DiffSal: Joint Audio and Video Learning for Diffusion Saliency Prediction

Anonymous CVPR submission

Paper ID 13555

Abstract

*Audio-visual saliency prediction can draw support from diverse modality complements, but further performance enhancement is still challenged by customized architectures as well as task-specific loss functions. In recent studies, denoising diffusion models have shown more promising in unifying task frameworks owing to their inherent ability of generalization. Following this motivation, a novel **Diff**usion architecture for generalized audio-visual **Sal**iency prediction (DiffSal) is proposed in this work, which formulates the prediction problem as a conditional generative task of the saliency map by utilizing input audio and video as the conditions. Based on the spatio-temporal audio-visual features, an extra network Saliency-UNet is designed to perform multi-modal attention modulation for progressive refinement of the ground-truth saliency map from the noisy map. Extensive experiments demonstrate that the proposed DiffSal can achieve excellent performance across six challenging audio-visual benchmarks, with an average relative improvement of 6.3% over the previous state-of-the-art results by six metrics.*

1. Introduction

With the functionality of visual and auditory sensory systems, human beings can quickly focus on the most interesting areas during their daily activities. Such a comprehensive capability of visual attention in multi-modal scenarios has been explored by numerous researchers and referred to as an *audio-visual saliency prediction* (AVSP) task. Based on the related techniques, many valuable practical applications have come into utility ranging from video summarization [29] and compression [65] to virtual reality [15] and augmented reality [46].

Significant efforts have been dedicated to advancing studies by concentrating on elevating the quality of multi-modal interaction and refining the generalizability of model structures in this domain. Among the prevalent AVSP approaches, as depicted in Figure 1(a), localization-based methods [38, 39, 50] have gained a lot of attention. These methods typically consider the sounding objects as saliency targets in the scene and transform the saliency prediction task into a spatial sound source localization problem. Even though the semantic interactions between audio and visual modalities have been considered in these methods, their focus on a generalized network structure is still limited and inevitably results in constrained performance.

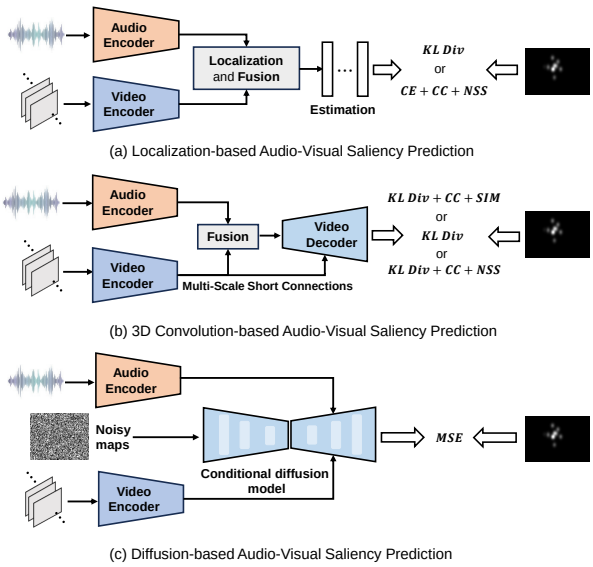

Figure 1. Comparison of conventional audio-visual saliency prediction paradigms and our proposed diffusion-based approach. Both the localization-based and 3D convolution-based methods use tailored network structures and sophisticated loss functions to predict saliency areas. Differently, our diffusion-based approach is a generalized audio-visual saliency prediction framework using simple *MSE* objective function.

In contrast, recent 3D convolution-based methods [10, 30, 58, 61] exhibit superior performance in predicting audio-visual saliency maps, as illustrated in Figure 1(b). However, these methods require customized architectures with built-in inductive biases tailored for saliency prediction tasks. For instance, Jain *et al.* [30] and Xiong *et al.* [58] both embrace a 3D encoder-decoder structure akin to UNet, but integrate their empirical designs into the decoder. Moreover, both localization-based and 3D convolution-

based methods employ sophisticated loss functions, contributing to a more intricate audio-visual saliency modeling paradigm.

Effective audio-visual interaction and the generalized saliency network are two essential factors for the seamless application of AVSP technology in the real world. Unfortunately, an in-depth exploitation of both challenges in the existing works is far from sufficient. Inspired by the strong generalization capabilities, diffusion models [25, 26, 43] can be employed as a *unified* framework for generative tasks with class labels [17], text prompts [20], images [11], and even sounds [45] as the conditions for modeling. However, it remains an open question how to design a diffusion model that satisfies the effective audio-visual interaction and the generalized saliency network.

