   To retain modality-specific structure and prevent early fusion loss, we apply FFNs separately to visual and audio streams. For the visual stream, two FFN layers are followed by a residual connection and layer normalization [107]:

$$F_v^1 = \text{FFN}_{v1}(F_v), \quad F_v^2 = \text{FFN}_{v2}(F_v^1), \quad F_{v\_out} = \text{LayerNorm}(F_v^2 + F_v) \tag{3.25}$$

   For the audio stream, the FFN transformation and residual normalization are similarly defined:

$$F_a^1 = \text{FFN}_a(F_a), \quad F_{a\_out} = \text{LayerNorm}(F_a^1 + F_a) \tag{3.26}$$

   The resulting $f_{v\_out} \in \mathbb{R}^{T \times D_v}$ and $f_{a\_out} \in \mathbb{R}^{T \times D_a}$ — temporally aligned and dimensionally enriched are passed into the transformer's multi-head attention block. ReLU is used within the FFNs to introduce non-linearity and improve learning capacity.

3. **Bidirectional Efficient Cross Attention (BECA)** In this section, we introduce a bidirectional efficient cross-attention(BECA) audio visual mechanism designed to implicitly capture the temporal correspondence between two modalities by reducing its diamension while preserving both local and global spatial information, thus improving saliency prediction performance. This mechanism operates by allowing visual features from one modality to concentrate on and incorporate data from the audio characteristics of another modality, or vice versa. Specifically, features of one modality are converted into queries and values, while features of the other modality be-

come keys. The queries are used to identify pertinent segments of the other modality by calculating the attention scores using the keys and then integrating this information through the values.

Before feeding the computed features into the multihead attention, a learnable temporal embedding is introduced to encode their position and order. TE_a and TE_v serve as learnable temporal embeddings for audio and visual stream features respectively.

$$F_a = F'_a + TE_a, \text{ and } F_v = F_{v\_out} + TE_v \tag{3.27}$$

Visual Bidirectional Efficient Cross Attention(VBECA) (Q from Visual, where as K and V from Audio): The attention mechanism for visual features moderated by spatio temporal compressed(STC) [108] audio features is described by:,

$$\mathbf{Q}_v = \mathbf{F}_v \mathbf{W}_Q \tag{3.28}$$

$$\mathbf{K}_a = \text{STC}(\mathbf{F}_a)\mathbf{W}_K \tag{3.29}$$

$$\mathbf{V}_a = \text{STC}(\mathbf{F}_a)\mathbf{W}_V \tag{3.30}$$

$$\text{Attention} = \text{softmax}\left(\frac{\mathbf{Q}_v \mathbf{K}_a^\top}{\sqrt{d_k}}\right)\mathbf{V}_a \tag{3.31}$$

where:

Q_v: Query matrix derived from visual features.

K_a : Key matrix derived from audio features.

V_a : Value matrix derived from visual features.

W_Q, W_K, W_v : Learnable projection matrices.

This structure adjusts the visual feature set V_v by incorporating attention

scores that emerge from the cross-modal interaction between visual Query Q_v and audio Key K_a, thereby integrating audio-informed attention into the visual processing stream.

Audio Bidirectional Efficient Cross Attention(ABECA) (Q and V from Audio, K from Visual): In a similar fashion, the attention score for audio features, leveraging STC [108] visual features for reference, is given by:

$$\mathbf{Q}_a = \mathbf{F}_a \mathbf{W}_Q \tag{3.32}$$

$$\mathbf{K}_v = \text{STC}(\mathbf{F}_v)\mathbf{W}_K \tag{3.33}$$

$$\mathbf{V}_v = \text{STC}(\mathbf{F}_v)\mathbf{W}_V \tag{3.34}$$

$$\text{Attention} = \text{softmax}\left(\frac{\mathbf{Q}_a \mathbf{K}_v^\top}{\sqrt{d_k}}\right)\mathbf{V}_v \tag{3.35}$$

where:

Q_a : Query matrix obtained from audio features.

K_v : Key matrix derived from visual features.

V_v : Value matrix computed from audio features.

This setup enables the audio stream to focus on its own values V_a, while being modulated by the Key matrix from visual features K_v, thereby incorporating visual context into the audio representation.

3DCNN + Norm: convolutions are employed for each streams to grasp local context. This generally involves depthwise separable convolutions for improved efficiency. This approach helps to capture detailed local features, thus complementing the global attention provided by the Transformer.

$$Y_v = \text{DepthwiseConv}\left(\text{GLU}\left(F_v^{(i)}\right)\right) + F_v^{(i)} \tag{3.36}$$

$$Y_a = \text{DepthwiseConv}\left(\text{GLU}\left(F_a^{(i)}\right)\right) + F_a^{(i)} \tag{3.37}$$

Additionally, another feedforward module with layer normalization is applied after the convolution layer. This is the second instance of a feedforward module, and again it includes residual connections.

4. **Gated AV Cross Attention Fusion(AV fusion)** After applying the multimodal cross-semantic interaction architecture separately to both audio and visual streams, we employ a Gated Fusion mechanism, ensuring an adaptive combination of both modalities while regulating their importance dynamically.

   To Compute Gating Weights a gating function, modeled by a learnable neural network (MLP or Conv1x1), generates an adaptive weighting factor $\boldsymbol{\alpha}$ for each fusion pathway:

$$\alpha = \sigma\left(\text{MLP}(\boldsymbol{F_{va}}, \boldsymbol{F_{av}})\right) \tag{3.38}$$

   where:

   - $\sigma(x) = \frac{1}{1+e^{-x}}$ is the **sigmoid activation function**, ensuring that $\alpha \in [0, 1]$.
   - $\text{MLP}(\boldsymbol{F_{va}}, \boldsymbol{F_{av}})$ learns nonlinear interactions between bidirectional attention outputs.
   - α adjusts the relative contribution of $\boldsymbol{F_{va}}$ and $\boldsymbol{F_{av}}$ in the fusion process.

   Then perform gated fusion of bidirectional outputs using the learned gating weight $\boldsymbol{\alpha}$, the final gated fusion representation is computed:

$$\boldsymbol{F}_{\text{fused}} = \alpha \cdot \boldsymbol{F_{va}} + (1 - \alpha) \cdot \boldsymbol{F_{av}} \tag{3.39}$$

where:

- If $\alpha \approx 1$, fusion is more based on visual-to-audio features ($\boldsymbol{F_{va}}$), improving the audio modality.
- If $\alpha \approx 0$, the fusion emphasizes audio-to-visual features ($\boldsymbol{F_{av}}$), refining visual saliency.

This adaptive weighting ensures robust multimodal fusion based on learned feature dependencies.

Skip Connection: are used to preserve critical spatial details from the visual encoder that might be lost during cross-modal fusion.

Downsample visual features f_{vi} to match F_{fused}'s spatial dimensions:

$$f_{vi,\text{down}} = \text{Conv3D}\left(f_{vi}; \text{kernel} = 2^i, \text{stride} = 2^i\right) \tag{3.40}$$

Residual addition with learnable scalar α:

$$F_{avi} = F_{fused} + \beta \cdot f_{vi,\text{down}}, \quad \beta \in [0, 1] \tag{3.41}$$

3.2.3.4 Diffusion Decoder Architecture

The decoder consists of four conditional denoising modules (CDM) that refine a noised saliency map into a predicted saliency map, leveraging early fused audiovisual features and temporal guided noise encoding. It operates top-down across four stages (deepest to shallowest), integrating noise features with fused audiovisual features through upsampling, gated fusion, and spatio-temporal cross-attention. Components of each CMD include:

A. Input Features

- **Noise Features** (f_{s_i}): Extracted from a temporal-guided 3D ResNet noise encoder processing a noised saliency map S_t at diffusion time step t.

- **Fused Audiovisual Features** (f_{out_i}): Generated by the multiscale cross-modal fusion module using bidirectional cross-attention and gated fusion.

B. Decoder Stages The decoder operates across four stages ((i = 4, 3, 2, 1), refining the saliency map hierarchically. For each stage i (from $i = 4$ to $i = 1$), the decoder performs the following steps:

Compute Fused Context ($f_{\mathbf{con}_i}$)

- **For the Deepest Stage ($i = 4$)**:

$$f_{\text{con}_4} = f_{\text{out}_4} \tag{3.42}$$

At the coarsest scale, the fused context is directly the fused audiovisual feature, providing initial multimodal guidance for denoising.

- **For Other Stages ($i < 4$)**:
  - **Stage 1: Spatial Upsampling**: To align the spatial resolution of the previous stage's output $f_{\text{dec}_{i+1}}$ with f_{out_i}.

$$f^{\text{up}}_{\text{dec}_{i+1}} = \text{Upsample}(f_{\text{dec}_{i+1}}) \tag{3.43}$$

transposed convolution (ConvTranspose3D), learns task-specific patterns.

  - **Stage 2: Gated Fusion**: Adaptively combine f_{out_i} and $f^{\text{up}}_{\text{dec}_{i+1}}$.

$$A_i = \text{Conv1x1}(f_{\text{out}_i}), \quad B_i = \text{Conv1x1}(f^{\text{up}}_{\text{dec}_{i+1}}) \tag{3.44}$$

$$\text{gate}_i = \sigma(\text{Conv1x1}(\text{concat}(A_i, B_i))) \tag{3.45}$$

$$f_{\text{con}_i} = \text{gate}_i \cdot A_i + (1 - \text{gate}_i) \cdot B_i \tag{3.46}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. A 1x1 convolution ad-

justs channel dimensions, enabling feature mixing. The gate, inspired by Gated Multimodal Units [30], dynamically weights contributions, enhancing multimodal integration.

Stage 3: Spatio-Temporal Cross-Attention To refine f_{s_i} by focusing on relevant parts of f_{con_i}, aligning noisy saliency patterns with audiovisual cues:

$$Q_i = f_{s_i} \cdot W_Q, \quad K_i = f_{\text{con}_i} \cdot W_K, \quad V_i = f_{\text{con}_i} \cdot W_V \tag{3.47}$$

$$\text{scores}_i = \frac{Q_i \cdot K_i^T}{\sqrt{d_k}} \tag{3.48}$$

$$\text{attn}_i = \text{softmax}(\text{scores}_i) \tag{3.49}$$

$$f_{\text{dec}_i} = \text{attn}_i \cdot V_i \tag{3.50}$$

where W_Q, W_K, W_V are learnable weight matrices, and d_k is the key dimension. The dot product computes similarity, scaled to stabilize softmax, as in transformer mechanisms. Cross-attention captures interactions, enhancing prediction accuracy.

Stage 4: Temporal 3D CNN To capture temporal dependencies in f_{dec_i} [39]:

$$f_{\text{dec}_i} = \text{Temporal3DCNN}(f_{\text{dec}_i}) \tag{3.51}$$

The kernel processes 3 frames temporally, ensuring the saliency map reflects dynamic changes.

Stage 5: Pass to Next Decoder Stage The output f_{dec_i} is passed to the stage $i-1$, ensuring progressive refinement from coarse to fine details.

Final Output After the final decoder stage, the output feature map is passed through a CNN layer to predict the final saliency map. After stage $i = 1$:

$$S_{\text{pred}} = \sigma\left(\text{Conv}\left(f_{\text{dec}_1}\right)\right), \quad \text{with out\_channels} = 1 \tag{3.52}$$

The Conv3D reduces to a single channel, and the sigmoid normalizes to $[0, 1]$, producing $S_{\text{pred}} \in \mathbb{R}^{B\times 1\times H\times W}$, a heatmap highlighting attention areas.

3.2.4 Saliency Loss

Our saliency model aims to predict attention maps that highlight the most visually important regions in an image. To train the model effectively, we measure the discrepancy between the predicted saliency maps $\hat{S}$ and the ground truth S using a weighted mean squared error (MSE) loss[109]. This loss encourages the model to generate accurate, well-localized saliency maps that closely match human-annotated ground truth.

The main loss is defined as

$$\text{Loss}_{\text{main}} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\hat{S}_{i,c,h,w} - S_{i,c,h,w} \right)^2, \tag{3.53}$$

where:

N**:** the number of samples (or patches) in the batch.

C**:** the number of channels (e.g. $C = 1$ for grayscale saliency maps).

H **and** W**:** the height and width of the saliency maps; here, the maps are resized to 224×224, yielding $224 \times 224 = 50\,176$ pixels.

$\hat{S}_{i,c,h,w}$ **and** $S_{i,c,h,w}$**:** the predicted saliency values and ground truth at the pixel position (c, h, w) for the ith sample.

λ**:** the loss weight scaling factor.

Example: For a single sample ($N = 1$) with one channel ($C = 1$) at a resolution of 224×224 pixels, assume that the average error per pixel is 0.05, so the squared error per pixel is $0.05^2 = 0.0025$. The summed loss per image is then

$$50\,176 \times 0.0025 \approx 125.44.$$

With $\lambda = 1$, the final $\text{Loss}_{\text{main}}$ for that sample is approximately 125.44. This example illustrates how small pixel-level errors can aggregate to a substantial loss when summed over all pixels, guiding the model to refine its saliency predictions.

3.2.5 Training Process

The **MUBiC** framework trains a diffusion-based model for audiovisual saliency prediction (AVSP) by reversing a noise-adding process. It processes video frames and audio signals to generate a unified multi-scale feature $\mathbf{f}_{av} = \{\mathbf{f}_{av,i}\}_{i=1}^{4}$ through bidirectional cross-attention and gated fusion, which conditions the diffusion decoder to denoise noisy saliency maps. The training process optimizes the diffusion decoder to predict ground-truth saliency maps from noisy inputs, leveraging a Denoising Diffusion Probabilistic Model (DDPM).

Denoising Diffusion Probabilistic Model (DDPM): DDPM is a generative framework that models data generation as a Markovian process [110]. It consists of:

- **Forward Process**: Gradually adds Gaussian noise to a ground-truth saliency map S_0 over T steps:

$$q(S_t|S_{t-1}) = \mathcal{N}\left(S_t \,|\, \sqrt{1-\beta_t}S_{t-1},\, \beta_t\mathbf{I}\right) \tag{3.54}$$

$$S_t = \sqrt{\hat{\alpha}_t}S_0 + \sqrt{1-\hat{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}) \tag{3.55}$$

  where $\hat{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, using a cosine noise schedule [104].

- **Reverse Process**: Learns to denoise from $S_T \sim \mathcal{N}(0,\mathbf{I})$ to S_0 using a neural network:

$$p_\psi(S_{t-1}|S_t) = \mathcal{N}\left(S_{t-1} \,|\, \mu_\psi(S_t,t),\, \Sigma_t\right) \tag{3.56}$$

  where μ_ψ is a learned denoising network and Σ_t is typically fixed [110].

In MUBiC, DDPM is conditioned on $\mathbf{f}_{av}$, enabling the diffusion decoder to generate saliency maps guided by audiovisual cues [28].

The diffusion decoder predicts the ground-truth saliency map S_0 from a noisy map S_t, conditioned on $\mathbf{f}_{av}$. The loss is the mean squared error (MSE):

$$\mathcal{L} = |S_0 - g_\psi(S_t, t, \mathbf{f}_{av})|^2 \tag{3.57}$$

Algorithm 1 MUBiC Training

Require: Video frames: $I = [I_1, \ldots, I_{T_v}]$, Audio: A, Total steps: T, Ground-truth maps: S_0

Initialize diffusion decoder parameters ψ, $\alpha_i = 1$, Conv weights
repeat
  $\{\mathbf{f}_v^i\}_{i=1}^4 \leftarrow \text{MViT}(I)$
  $\mathbf{f}_a \leftarrow \text{VGGishTransformer}(A)$
  for each scale $i = 1$ to 4 **do**
    $\mathbf{f}_a^i \leftarrow \text{TransConv3D}(\mathbf{f}_a, (T_v^i, h_v^i, w_v^i, C_a^i))$
    $\mathbf{f}_{vc}^i \leftarrow \text{ECA}(Q = \mathbf{f}_v^i W_Q, K = \text{STC}(\mathbf{f}_a^i) W_K, V = \text{STC}(\mathbf{f}_a^i) W_V)$
    $\mathbf{f}_{ac}^i \leftarrow \text{ECA}(Q = \mathbf{f}_a^i W_Q, K = \text{STC}(\mathbf{f}_v^i) W_K, V = \text{STC}(\mathbf{f}_v^i) W_V)$
    $\mathbf{x}_i \leftarrow \text{concat}(\mathbf{f}_{vc}^i, \mathbf{f}_{ac}^i)$
    $\text{gate}_i \leftarrow \sigma(\text{Conv}_{1\times1\times1}(\mathbf{x}_i))$
    $\mathbf{f}_{\text{fused}}^i \leftarrow \text{gate}_i \odot \mathbf{f}_{vc}^i + (1 - \text{gate}_i) \odot \mathbf{f}_{ac}^i$
    $\mathbf{f}_{av,i} \leftarrow \mathbf{f}_{\text{fused}}^i + \alpha_i \cdot \text{Conv}_{1\times1\times1}(\mathbf{f}_v^i)$
  end for
  $\mathbf{f}_{av} \leftarrow \{\mathbf{f}_{av,i}\}_{i=1}^4$
  Sample $t \sim \text{Uniform}(1, \ldots, T)$
  Compute $\hat{\alpha}_t \leftarrow \prod_{s=1}^t (1 - \beta_s)$
  Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $S_t \leftarrow \sqrt{\hat{\alpha}_t} S_0 + \sqrt{1 - \hat{\alpha}_t}\epsilon$
  $\{\mathbf{X}_i\}_{i=1}^4 \leftarrow \text{ResNetEncode}(S_t)$
  for each stage $i = 4$ to 1 **do**
    if $i < 4$ **then**
      $\mathbf{f}_{\text{up},i+1} \leftarrow \text{Upsample}(\mathbf{f}_{\text{dec}(i+1)})$
    else
      $\mathbf{f}_{\text{up},i+1} \leftarrow \mathbf{0}$
    end if
    $\mathbf{f}_{\text{con}i} \leftarrow \sigma(W_g \cdot (\mathbf{f}_{av,i} + \mathbf{f}_{\text{up},i+1})) \odot \mathbf{f}_{av,i} + (1 - \sigma(W_g \cdot (\mathbf{f}_{av,i} + \mathbf{f}_{\text{up},i+1}))) \odot \mathbf{f}_{\text{up},i+1}$
    $\mathbf{f}_{\text{dec}i} \leftarrow \text{Attention}(Q = \mathbf{X}_i W_Q, K = \mathbf{f}_{\text{con}i} W_K, V = \mathbf{f}_{\text{con}i} W_V)$
    $\mathbf{f}_{\text{dec}i} \leftarrow \text{Conv3D}(\mathbf{f}_{\text{dec}i})$
  end for
  $\bar{S}_0 \leftarrow \sigma(\text{Conv}_{3D}(\mathbf{f}_{\text{dec1}}))$
  $\mathcal{L} \leftarrow \|\bar{S}_0 - S_0\|^2$
  Update ψ, α_i, Conv weights via gradient descent on $\nabla\mathcal{L}$
until convergence

The training algorithm initializes diffusion decoder parameters ψ, residual scalars $\alpha_i = 1$, and convolution weights to prepare for gradient-based optimization, with MSE loss guiding α_i to balance visual feature contributions (Line 1). Visual features $\{f_v^i\}_1^4$ are extracted using MViTv2, and audio feature f_a is derived via VGGish+Transformer, followed by upsampling f_a to f_a^i using transposed convolution for scale alignment (Lines 3–6). Bidirectional Efficient Cross Attention (ECA) with Spatio-Temporal Compression (STC) computes f_{vc}^i and f_{ac}^i, followed by gated fusion and residual connection to produce fused features $f_{av,i}$ (Lines 7–12). The forward diffusion process samples noise ϵ and step t to corrupt ground-truth saliency maps S_0 into S_t, which are encoded into $\{X_i\}_1^4$ using ResNet (Lines 14–19). Decoder stages iteratively refine X_i through upsampling, gated fusion, standard attention, and 3D convolution, culminating in predicting $\bar{S}_0$ and updating parameters via MSE loss (Lines 20–32). This process ensures robust audiovisual saliency prediction by leveraging hierarchical feature fusion and denoising.

3.2.6 Inference Process

MUBiC's inference process generates a saliency map S_{pred} by iteratively denoising a random noisy map $S_T \sim \mathcal{N}(0, \mathbf{I})$, conditioned on $\mathbf{f}_{av}$, using the trained diffusion decoder. It employs the Denoising Diffusion Implicit Model (DDIM) with adaptive steps for efficiency [111].

DDPM and DDIM in Inference: DDPM's reverse process iteratively refines noise to approximate the data distribution [110]. DDIM accelerates this with a deterministic update:

$$S_{t-1} = \sqrt{\hat{\alpha}t-1}g\psi(S_t, t, \mathbf{f}av) + \sqrt{1-\hat{\alpha}t-1} \cdot \frac{S_t - \sqrt{\hat{\alpha}t}g\psi(S_t, t, \mathbf{f}_{av})}{\sqrt{1-\hat{\alpha}_t}} \tag{3.58}$$

MUBiC uses DDIM with an adaptive number of steps, predicted by an MLP

based on $\mathbf{f}_{av,1}$ entropy [28].