In this work, we present a conditional **Diff**usion model for generalized audio-visual **Sal**iency prediction (DiffSal), which aims to tackle these two challenges simultaneously, as illustrated in Figure 1(c). Our DiffSal utilizes the input video and audio as the conditions to reformulate the prediction problem as a conditional generative task of the saliency map. During the training phase, the model is fed the video and audio cues as well as a degraded saliency map, which has been obtained from the ground-truth with varying degrees of injected noise. By constructing a two-stream encoder to explore audio and video feature pairs with spatio-temporal coherence, the obtained similar pixel-wise multi-modal features can be utilized to guide the diffusion model generation process. In addition, a novel network Saliency-UNet is employed to recover the original saliency maps from noisy inputs, which utilizes information from spatio-temporal audio and video features as the conditions. To explore the latent semantic associations between audio and video features, an effective multi-modal interaction mechanism is proposed. The entire DiffSal framework employs a *simple* mean square error loss to predict ground-truth saliency maps from random noise. During the inference phase, following the reversed diffusion process, DiffSal performs multi-step denoising to generate predictions based on randomly generated noisy saliency maps.

Benefiting from such a diffusion-based framework, we demonstrate two distinct properties that appeal to the AVSP task. (**i**) In contrast to existing methods with spatio-temporal visual branching [30, 50, 58], DiffSal enables spatio-temporal modeling of audio and video, and can be generalized to audio-only, video-only, as well as audio-visual scenarios. (**ii**) Thanks to the iterative denoising property of the diffusion model, DiffSal can iteratively reuse Saliency-UNet to improve performance without retraining.

To summarize, our main contributions are: (1) We formulate the saliency prediction task as a conditional generative problem and propose a novel conditional diffusion saliency model, which is beneficial from the generalized network structure and effective audio-visual interaction. (2) We demonstrate two properties of DiffSal that are effective on saliency prediction: the ability to be applied to either uni-modal or multi-modal scenarios, and to perform flexible iterative refinement without retraining. (3) Extensive experiments have been conducted on six challenging audio-visual benchmarks and the results demonstrate that DiffSal achieves excellent performance, exhibiting an average relative improvement of 6.3% over the previous state-of-the-art results across four metrics.

2. Related Work

2.1. Audio-Visual Saliency Prediction

For audio-visual saliency prediction, different strategies for multi-modal correlation modeling have been proposed to estimate the saliency maps over consecutive frames. Early solutions [38, 39] attempted to localize the moving-sounding target by canonical correlation analysis(CCA) to establish the cross-modal connections between the two modalities. With the advent of deep learning, Tsiami *et al*. [50] continued the localization-based approach by extracting audio representation using SoundNet [4], and further performed spatial sound source localization through bilinear operations. Unfortunately, these methods exhibited sub-optimal performance only and thus led to the emergence of more effective 3D convolution-based approaches [30, 48, 58] based on the encoder-decoder network frameworks. Jain *et al*. [30] and Xiong *et al*. [58] both embrace the UNet-style encoder-decoder structure by incorporating their empirical designs into the decoder. Moreover, Chang *et al*. [10] employs a complex hierarchical feature pyramid network to aggregate deep semantic features. Considering that both the localization-based and 3D convolution-based methods use tailored network structures and sophisticated loss functions to predict saliency areas. In this study, by formulating the task as a conditional generation problem, a novel conditional diffusion model is proposed for generalized audio-visual saliency prediction.

2.2. Diffusion Model

Recently, diffusion models have gained significant traction in the field of deep learning. During diffusion modeling, the Markov process is employed to introduce noise into the training data followed by a training of deep neural networks to reverse it. Thanks to the high-quality generative results and strong generalization capabilities, diffusion models have achieved an impressive performance in generative tasks, such as image generation [3, 6, 7, 12, 16], image-to-image translation [32, 44, 52, 56, 63], video generation [23, 27, 60], text-to-image synthesis [20, 42, 62], and *etc*. Beyond generative tasks, diffusion models have proven to be highly effective in various computer vision tasks. For in-