Algorithm 2 MUBiC Inference

Require: Video frames: I, Audio: A, Total steps: T
Ensure: Predicted saliency map: S_{pred}

$\{\mathbf{f}_v^i\}_{i=1}^4 \leftarrow \text{MViT}(I)$
$\mathbf{f}_a \leftarrow \text{VGGishTransformer}(A)$
for each scale $i = 1$ to 4 **do**
  $\mathbf{f}_a^i \leftarrow \text{TransConv3D}(\mathbf{f}_a, (T_v^i, h_v^i, w_v^i, C_a^i))$
  $\mathbf{f}_{vc}^i \leftarrow \text{ECA}(Q = \mathbf{f}_v^i W_Q, K = \text{STC}(\mathbf{f}_a^i) W_K, V = \text{STC}(\mathbf{f}_a^i) W_V)$
  $\mathbf{f}_{ac}^i \leftarrow \text{ECA}(Q = \mathbf{f}_a^i W_Q, K = \text{STC}(\mathbf{f}_v^i) W_K, V = \text{STC}(\mathbf{f}_v^i) W_V)$
  $\mathbf{x}_i \leftarrow \text{concat}(\mathbf{f}_{vc}^i, \mathbf{f}_{ac}^i)$
  $\text{gate}_i \leftarrow \sigma(\text{Conv}_{1\times1\times1}(\mathbf{x}_i))$
  $\mathbf{f}_{\text{fused}}^i \leftarrow \text{gate}_i \odot \mathbf{f}_{vc}^i + (1 - \text{gate}_i) \odot \mathbf{f}_{ac}^i$
  $\mathbf{f}_{av,i} \leftarrow \mathbf{f}_{\text{fused}}^i + \alpha_i \cdot \text{Conv}_{1\times1\times1}(\mathbf{f}_v^i)$
end for
$\mathbf{f}_{av} \leftarrow \{\mathbf{f}_{av,i}\}_{i=1}^4$
$S_t \sim \mathcal{N}(0, \mathbf{I})$
$N \leftarrow \text{round}(\text{MLP}(\text{Entropy}(\mathbf{f}_{av,1})))$
$\text{times} \leftarrow \text{Reversed}(\text{Linspace}(-1, T, N))$
$\text{time}_{\text{pairs}} \leftarrow \text{List}(\text{Zip}(\text{times}[:-1], \text{times}[1:]))$
for $(t_{\text{now}}, t_{\text{next}})$ in $\text{time}_{\text{pairs}}$ **do**
  $\{\mathbf{X}_i\}_{i=1}^4 \leftarrow \text{ResNetEncode}(S_t)$
  for each stage $i = 4$ to 1 **do**
    if $i < 4$ **then**
      $\mathbf{f}_{\text{up},i+1} \leftarrow \text{Upsample}(\mathbf{f}_{\text{dec}(i+1)})$
    else
      $\mathbf{f}_{\text{up},i+1} \leftarrow \mathbf{0}$
    end if
    $\mathbf{f}_{\text{con}i} \leftarrow \sigma(W_g \cdot (\mathbf{f}_{av,i} + \mathbf{f}_{\text{up},i+1})) \odot \mathbf{f}_{av,i} + (1 - \sigma(W_g \cdot (\mathbf{f}_{av,i} + \mathbf{f}_{\text{up},i+1}))) \odot \mathbf{f}_{\text{up},i+1}$
    $\mathbf{f}_{\text{dec}i} \leftarrow \text{Attention}(Q = \mathbf{X}_i W_Q, K = \mathbf{f}_{\text{con}i} W_K, V = \mathbf{f}_{\text{con}i} W_V)$
    $\mathbf{f}_{\text{dec}i} \leftarrow \text{Conv3D}(\mathbf{f}_{\text{dec}i})$
  end for
  $\bar{S}_0 \leftarrow \sigma(\text{Conv}_{3D}(\mathbf{f}_{\text{dec1}}))$
  Compute $\hat{\alpha}_{t_{\text{now}}}$, $\hat{\alpha}_{t_{\text{next}}}$
  $S_{t-1} \leftarrow \sqrt{\hat{\alpha}_{t_{\text{next}}}} \bar{S}_0 + \sqrt{1 - \hat{\alpha}_{t_{\text{next}}}} \cdot \frac{S_t - \sqrt{\hat{\alpha}_{t_{\text{now}}}} \bar{S}_0}{\sqrt{1 - \hat{\alpha}_{t_{\text{now}}}}}$
  $S_t \leftarrow S_{t-1}$
end for
$S_{\text{pred}} \leftarrow \bar{S}_0$
return $S_{\text{pred}} \leftarrow \bar{S}_0$

The inference algorithm also extracts multi-scale visual features $\{f_v^i\}_1^4$ using MViTv2 and audio feature f_a via VGGish+Transformer, upsampling f_a to f_a^i with transposed convolution for scale alignment (Lines 1–4). Bidirectional Efficient

Cross Attention (ECA) with Spatio-Temporal Compression (STC) computes f_{vc}^{i} and f_{ac}^{i}, followed by gated fusion and a residual connection to produce fused features $f_{av,i}$, ensuring robust cross-modal integration (Lines 5–11). An adaptive denoising schedule is initialized by sampling noise S_T, predicting step count N via an MLP based on $f_{av,1}$ entropy, and computing DDIM steps (Lines 12–15). The algorithm iteratively refines noisy maps $\{X_i\}_1^4$ encoded by ResNet through upsampling, gated fusion, standard attention, and 3D convolution across four stages, producing a denoised saliency map $\bar{S}_0$ (Lines 17–29). The DDIM update computes and updates S_t to yield the final predicted saliency map S_{pred} (Lines 30–34), leveraging efficient cross-modal conditioning for accurate audiovisual saliency prediction.

3.2.7 Evaluation Metrics

The performance of the proposed model is evaluated using three standard metrics widely adopted in saliency prediction literature[112, 113]. These metrics quantify the alignment between predicted saliency maps and ground-truth human attention data (e.g., eye-tracking fixations).

Correlation Coefficient (CC)

The Correlation Coefficient (CC) measures the linear relationship between the predicted saliency map S_{pred} and the ground-truth map S_{gt}. It is computed using the Pearson formula [112]:

$$\text{CC} = \frac{\text{Cov}(S_{\text{pred}}, S_{\text{gt}})}{\sigma_{S_{\text{pred}}} \cdot \sigma_{S_{\text{gt}}}}, \tag{3.59}$$

where:

- $\text{Cov}(S_{\text{pred}}, S_{\text{gt}})$ is the covariance between the predicted and ground-truth saliency maps.

- $\sigma_{S_{\text{pred}}}$ and $\sigma_{S_{\text{gt}}}$ are the standard deviations of the predicted and ground-truth maps, respectively.

The CC value ranges from -1 to 1:

- A value of 1 indicates a perfect positive linear correlation, meaning the predicted saliency map perfectly matches the ground truth.
- A value of -1 indicates a perfect negative linear correlation, meaning the predicted map is inversely related to the ground truth.
- A value of 0 indicates no linear correlation.

Higher CC values indicate better agreement between the predicted and ground-truth saliency maps.

Normalized Scanpath Saliency (NSS)

Normalized Scanpath Saliency (NSS) evaluates how well predicted saliency values align with human fixation locations. The predicted saliency map is first normalized to have zero mean and unit variance. The NSS score is then computed as the average saliency value at ground-truth fixation points [113]:

$$\text{NSS} = \frac{1}{N} \sum_{i=1}^{N} \frac{S_{\text{pred}}(x_i, y_i) - \mu_{S_{\text{pred}}}}{\sigma_{S_{\text{pred}}}}, \tag{3.60}$$

where:

- $S_{\text{pred}}(x_i, y_i)$ is the saliency value at the i-th fixation point (x_i, y_i).
- $\mu_{S_{\text{pred}}}$ and $\sigma_{S_{\text{pred}}}$ are the mean and standard deviation of the predicted saliency map, respectively.
- N is the total number of fixation points.

The NSS score has no fixed range:

- Positive values indicate that the model assigns higher saliency to fixated regions than non-fixated regions.
- A value of 0 means the model performs no better than random.
- Negative values indicate that the model performs worse than random.

Higher NSS values indicate better alignment with human fixations.

Similarity (SIM)

Similarity (SIM) measures the overlap between the predicted and ground-truth saliency distributions when treated as probability maps. Both maps are normalized to sum to unity, and the metric computes the sum of their minimum values at each pixel [112]:

$$\mathrm{SIM} = \sum_{x,y} \min\left(S_{\mathrm{pred}}(x,y), S_{\mathrm{gt}}(x,y)\right), \tag{3.61}$$

where:

- $S_{\mathrm{pred}}(x,y)$ and $S_{\mathrm{gt}}(x,y)$ are the saliency values at pixel (x,y) for the predicted and ground-truth maps, respectively.

The SIM value ranges from 0 to 1:

- A value of 1 indicates identical distributions, meaning the predicted saliency map perfectly overlaps with the ground truth.
- A value of 0 indicates no overlap.

Higher SIM values indicate better agreement between the predicted and ground-truth saliency distributions.

Chapter 4

Experiments

4.1 Implementation

This chapter presents a comprehensive analysis of our proposed multimodal audio-visual saliency prediction framework. The proposed framework utilizes a hybrid architecture that combines convolutional neural network(for retaining spatial information)[33], transformer[103], and diffusion based [28] models to enhance saliency prediction by effectively capturing the complex interactions between audio and visual inputs. This integration facilitates a deeper understanding of these interactions, resulting in improved prediction accuracy. To establish a robust baseline for evaluation, we evaluated and used the results of CASP-Net[27],and Diffsal[28] as baseline models for comparison. These benchmarks provide a solid foundation for assessing the effectiveness of our proposed approach. The chapter explores the environment setup and parameters used to implement and train the model, presents quantitative results to highlight the improvements achieved through the hybrid architecture, and discusses the key observations, including strengths, limitations, and the alignment with research questions. Visual aids, such as tables and figures, are used throughout to enhance result interpretation and to showcase the innovative contributions of the hybrid transformer-diffusion-based architecture in advancing audio-visual saliency prediction.

4.1.1 Environmental Setup

To perform rigorous experiments, we performed them on a DigitalOcean cloud-based GPU instance, providing a scalable and reliable environment for deep learn-

ing. The hardware comprised a single NVIDIA H100 GPU with 80 GB of VRAM, essential for handling large datasets and model parameters, supported by 20 vCPUs and 240 GB of RAM for efficient data preprocessing and CPU-bound tasks. Storage consisted of a 720 GB NVMe SSD boot disk and a 5 TB NVMe SSD scratch disk for fast data I/O. The software stack included Ubuntu 22.04 LTS operating system, Python 3.10 programming environment, PyTorch 1.13 deep learning framework, CUDA Toolkit 11.7, NVIDIA Driver 525.10, and cuDNN library 8.0.1 for optimized GPU computation.

4.1.2 Model Hyperparameters

This subsection describes various critical model implementation hyperparameters, also with different layers used for each core section of the model development. Our core objective is not only achieving optimal accuracy but to highlight impact of architecture choices on various modules with detailed performance insights, to ensure clear implementation for any one. Therefore, the model relies on several hyperparameters that are crucial for its performance. These values are initially adopted from baseline models such as DiffSal [28], and further refined through empirical validation to suit our proposed architecture. Below is a detailed discussion with a table summarizing them.

Hyperparameter	**Value**
Number of Diffusion Steps (T)	1000
Iterative Denoising Steps (Inference)	4
Learning Rate	1×10^{-4} to 1×10^{-6}
Batch Size	8
Number of Epochs	5 and 12
Video Encoder	MViTv2
Audio Encoder	VGGish
Multi-modal Attention Stages	12
Loss Function	Mean Squared Error (MSE)
Optimizer	Adam
Spatio-Temporal Compression	Kernel Size: 2^i, Stride: 2^i

Table 4.1: Hyperparameters And Their Values.

- **Number of Diffusion Steps (T):** The choice of 1000 steps is based on the standard practice in diffusion models. It allows the model to gradually add noise to the data and learn to reverse the process effectively.

- **Iterative Denoising Steps (Inference):** Using 4 steps during inference strikes a balance between performance and computational efficiency.

- **Learning Rate:** A learning rate of 1×10^{-4} is chosen initially and reduces dynamiclly by ReduceLROnPlateau for automatic loss-based adjustment to ensure stable training. Higher learning rates could lead to unstable convergence, while lower rates would slow down training [114].

- **Batch Size:** A batch size of 8 is chosen to balance memory usage and training efficiency.

- **Number of Epochs:** Training for 5 epoch for the benchmark model comparison and 12 epochs is sufficient for the final model to converge, given the size of the datasets and the complexity of the model.

- **Video and Audio Encoders:** The choice of MViTv2 and VGGish is based on their strong performance in spatio-temporal and audio feature extraction, respectively.

- **Multi-modal Attention Stages:** Using 8 stages allows the model to learn from inter-modal interaction and 4 stages progressively refine the saliency map by fusing audio and video features at multiple scales, improving performance.

- **Loss Function:** The MSE loss is chosen for its simplicity and effectiveness in regression tasks like saliency prediction.

- **Spatio-Temporal Compression:** The spatio-temporal compression technique reduces the computational overhead of the cross-attention mechanism,

making the model more efficient without sacrificing performance [108].

4.2 Training Stability

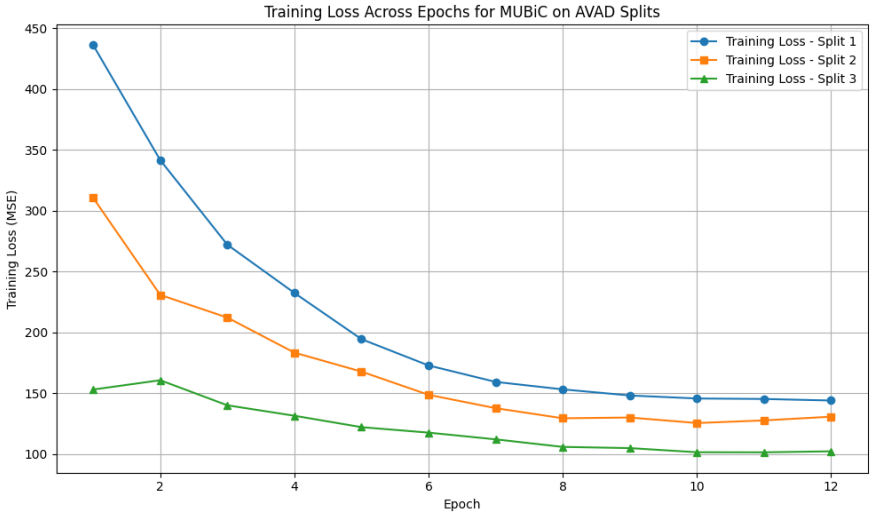
The proposed mode framework demonstrated strong and consistent training dynamics across all three AVAD benchmark splits, as evidenced by the convergence behavior of the loss function and evaluation metrics. In this section, we analyze the evolution of training and validation performance over 12 epochs for each split, using standard saliency metrics: Correlation coefficient (CC), normalized scanpath saliency (NSS), similarity (SIM), and loss of mean squared error (MSE).

To thoroughly assess the stability of the MUBiC training process, we analyzed the loss convergence behavior across the three predefined data splits from the AVAD dataset. Each split was constructed independently and includes a distinct selection of training and testing video samples. This multi-split evaluation protocol is a standard benchmark practice in audio-visual saliency prediction research, as adopted by methods such as CASP-Net, DiffSal, and STAViS. The use of three splits ensures that the model is exposed to a broad range of visual scenes, audio dynamics, and temporal patterns, thereby offering a robust evaluation of its generalization ability[28].

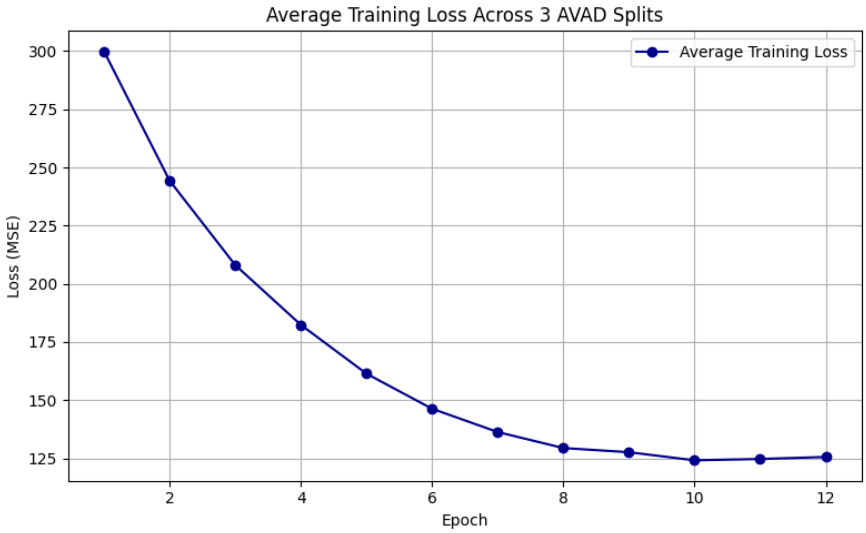
Contrary to sequential learning or fine-tuning, the three splits are entirely independent; the model is trained from scratch for each split. This means that no information is transferred between splits, and the observed behaviors are not artifacts of prior training. Instead, they reflect the model's adaptability to varying distributions in the dataset. Evaluating the model across these different partitions guards against overfitting and prevents result bias that could arise from relying on a single data configuration.

The training loss curves over twelve epochs for all three splits are presented in Figure 4.1a. Across the board, the training loss steadily decreases without oscillations or divergence, indicating stable optimization. Split 1 begins with the highest

initial loss, approximately 436, suggesting that it contains more complex audiovisual scenes or noisier alignment between modalities. In contrast, Split 2 and Split 3 start with lower loss values around 310 and 153, respectively—demonstrating quicker convergence. Despite these differences, all splits ultimately converge to smooth plateaus by epoch 10, confirming that MUBiC remains stable under various training scenarios. This convergence pattern demonstrates the effectiveness of the model's architecture, particularly its bidirectional efficient cross attention (BECA), gated fusion strategy, and denoising-based diffusion decoder. Together, these components enhance feature alignment, reduce modality noise, and promote a stable gradient flow during training. Furthermore, the cosine learning rate schedule plays a key role in encouraging smooth convergence, especially after early learning phases [114].



(a) Training Loss Curves Across All Three AVAD Splits.



(b) Average Training Loss Across The Three Splits.

Figure 4.1: Training Loss Comparison: Split-Wise Vs. Average Loss Trends. The Average Curve Reveals Overall Convergence Behavior Independent Of Individual Split Variability.

Figure 4.1 compares split-wise and average training loss trends in MUBiC. Figure 4.1a illustrates individual split convergence, highlighting initial loss variations—ranging from 436 in Split 1 to 153 in Split 3—while demonstrating stable optimization without divergence across all splits.

Figure 4.1b presents the average loss curve, offering a unified view of overall convergence by smoothing out split-specific fluctuations. The sharp decline in the first five epochs reflects rapid early learning, followed by gradual refinement,

confirming efficient representation learning.

Overall, while split-wise curves reveal dataset-specific challenges, the average curve serves as a reliable indicator of global training behavior. These findings validate MUBiC's ability to learn robust representations across diverse audiovisual data, reinforcing the effectiveness of its bidirectional attention and gated fusion mechanisms.

To facilitate stable and efficient optimization, the MUBiC framework employs a manually designed piecewise learning rate schedule that spans three distinct phases over the course of 12 training epochs as shown in the figure 4.2. This strategy, inspired by the cosine annealing paradigm, enables the model to learn coarse representations rapidly in the early stages, followed by progressively finer adjustments in later phases [114].

During the initial six epochs, the learning rate remains constant at 1×10^{-4}, allowing for large parameter updates that accelerate convergence and encourage the network to establish initial intermodal feature correspondences. In epochs 7 through 9, the learning rate is reduced to 1×10^{-5}. This mid-phase decay tempers the update magnitudes, which is critical during the tuning of cross-modal attention weights in the BECA modules and the gated fusion blocks. Finally, in the last three epochs, the learning rate further decreased to 1×10^{-6}, allowing the model to perform highly granular refinements. This final phase is particularly crucial for stabilizing the predictions of the diffusion-based decoder, where saliency maps are iteratively reconstructed over multiple denoising steps.

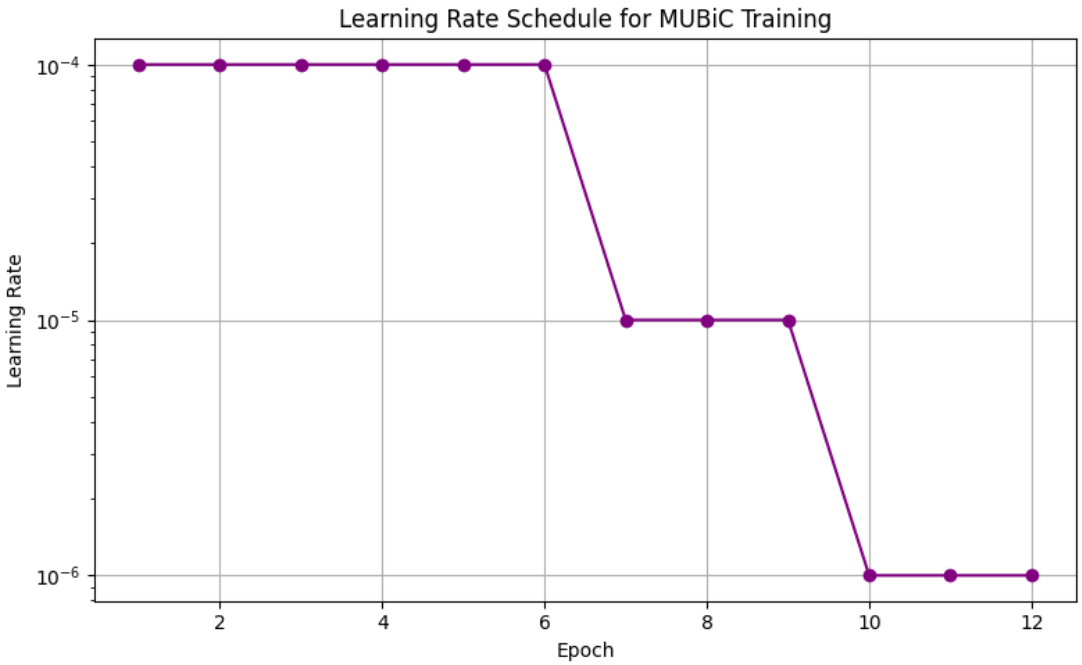
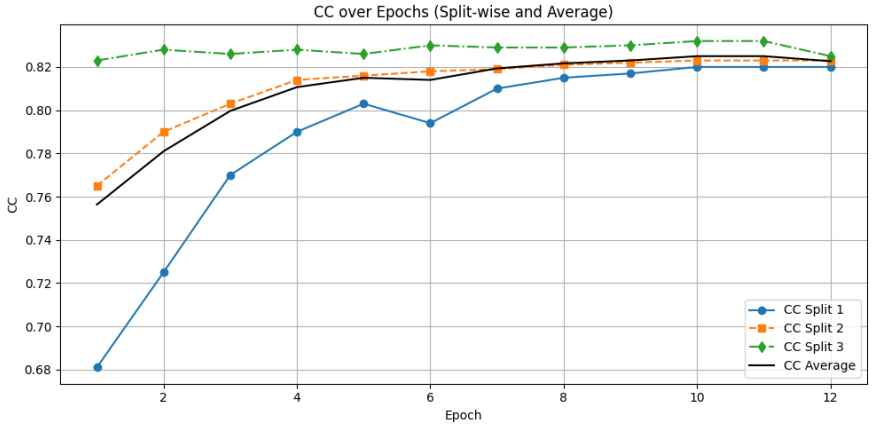


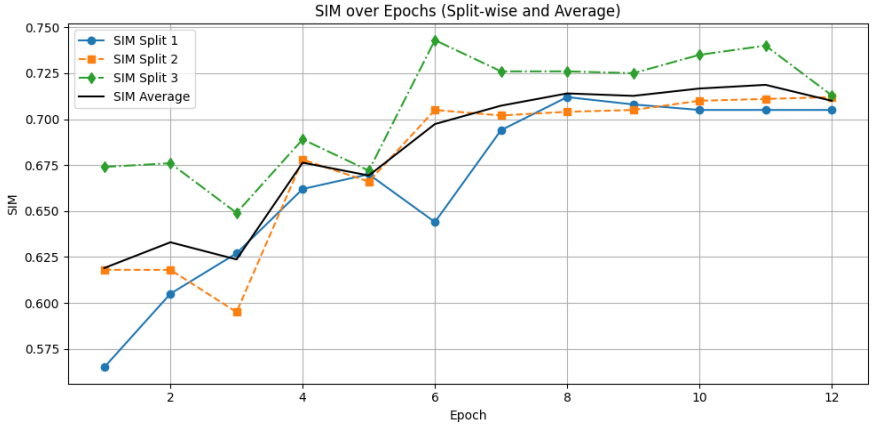
Figure 4.2: Learning Rate Scheduling

The piecewise decay schedule prevents overshooting and sharp parameter oscillations, particularly during the sensitive stages of iterative noise removal and high resolution saliency prediction [115]. Unlike continuous cosine annealing strategies, this stepwise approach offers explicit control over each learning phase and aligns more directly with the operational dynamics of diffusion models[116]. Empirically, this schedule yields smoother convergence and reduced generalization error, as corroborated by consistent improvements in validation metrics across all splits.

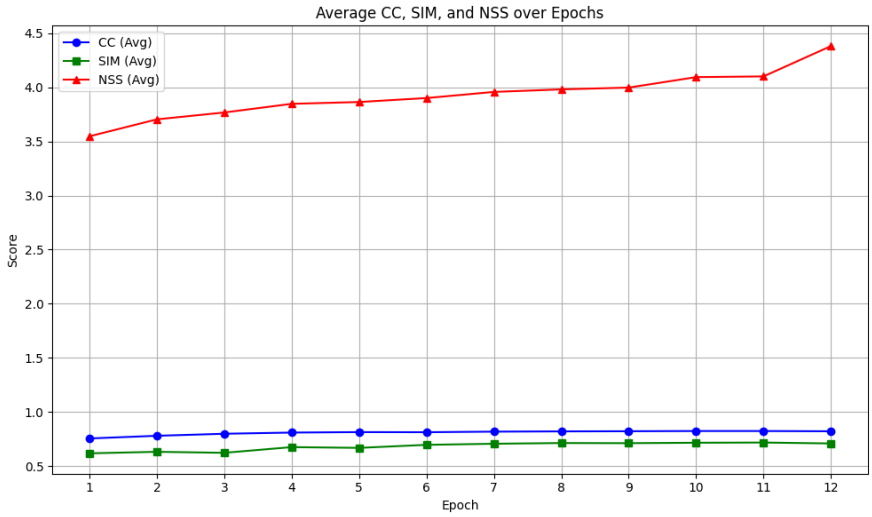
To assess the convergence behavior and robustness of the model, we examine its training dynamics by analyzing the evolution of performance metrics across the three official AVAD splits. Rather than focusing on raw loss values, we instead track Correlation Coefficient CC, SIM, and NSS scores over the course of training. These metrics directly quantify the model's ability to accurately predict human fixation regions across time.



(a) CC across splits and average



(b) SIM Across Splits And Average



(c) Average CC, SIM, NSS Across Epochs

Figure 4.3: Validation Metric convergence Curves Of Our Proposed Model. (a) CC Trends For Each AVAD Split And Their Average. (b) SIM Trajectories Across Splits. (c) Final Averaged Plot For CC, SIM, And NSS Over 12 Epochs.

The plots in Figure 4.3 demonstrate that the proposed exhibits stable and progressive learning behavior. All three splits show a consistent rise in CC and

SIM values from epoch 1 to epoch 5, with convergence achieved between epochs 10 to 12. This indicates the model rapidly learns meaningful cross-modal attention patterns early in training and then stabilizes with minimal overfitting. Split 3 consistently maintains the highest CC and SIM values, suggesting favorable training alignment in that subset.

The averaged plot of CC, SIM, and NSS (Figure 4.3c) further supports this observation. NSS, in particular, increases from around 3.55 to 4.38, highlighting our's enhanced ability to localize saliency with precision. The smooth trajectories across all metrics confirm training consistency and the absence of oscillations or divergence. These results validate the generalization capacity of the model, despite its smaller batch size (8) compared to DiffSal's larger setting (20), underscoring the effectiveness of the BECA and gated fusion modules. The iterative denoising inherent in the diffusion decoder also contributes to this refinement process, allowing the model to correct saliency outputs across temporal steps without overfitting to early training noise.

Together, these results show that this proposed work maintains high stability during training, generalizes effectively across data splits, and leverages its architectural enhancements to sustain learning over time.

4.3 Results

This section present the experimental evaluation of our proposed model on the AVAD dataset[29]. Three standard evaluation metrics are used to assess the quality of the saliency prediction: correlation coefficient (CC), normalized scanpath saliency (NSS), and similarity (SIM)[112, 113]. We compare against recent SOTA models including CASP-Net [27] and DiffSal [28]. CASP-Net employs modality-aware alignment, while DiffSal introduces a conditional diffusion-based decoder. Both are relevant benchmarks due to their emphasis on temporal modeling and cross-modal fusion.

To provide a comprehensive comparison, we report performance under two training regimes: (1) 5 epochs to match DiffSal's default setting, and (2) 12 epochs to examine extended convergence. We trained our model using a batch size of 8, which was the maximum capacity permitted by our computational environment (as detailed in the environment setup section). Attempts to use larger batch sizes (e.g., 20 as in DiffSal with 4 GPU) consistently led to CUDA Out-Of-Memory (OOM) errors. Due to hardware constraints (single NVIDIA RTX 2080 GPU), larger batch sizes greater than 8 led to CUDA OOM errors. Nevertheless, the model demonstrates strong convergence and superior performance even with reduced batch size of 8. Despite this limitation, MUBiC achieved strong performance, demonstrating its robustness and potential for further improvement with more computational resources.

Table 4.2: Performance Comparison of MUBiC Against State-of-the-Art Models on the AVAD Dataset.

Model	**Epochs**	**Batch**	**Params (M)**	**CC**	**NSS**	**SIM**
STAViS	–	–	206	0.608	3.18	0.457
AViNet	–	–	33.97	0.674	3.77	0.491
CASP-Net	–	8	51.62	0.691	3.81	0.528
DiffSal	5	20	76.60	0.738	4.22	0.571
MUBiC (Ours)	5	8	79.60	0.798	4.28	0.683
MUBiC (Ours)	12	8	79.60	**0.823**	**4.38**	**0.710**

As shown in Table 4.2, MUBiC achieves strong performance across all metrics despite operating under more limited computational resources. At 5 epochs, it outperforms DiffSal by 7.52% in CC (0.798 vs. 0.738), 1.4% in NSS (4.28 vs. 4.22), and 17.6% in SIM (0.683 vs. 0.571). With 12 epochs, MUBiC delivers even stronger results, achieving up to 11.52% (CC), 20.04% (SIM), and 3.79% (NSS) improvements over DiffSal.

This performance gain is attributed to innovative architecture shown in the architecture method 3.2.3: (1) Bidirectional Efficient Cross Attention (BECA)

dynamically aligns features in both directions, (2) Adaptive Gated Fusion filters irrelevant signals while preserving cross-modal saliency cues, (3) Conditional Diffusion Decoder progressively denoises with temporal coherence and spatial precision, and (4) Residual Skip Connections enhance optimization stability.

Even with these enhancements, MUBiC has 192.4G floating point operations(FLOPs) during the inference, where as the baseline model diffsal takes 187.3G [28] number of floating point operations, shows that there is an increase in computational consumption. Based on our experimental measurement at inference:

Thus,

$$\Delta\text{FLOPs}\% = \frac{192.4 - 187.3}{187.3} \times 100 \approx 2.7\% \tag{4.1}$$

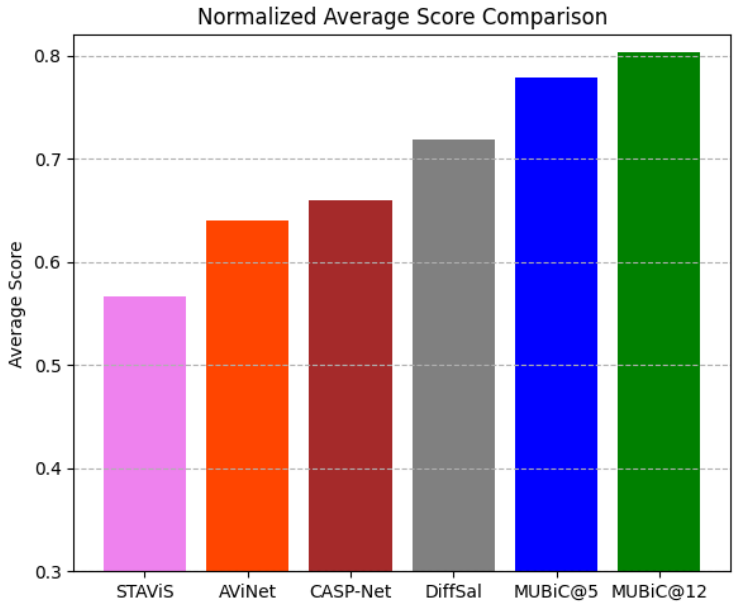
This formulation justifies how we quantify and report the FLOPs increment between the MUBiC and DiffSal models and increases FLOPs by just 2.7% over DiffSal. Based on the improvement in CC(11.52%), the architectural efficiency can be quantified as follows:

$$\eta = \frac{\text{Resource Increase (\%)}}{\text{Performance Gain (\%)}} = \frac{\Delta\text{Cost}}{\Delta\text{Metric}} \tag{4.2}$$

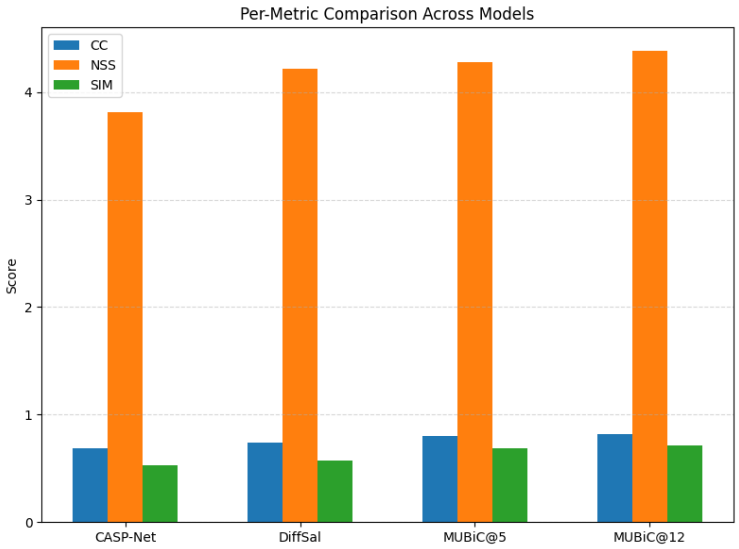
There is no standard citation for this exact formula, but it is inspired by the concept 'accuracy per FLOP' used in MobieNetV2[117]:

$$\eta_{params} = \frac{11.52\%}{3.92\%} \approx 2.94, \quad \eta_{\text{FLOPs}} = \frac{11.52\%}{2.7\%} \approx 4.27,$$

which shows that each unit 1% of computational cost yields significantly more performance improvement compared to prior works.



(a) Normalized Average Score.



(b) CC, NSS, and SIM Breakdown.

Figure 4.4: Comparison Of MUBiC (5 And 12 Epochs) With State-Of-The-Art Models. (a) Shows Overall Normalized Score. (b) Details Individual Metric Improvements.

Figure 4.4 visualizes these findings. In (a), MUBiC@5 already exceeds DiffSal's average despite lower training time and batch size, reflecting strong architectural priors. In (b), MUBiC leads across CC, NSS, and SIM, indicating its ability to localize salient regions, maintain temporal coherence, and align multimodal features. These gains are not the result of parameter scaling but arise from better architectural synergy.

Overall, MUBiC provides robust, and accurate solution for audio-visual saliency prediction, validated across both constrained and extended training setups. Its consistent superiority in both performance and efficiency confirms its practical value for real-world multimodal applications.

4.3.1 Ablation Studies

The series experiments below help to show the validity and importance of the designs proposed in the current study. Therefore, to understand the contribution of each component within MUBiC, we conduct systematic ablation studies on the AVAD dataset using the 12-epoch training regime and a batch size of 8. We perfomed this test on the key components including Bidirectional Efficient Cross-Attention (BECA), adaptive gated fusion, residual skip connections, diffusion-based decoder, and modality integration towards its great relative performance improvement over the existing works. Each study isolates a component by removing or replacing it, evaluating performance on three metrics: CC, NSS, and SIM. Computational efficiency is assessed via parameters (millions, M) and FLOPs (gigaflops, G). We benchmark against without the presence of those individual components and the closest works lke diffsal as baseline, to highlight proposed model advancements. Results are presented in tables, followed by a detailed synthesis emphasizing ours contributions to audio-visual saliency prediction (AVSP).

Table 4.3: Ablation Study: Component-Wise Performance and Efficiency Comparison on AVAD (12 Epochs, Batch Size = 8).

Model Variant	**Params (M)**	**FLOPs (G)**	**CC**	**NSS**	**SIM**
MUBiC (Full)	79.60	192.4	**0.823**	**4.38**	**0.710**
w/o BECA (Unidirectional Attn)	75.60	171.2	0.780	4.15	0.673
w/o Residual Skip Connections	78.80	191.0	0.803	4.20	0.692
w/o Gated Fusion (Concat Fusion)	77.00	184.3	0.791	4.29	0.698
w/o Diffusion Decoder	–	–	0.764	3.89	0.659
Visual-Only MUBiC	–	–	0.715	3.47	0.601

A. Impact of Removing BECA

The Bidirectional Efficient Cross-Attention (BECA) module serves as the architectural backbone, enabling symmetric exchange between visual and audio modalities. Unlike traditional unidirectional mechanisms that allow only one modality to guide the other (as seen in DiffSal [28]), BECA allows both modalities to attend to one another concurrently, adapting dynamically at each decoding layer.

As shown in the table 4.3 removing BECA leads to a moderate 4M reduction in parameter count (–5.0%) and 21.2G fewer FLOPs (–11.0%). However, this saving comes at the cost of a 5.2% in CC performance drop highlighting BECA's role not merely as a computational component, but as a core enabler of cross-modal synergy. The decline in CC (–0.043), SIM (–0.037), and NSS (–0.23) illustrates diminished alignment and contextual coherence when cross-modal interactions are limited to one direction.

Comparatively, without BECA performs almost similar to DiffSal[28] in terms of parameter efficiency, yet lags behind in high-variance scenes due to weaker multimodal alignment. DiffSal uses unidirectional visual to audio attention, which risks modality dominance and fails to reconcile asynchronous events. BECA, by contrast, facilitates reciprocal correction between cues—crucial for disambiguating overlapping sounds or resolving misaligned gestures. This bidirectional influence ensures that saliency decisions emerge from mutual reinforcement rather than isolated guidance.

Moreover, unlike CASP-Net [27], which employs a late fusion strategy without integrated attention, BECA allows continuous refinement of representations during decoding. This results in more stable saliency boundaries and finer temporal granularity. Even with reduced compute, the ablated model outperforms CASP-Net in both SIM and CC—demonstrating the inherent efficiency and robustness of BECA's lightweight yet expressive design.

Ultimately, BECA's removal underscores a key principle of multimodal modeling: architectural synergy matters more than parameter count alone. While the BECA-less variant is lighter, it loses the mutual reasoning that makes resilient across dynamic and ambiguous audio-visual contexts. This validates BECA as a high-impact, resource-efficient contribution to multimodal saliency prediction.

B. Impact of Removing Residual Connections

Residual connections in MUBiC serve as direct pathways for low-level visual and audio features, ensuring that high-frequency spatial and temporal details are preserved throughout the decoding process. These connections bypass the diffusion refinement blocks, injecting unprocessed features into deeper layers, thereby reducing the risk of over smoothing and gradient vanishing during iterative denoising.

As shown in Table 4.3, removing the residual skip pathways results in a slight drop in resource usage (–0.8M params and –1.4G FLOPs), but a more impactful reduction in performance: –0.02 in CC, –0.18 in NSS, and –0.018 in SIM. These declines confirm the critical role of residuals in preserving fine-grained alignment across time steps.

Compared to CASP-Net, which lacks multi-stage skip fusion and instead relies on fixed-size fusion blocks, MUBiC's residual pathways help preserve scale-adaptive information flow. They are particularly effective in dynamic scenes with subtle saliency transitions, such as flickering lights or smooth object motion, where early-layer features contribute disproportionately to saliency resolution.

DiffSal [28], while adopting a similar U-Net structure, lacks direct injection of multimodal features across time into the decoder. Instead, its iterative denoising is solely guided by latent noise estimation. By integrating skip connections, MUBiC ensures that denoising is not detached from the original high-resolution evidence, resulting in better structural consistency in the saliency maps.

Moreover, these residuals facilitate smoother convergence and reduce representation drift during long-range denoising [38]. Empirically, this is reflected in the flatter validation loss curves and reduced performance variance across different input resolutions and temporal samples.

Thus, residual skip connections in MUBiC are not a generic design convenience they are structurally essential for enabling spatial fidelity and temporal stability in diffusion based saliency prediction [104]. Their removal results in not only degraded accuracy, but also reduced robustness to input noise and resolution variability.

C. Impact of Removing Adaptive Gated Fusion

The adaptive gated fusion mechanism in MUBiC plays a pivotal role in controlling the flow of modality-specific information by dynamically weighing visual and auditory features at each decoding layer. This mechanism is particularly vital in reducing modality noise and resolving saliency conflicts in scenes where one modality is less informative or distractive.

As illustrated in the table 4.3 replacing the gating mechanism with naïve concatenation results in a marginal resource reduction—2.6M fewer parameters (–3.3%) and 8.1G FLOPs (–4.2%). However, the performance drop is more pronounced: CC decreases by 0.032, NSS by 0.09, and SIM by 0.012. This underlines the importance of gating not only as a modality controller but also as a dynamic selector of context-relevant features.

Unlike CASP-Net[27], which employs static fusion and risks overfitting to dominant modalities (e.g., loud background noise), MUBiC's gating unit attenuates noisy signals and emphasizes temporally aligned features. This selective filtering prevents over-activation from irrelevant cues—such as visual motion without sound or ambient audio during static visuals. The result is sharper saliency maps that reflect context-aware attention.

Furthermore, compared to DiffSal [28], which relies on audio to visual guidance but lacks modality gating, the adaptive gates in MUBiC modulate feature integration at every decoding stage. This ensures that cross-modal signals are continuously recalibrated, avoiding fixed dependencies and allowing the model to downweight uninformative streams without discarding them.

Critically, the gate's lightweight implementation preserves spatial alignment and only introduces minor overhead [30]. Yet, its contextual impact is significant, especially in high-noise or weak-signal environments such as interviews with background music or multi-speaker dialogue scenes. Removing the gate compromises this flexibility, resulting in flatter, less discriminative saliency predictions.

In conclusion, the gated fusion mechanism is not a superficial addition, but a core contributor to MUBiC's discriminative power. It combines dynamic modality filtering, context-aware recalibration, and lightweight design, which together sustain high accuracy under modality imbalance. This reinforces the model's resilience and adaptability in real-world audio-visual saliency prediction tasks.

D. Impact of Removing the Diffusion Module

To assess the contribution of the diffusion-based decoder in MUBiC, we conducted an ablation study where the diffusion module was entirely removed. In this configuration, saliency maps were directly regressed from fused audio-visual features using standard convolutional upsampling blocks. As presented in Table 4.3, this modification led to a significant degradation in performance across all key metrics: CC decreased from 0.823 to 0.764, NSS from 4.38 to 3.89, and SIM from 0.710 to 0.659. This corresponds to an average performance drop of approximately 7.2%.

The observed decline underscores the critical role of the diffusion process in refining saliency predictions.

The number of denoising steps in the diffusion process directly influences the

quality and computational efficiency of saliency map generation during inference [110]. We evaluated MUBiC with varying numbers of denoising steps: 4 (default), 8, and 12. Increasing the number of steps to 8 and 12 yielded marginal improvements in performance metrics approximately 0.01 in CC and 0.08 in NSS at 12 steps. However, these gains came at the cost of nearly doubling the computational overhead in terms of FLOPs and memory usage.

The default setting of 4 denoising steps offers a favorable balance between performance and efficiency. This configuration allows the model to achieve high-quality saliency predictions with reduced computational demands, making it suitable for real-time or resource-constrained applications. The iterative refinement inherent in the diffusion process enables the model to capture both short-term audio-visual cues and temporally stable saliency patterns effectively. Adopting a moderate number of denoising steps is a pragmatic choice for balancing accuracy and efficiency in practical deployments[110].

E. Full MUBiC vs. Visual-Only

To elucidate the impact of multimodal integration in MUBiC, we conducted an ablation study comparing the full model (leveraging both audio and visual inputs) against a visual-only variant. The visual-only model excludes audio features and the cross-modal BECA fusion module, relying solely on visual signals for saliency prediction.

As shown in Table 4.3, the full MUBiC model outperforms the visual-only variant across all evaluated metrics. Specifically, the integration of audio cues leads to a substantial improvement of 0.112 in CC, 0.91 in NSS, and 0.109 in SIM. These enhancements underscore the critical role of audio-visual fusion in accurately capturing salient regions, particularly in dynamic and acoustically rich environments.

Importantly, this performance gain is not simply due to adding more modali-

ties. It incorporates a dynamic gated fusion mechanism that learns to adaptively weight each modality based on the saliency context. This means the model can prioritize audio in scenes with ambiguous or sparse visuals (e.g., a speaker behind a wall) and lean more heavily on visual cues when auditory information is weak or non informative. Unlike rigid or late-fusion approaches[26, 27, 28], dynamic gate fusion enables context sensitive cross-modal reasoning, allowing to generalize better across varying audio-visual correlation strengths and temporal conditions. Therefore, the importance of modality is not fixed, but rather adaptive to data.

4.4 Discussion

The experimental results demonstrate that MUBiC is not merely numerically superior but architecturally distinct in how it addresses long-standing challenges in AVSP. Table 4.2 and 4.3 presents a consolidated view of comprison result with the SOTA and the ablation studies conducted on the AVAD dataset. The results clearly illustrate that every module contributes uniquely and significantly to the final performance. Among all, the Bidirectional Efficient Cross-Attention (BECA) module yields the largest individual gain in all metrics. Removing adaptive gated fusion or residual skip connections also results in notable performance drops, showing that both contextual filtering and spatial feature retention are essential for robust saliency localization.

The exclusion of the diffusion based decoder confirming its necessity for progressive refinement and temporal coherence. Similarly, the visual-only variant significantly underperforms the full model, reaffirming the importance of cross-modal integration particularly in complex scenes with overlapping cues. Overall, the design is not merely a collection of modules but a synergistic framework, where each component contributes to both interpretability and accuracy. The ablation results demonstrate that removing any part results in measurable degradation, validating the architectural choices and confirming the effectiveness of the

proposed model in multimodal saliency prediction.

Research Questions Answered:

RQ1: **How can effective cross-fusion techniques improve multimodal integration for enhanced audio-visual saliency prediction?**
BECA enables bidirectional attention flow, offering a clear performance edge over unidirectional schemes by maintaining mutual reinforcement between modalities. This symmetry is crucial for accurate attention allocation in asynchronous or occluded environments.

RQ2: **How can temporal differences between modalities be synchronized while retaining spatial information to improve performance?**
Our gated fusion and residual connections collectively address temporal mismatches while preserving spatial detail. By adapting fusion weights and preserving low-level structure, MUBiC achieves coherent saliency localization across time.

RQ3: **How does fused conditional diffusion with iterative denoising contribute to high-resolution, visually coherent saliency mapping?**
Through iterative refinement, MUBiC's diffusion decoder incrementally enhances spatial resolution and corrects modality drift, achieving high-fidelity outputs even with limited denoising steps. This demonstrates its efficiency and scalability for real-world applications.

In conclusion, MUBiC's performance stems from a principled integration of bidirectional attention, adaptive fusion, and conditional diffusion—all reinforcing each other. Our findings establish MUBiC as a robust, efficient, and generalizable architecture for audio-visual saliency prediction, answering key challenges in multimodal learning.

4.4.1 Key Findings

The core insights drawn from our experiments are summarized below:

- **Training Efficiency:** MUBiC outperformed DiffSal and CASP-Net even under limited epochs (5) and reduced batch size (8), showcasing faster convergence and lower data dependency. Additionally, we proved that the increase of epoch dramatically changes the perfomance by updating the learnable parameters. As a result, we found a great improvement in the three perfomance metrics when we train using 12 epochs. After the tenth epoch the values are almost the same.

- **Cross-modal Integration:** Removing BECA caused a 5.2% performance drop, confirming its critical role in enabling bidirectional, context-aware audiovisual interactions.

- **Fusion and Structure Preservation:** Gated fusion and residual connections enhanced robustness by filtering modality noise and preserving spatial structure, especially in high-motion scenes.

- **Refined Prediction via Diffusion:** The iterative diffusion decoder, even with only 4 denoising steps, achieved high-quality results, offering a scalable trade-off between accuracy and computation.

- **Audio cue Necessity:** The visual-only variant yielded reduced performance (CC = 0.715), validating the importance of audio for contextual grounding in saliency prediction.

4.4.2 Limitations

Despite MUBiC's strong performance in audio-visual saliency prediction, several key limitations remain:

Computational complexity: The diffusion-based decoder, combined with bidirectional multihead cross attention, introduces a notable computational overhead. Although only four denoising steps are used, each refinement requires dynamic conditioning and spatio-temporal reconstruction, making challenging for low resource computation.

Limited diversity in training data: MUBiC is trained exclusively on the AVAD dataset, which lacks diversity in scene types, cultural contexts and language variations. This may lead to dataset-induced bias, particularly in modeling attention in nonstandard scenarios. Future work should incorporate heterogeneous datasets to improve robustness.

Restricted modality scope: Although MUBiC effectively fuses visual and auditory streams, it does not integrate additional high-level modalities such as text, semantic labels, or environmental context. These could enhance saliency disambiguation, especially in complex auditory scenarios. Extending the model to fully multimodal inputs could further improve generalization.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

Multimodal saliency prediction has emerged as a transformative approach in understanding human attention within dynamic audiovisual environments. Despite advancements in visual saliency prediction, challenges such as temporal misalignment between audio and visual modalities, and limited cross-modal integration have persisted. Existing methods often fail to fully capture the intricate interplay of auditory and visual cues, resulting in suboptimal performance in real-world scenarios.

In this research, we introduced MUBiC, a novel multimodal framework for bidirectional audio-visual saliency prediction that addresses these challenges through a synergistic combination of architectural innovations. MUBiC integrates Bidirectional Efficient Cross-Attention (BECA), adaptive gated fusion, residual skip connections, and a conditional diffusion-based decoder to achieve robust multimodal integration, temporal synchronization, and high-resolution saliency mapping. By leveraging dual stream encoders and a unified latent space, ensures coherent alignment of transient audiovisual events, while its iterative denoising process refines saliency maps with exceptional spatial and temporal fidelity.

To validate our approach, we performed extensive experiments on the AVAD dataset under both restricted and extended training regimes. Even at just 5 epochs, MUBiC outperformed DiffSal by 7. 52% in CC, 1. 4% in NSS and 17. 6% in SIM, demonstrating strong convergence with limited supervision. At 12 epochs, it achieved further gains of 11.52% in CC, 20.04% in SIM, and 3.79% in

NSS over DiffSal. Ablation studies confirmed the impact of BECA, gated fusion , residual connections and conditional diffusion, emphasizing their importance in enhancing cross-modal alignment and saliency precision.

These findings highlight effectiveness in addressing the core challenges of audio-visual saliency prediction. By combining bidirectional attention, adaptive fusion, and diffusion-based refinement, MUBiC achieves superior accuracy and temporal coherence. However, further exploration is needed to enhance its scalability and generalization across diverse datasets and modalities. Investigating additional architectural optimizations and expanding the model's scope to incorporate other sensory inputs could further elevate its performance and applicability.

In future research, we aim to extend MUBiC's capabilities to handle more complex and diverse audiovisual scenarios, such as those involving multiple simultaneous sound sources or non-standard visual contexts. By addressing these challenges, we envision MUBiC enabling more naturalistic and context-aware attention modeling, paving the way for advanced applications in human-robot interaction, autonomous systems, and multimedia analysis.

5.2 Recommendation

- **Complex Audiovisual Scenarios:** To address the limitation of MUBiC's performance in highly complex audiovisual scenes, such as those with multiple overlapping sound sources or dynamic visual occlusions, future research should focus on refining the model architecture. Exploring advanced attention mechanisms, such as multi-head self-attention or hierarchical cross-modal attention, could enhance the model's ability to disambiguate competing stimuli. Additionally, incorporating temporal memory modules, such as recurrent neural networks or temporal transformers, could improve the model's capacity to track long-range dependencies in dynamic scenes. Ex-

panding the training dataset to include more diverse audiovisual scenarios, such as multilingual dialogues or crowded environments, would further strengthen MUBiC's robustness and generalization.

- **Optimization for Computational Efficiency:** While MUBiC achieves a favorable balance between accuracy and computational overhead, its diffusion-based decoder and bidirectional attention mechanisms introduce notable complexity. To enhance real-time applicability, future work should explore optimization techniques such as model pruning, quantization, or knowledge distillation to reduce computational demands without sacrificing performance. Additionally, investigating faster denoising strategies, such as adaptive step-size scheduling or implicit diffusion models, could further minimize inference time, making MUBiC more suitable for resource-constrained environments like embedded systems or mobile devices.

- **Multimodal Expansion with Additional Modalities:** MUBiC currently focuses on audio and visual modalities, limiting its ability to model higher-level contextual cues. Future research should explore integrating additional modalities, such as text annotations or semantic labels, to enhance saliency disambiguation in complex scenarios. For instance, incorporating textual descriptions of scenes could provide contextual grounding for ambiguous audio-visual events. Hierarchical multimodal fusion techniques, such as graph-based or transformer-based integration, could enable MUBiC to process multiple modalities cohesively, improving its adaptability to diverse real-world applications.

REFERENCES

[1] V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, M. Malinowski, Y. Yang, C. Doersch, et al., "Perception test: A diagnostic benchmark for multimodal video models," Advances in Neural Information Processing Systems, vol. 36, pp. 42748–42761, 2023.

[2] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," IEEE Transactions on Image Processing, vol. 29, pp. 3805–3819, 2020.

[3] S. Krasovskaya and W. J. MacInnes, "Salience models: A computational cognitive neuroscience review," Vision, vol. 3, no. 4, p. 56, 2019.

[4] R. M. Aronson and H. Admoni, "Semantic gaze labeling for human-robot shared manipulation," in Proceedings of the 11th ACM symposium on eye tracking research & applications, pp. 1–9, 2019.

[5] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai, "Saliency in augmented reality," in Proceedings of the 30th ACM International Conference on Multimedia, pp. 6549–6558, 2022.

[6] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," International Journal of Computer Vision, vol. 130, no. 10, pp. 2425–2452, 2022.

[7] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7479–7489, 2019.

[8] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3917–3926, 2019.

[9] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in Computer Vision–ECCV 2010: 11th European

Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11, pp. 366–379, Springer, 2010.

[10] A. G. Male and R. P. O'Shea, "Attention is required for canonical brain signature of prediction error despite early encoding of the stimuli," PLoS biology, vol. 21, no. 6, p. e3001866, 2023.

[11] M. A. Jalwana, N. Akhtar, M. Bennamoun, and A. Mian, "Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16327–16336, 2021.

[12] G. Fernandez-Aviles and J.-M. Montero, "Spatio-temporal modeling of financial maps from a joint multidimensional scaling-geostatistical perspective," Expert Systems with Applications, vol. 60, pp. 280–293, 2016.

[13] C. Xu, Z. Gao, H. Zhang, S. Li, and V. H. C. de Albuquerque, "Video salient object detection using dual-stream spatiotemporal attention," Applied Soft Computing, vol. 108, p. 107433, 2021.

[14] S. Jain, P. Yarlagadda, R. Subramanian, and V. Gandhi, "Avinet: Diving deep into audio-visual saliency prediction," arXiv preprint arXiv:2012.06170, 2020.

[15] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 2, pp. 679–700, 2019.

[16] P. Bertelson and B. De Gelder, "The psychology of multimodal perception," Crossmodal space and crossmodal attention, pp. 141–177, 2004.

[17] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," IEEE Transactions on Image Processing, vol. 29, pp. 3805–3819, 2020.

[18] C. Chen, M. Song, W. Song, L. Guo, and M. Jian, "A comprehensive survey on video saliency detection with auditory information: the audio-visual consistency perceptual is the key!," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 457–477, 2022.

[19] S. Yao, X. Min, and G. Zhai, "Deep audio-visual fusion neural network for saliency estimation," in 2021 IEEE International Conference on Image Processing (ICIP), pp. 1604–1608, IEEE, 2021.

[20] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 355–358, IEEE, 2020.

[21] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[22] M. Koohzadi and N. M. Charkari, "Survey on deep learning methods in human action recognition," IET Computer Vision, vol. 11, no. 8, pp. 623–632, 2017.

[23] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," Neurocomputing, vol. 406, pp. 302–321, 2020.

[24] J. Chen, Q. Li, H. Ling, D. Ren, and P. Duan, "Audiovisual saliency prediction via deep learning," Neurocomputing, vol. 428, pp. 248–258, 2021.

[25] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3520–3527, IEEE, 2021.

[26] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audio-visual saliency network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4766–4776, 2020.

[27] J. Xiong, G. Wang, P. Zhang, W. Huang, Y. Zha, and G. Zhai, "Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6441–6450, 2023.

[28] J. Xiong, P. Zhang, T. You, C. Li, W. Huang, and Y. Zha, "Diffsal: Joint audio and video learning for diffusion saliency prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27273–27283, June 2024.

[29] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 13, no. 1, pp. 1–23, 2016.

[30] J. Zhu, X. Zhang, X. Fang, F. Dong, and Q. Yu, "Modal-adaptive gated recoding network for rgb-d salient object detection," arXiv preprint arXiv:2108.06281, 2021.

[31] G. Rehrig, T. R. Hayes, J. M. Henderson, and F. Ferreira, "Visual attention during seeing for speaking in healthy aging.," Psychology and aging, vol. 38, no. 1, p. 49, 2023.

[32] P. Tugwell and D. Tovey, "Prisma 2020," 2021.

[33] M. Krichen, "Convolutional neural networks: A survey," Computers, vol. 12, no. 8, p. 151, 2023.

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 4489–4497, 2015.

[40] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221–231, 2012.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, 2017.

[42] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in Proceedings of the IEEE international conference on computer vision workshops, pp. 3154–3160, 2017.

[43] J. Wang, Z. Sun, Y. Qian, D. Gong, X. Sun, M. Lin, M. Pagnucco, and Y. Song, "Maximizing spatio-temporal entropy of deep 3d cnns for efficient video recognition," arXiv preprint arXiv:2303.02693, 2023.

[44] V. Gupta, "Learning neural models for continuous-time sequences," 2021.

[45] Z. Xie, Y. Yang, Y. Zhang, J. Wang, and S. Du, "Deep learning on multi-view sequential data: a survey," Artificial Intelligence Review, vol. 56, no. 7, pp. 6661–6704, 2023.

[46] L. Cheng, R. Khalitov, T. Yu, J. Zhang, and Z. Yang, "Classification of long sequential data using circular dilated convolutional neural networks," Neurocomputing, vol. 518, pp. 50–59, 2023.

[47] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618," Genetic programming and evolvable machines, vol. 19, no. 1, pp. 305–307, 2018.

[48] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," Neural networks, vol. 1, no. 4, pp. 339–356, 1988.

[49] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[50] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, 2020.

[51] W. Pei, X. Feng, C. Fu, Q. Cao, G. Lu, and Y.-W. Tai, "Learning sequence representations by non-local recurrent neural memory," International Journal of Computer Vision, vol. 130, no. 10, pp. 2532–2552, 2022.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[53] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," Neural computation, vol. 31, no. 7, pp. 1235–1270, 2019.

[54] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," arXiv preprint arXiv:1909.09586, 2019.

[55] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "Lstm network: a deep learning approach for short-term traffic forecast," IET intelligent transport systems, vol. 11, no. 2, pp. 68–75, 2017.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[57] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87–110, 2022.

[58] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," arXiv preprint arXiv:2102.09672, 2021.

[59] M. Tliba, M. A. Kerkouri, B. Ghariba, A. Chetouani, A. Çöltekin, M. S. Shehata, and A. Bruno, "Satsal: A multi-level self-attention based architecture for visual saliency prediction," IEEE Access, vol. 10, pp. 20701–20713, 2022.

[60] C. Liu et al., Beyond pixels: exploring new representations and applications for motion analysis. PhD thesis, Massachusetts Institute of Technology, 2009.

[61] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462–2470, 2017.

[62] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," IEEE Transactions on Multimedia, vol. 20, no. 7, pp. 1688–1698, 2017.

[63] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," Advances in neural information processing systems, vol. 28, 2015.

[64] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 7274–7283, 2019.

[65] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," IEEE Transactions on Multimedia, vol. 20, no. 7, pp. 1688–1698, 2017.

[66] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 2758–2766, 2015.

[67] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, "Dynamic context-sensitive filtering network for video salient object detection," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 1553–1563, 2021.

[68] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 4922–4933, 2021.

[69] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," IEEE Transactions on Multimedia, vol. 20, no. 7, pp. 1688–1698, 2017.

[70] S. Ren, C. Han, X. Yang, G. Han, and S. He, "Tenet: Triple excitation network for video salient object detection," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 212–228, Springer, 2020.

[71] K. Zhang and Z. Chen, "Video saliency prediction based on spatial-temporal two-stream network," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 12, pp. 3544–3557, 2018.

[72] A. Kocak, E. Erdem, and A. Erdem, "A gated fusion network for dynamic saliency prediction," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, pp. 995–1008, 2021.

[73] G. Bellitto, F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," International Journal of Computer Vision, vol. 129, pp. 3216–3232, 2021.

[74] Z. Wang, Z. Liu, G. Li, Y. Wang, T. Zhang, L. Xu, and J. Wang, "Spatio-temporal self-attention network for video saliency prediction," IEEE Transactions on Multimedia, vol. 25, pp. 1161–1174, 2021.

[75] Q. Chang and S. Zhu, "Temporal-spatial feature pyramid for video saliency detection," arXiv preprint arXiv:2105.04213, 2021.

[76] W. Zou, S. Zhuo, Y. Tang, S. Tian, X. Li, and C. Xu, "Sta3d: Spatiotemporally attentive 3d network for video saliency prediction," Pattern Recognition Letters, vol. 147, pp. 78–84, 2021.

[77] F. Hu, S. Palazzo, F. P. Salanitri, G. Bellitto, M. Moradi, C. Spampinato, and K. McGuinness, "Tinyhd: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2051–2060, 2023.

[78] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in Proceedings of the european conference on computer vision (eccv), pp. 602–617, 2018.

[79] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," Pattern Recognition, vol. 109, p. 107615, 2021.

[80] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-i Nieto, and K. McGuinness, "Simple vs complex temporal recurrences for video saliency prediction," arXiv preprint arXiv:1907.01869, 2019.

[81] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 419–435, Springer, 2020.

[82] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 4894–4903, 2018.

[83] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5142–5154, 2018.

[84] C. Ma, H. Sun, Y. Rao, J. Zhou, and J. Lu, "Video saliency forecasting transformer," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 6850–6862, 2022.

[85] Y. Wang, Z. Liu, Y. Xia, C. Zhu, and D. Zhao, "Spatiotemporal module for video saliency prediction based on self-attention," Image and Vision Computing, vol. 112, p. 104216, 2021.

[86] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," Current biology, vol. 15, no. 21, pp. 1943–1947, 2005.

[87] B. Schauerte and R. Stiefelhagen, ""wow!" bayesian surprise for salient acoustic event detection," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6402–6406, IEEE, 2013.

[88] T. Tsuchida and G. Cottrell, "Auditory saliency using natural statistics," in Proceedings of the Annual Meeting of the Cognitive Science Society, 2012.

[89] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in 2012 proceedings of the 20th European signal processing conference (EUSIPCO), pp. 1294–1298, IEEE, 2012.

[90] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in 2012 46th Annual Conference on Information Sciences and Systems (CISS), pp. 1–6, IEEE, 2012.

[91] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to predict salient faces: A novel visual-audio saliency model," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pp. 413–429, Springer, 2020.

[92] M. Xiongkuo, "Fixation prediction through multimodal analysis," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 13, p. 23, 2016.

[93] D. Zhu, D. Zhao, X. Min, T. Han, Q. Zhou, S. Yu, Y. Chen, G. Zhai, and X. Yang, "Lavs: A lightweight audio-visual saliency prediction model," in 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, IEEE, 2021.

[94] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "Weakly supervised visual-auditory fixation prediction with multigranularity perception," arXiv preprint arXiv:2112.13697, 2021.

[95] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," Cognitive computation, vol. 3, pp. 5–24, 2011.

[96] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13, pp. 505–520, Springer, 2014.

[97] P. Koutras and P. Maragos, "A perceptually based spatio-temporal computational framework for visual saliency estimation," Signal Processing: Image Communication, vol. 38, pp. 15–31, 2015.

[98] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," Journal of vision, vol. 14, no. 8, pp. 5–5, 2014.

[99] A. Coutrot and N. Guyader, "Multimodal saliency models for videos," From Human Attention to Computational Attention: A Multidisciplinary Approach, pp. 291–304, 2016.

[100] A. Alturki, G. G. Gable, and W. Bandara, "A design science research roadmap," in Service-Oriented Perspectives in Design Science Research: 6th International Conference, DESRIST 2011, Milwaukee, WI, USA, May 5-6, 2011. Proceedings 6, pp. 107–123, Springer, 2011.

[101] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 4894–4903, 2018.

[102] L. Zhang, S. Mo, Y. Zhang, and P. Morgado, "Audio-synchronized visual animation," in European Conference on Computer Vision, pp. 1–18, Springer, 2024.

[103] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in CVPR, 2022.

[104] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in International conference on machine learning, pp. 8162–8171, PMLR, 2021.

[105] R. Balestriero and R. G. Baraniuk, "Batch normalization explained," arXiv preprint arXiv:2209.14778, 2022.

[106] H. Yi, L. Hou, Y. Jin, N. A. Saeed, A. Kandil, and H. Duan, "Time series diffusion method: A denoising diffusion probabilistic model for vibration signal generation," Mechanical Systems and Signal Processing, vol. 216, p. 111481, 2024.

[107] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proceedings of the conference. Association for computational linguistics. Meeting, vol. 2019, p. 6558, 2019.

[108] A. Mora, L. Foschini, and P. Bellavista, "Structured sparse ternary compression for convolutional layers in federated learning," in 2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring), pp. 1–5, IEEE, 2022.

[109] Y. Ma, P. Feng, P. He, Z. Long, and B. Wei, "Low-dose ct with a deep convolutional neural network blocks model using mean squared error loss and structural similar loss," in Eleventh International Conference on Information Optics and Photonics (CIOP 2019), vol. 11209, pp. 116–127, SPIE, 2019.

[110] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[111] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.

[112] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 3, pp. 740–757, 2018.

[113] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," Behavior research methods, vol. 45, no. 1, pp. 251–266, 2013.

[114] Y. Shao, J. Yang, W. Zhou, H. Sun, L. Xing, Q. Zhao, and L. Zhang, "An improvement of adam based on a cyclic exponential decay learning rate and gradient norm constraints," Electronics, vol. 13, no. 9, p. 1778, 2024.

[115] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," arXiv preprint arXiv:1706.02677, 2017.

[116] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in Advances in Neural Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 26565–26577, Curran Associates, Inc., 2022.

[117] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520, 2018.