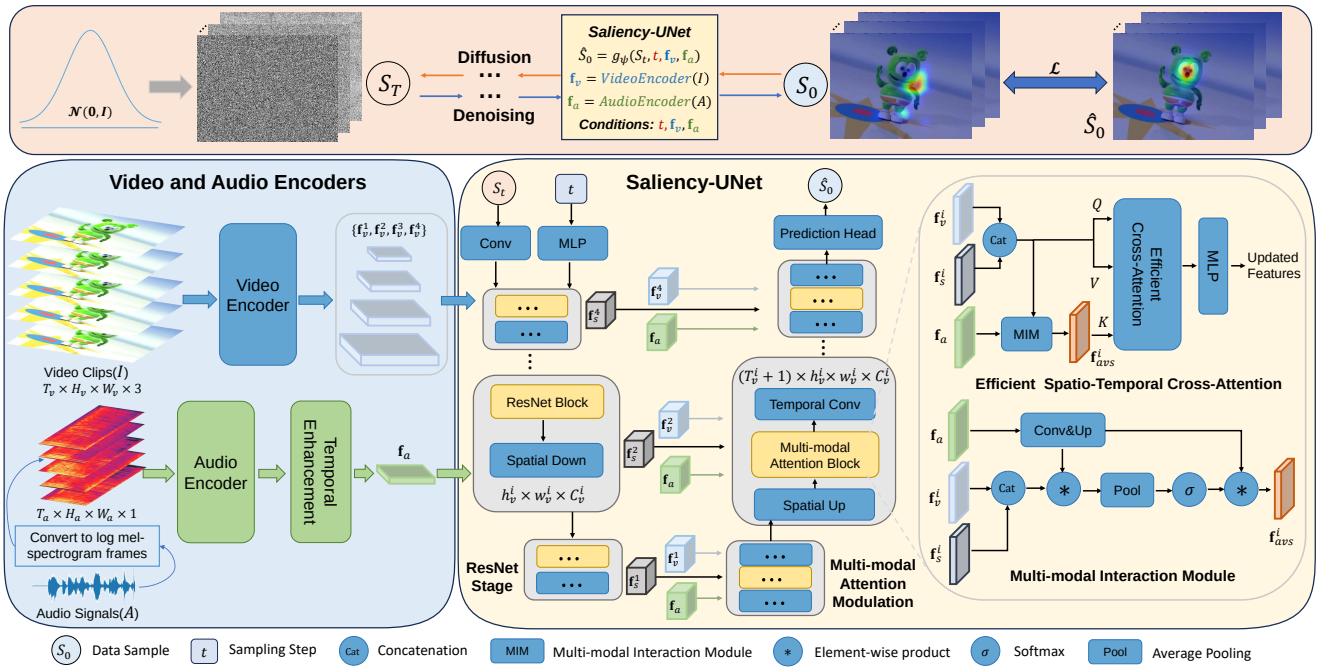

Figure 2. An overview of the proposed DiffSal framework. DiffSal first encodes spatio-temporal video features $\mathbf{f}_v$ and audio features $\mathbf{f}_a$ by the Video and Audio Encoders, respectively. Then the Saliency-UNet takes audio features $\mathbf{f}_a$ and video features $\mathbf{f}_v$ as the conditions to guide the network in generating the saliency map $\hat{S}_0$ from the noisy map S_t.

stance, DiffSeg [1] proposes a diffusion model conditioned on an input image for image segmentation. Chen *et al*. [11] propose a model named DiffusionDet, which formulates object detection as a generative denoising process from noisy boxes to object boxes. Subsequently, the pipeline of this model is extended by Gu *et al*. [21] by introducing noise filters during diffusion, as well as incorporating a mask branch for global mask reconstruction, which makes making DiffusionDet more applicable to instance segmentation tasks. To the best of our knowledge, there have been no previous successful attempts to apply diffusion models to saliency prediction, which inspires the proposed DiffSal in this work to explore the potential of diffusion models in the domain of audio-visual saliency prediction.

3. Preliminaries

Diffusion models [26] are likelihood-based models for points sampling from a given distribution by gradually denoising random Gaussian noise in T steps. In the forward diffusion process, the increased noises are added to a sample point x_0 iteratively as $x_0 \rightarrow \cdots \rightarrow x_{T-1} \rightarrow x_T$, to obtain a completely noisy image x_T. Formally, the forward diffusion process is a Markovian noising process defined by a list of noise scales $\{\bar{\alpha}_t\}_{t=1}^{T}$ as:

$$\begin{aligned} q(x_t|x_0) &:= \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \\ x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(0, \boldsymbol{I}), \end{aligned} \tag{1}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ and β_s denote the noise variance schedule [26], ϵ is the noise, $\mathcal{N}$ denotes normal distribution, x_0 is the original image, and x_t is noisy image after t steps of the diffusion process.

The reverse diffusion process aims to learn the posterior distribution $p(x_{t-1}|x_0, x_t)$ for x_{t-1} estimation given x_t. Typically, this can be done using a step-dependent neural network in multiple parameterized ways. Instead of directly predicting the noise ϵ, we choose to parameterize the neural network $f_\theta(x_t, t)$ to predict x_0 as [11]. For model optimization, a mean squared error loss is employed to match $f_\theta(x_t, t)$ and x_0:

$$\mathcal{L} = \|f_\theta(x_t, t) - x_0\|^2, t \in_R \{1, 2, \ldots, T\}, \tag{2}$$

where the step t is randomly selected at each training iteration. From starting with a pure noise $x_t \in \mathcal{N}(0, \boldsymbol{I})$ during inference stage, the model can gradually reduce the noise according to the update rule [47] using the trained f_θ as below:

$$\begin{aligned} x_{t-1} =& \sqrt{\bar{\alpha}_{t-1}} f_\theta(x_t, t) + \\ & \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\bar{\alpha}_t} f_\theta(x_t, t)}{\sqrt{1-\bar{\alpha}_t}} + \sigma_t \epsilon. \end{aligned} \tag{3}$$

Iteratively applying Eq. 3, a new sample x_0 can be generated from f_θ via a trajectory $x_T \rightarrow x_{T-1} \rightarrow \cdots \rightarrow x_0$. Specially, some improved sampling strategies skip such an operation in the trajectory to achieve better efficiency [47].

To control the generation process, the conditional information can be modeled and incorporated in the diffusion model as an extra input $f_\theta(x_t, t, \mathbb{C})$. The class labels [17], text prompts [20], and audio guidance [2] are the prevalent forms of conditional information documented in the literature.

4. Method

To tackle the challenges of effective audio-visual interaction and saliency network generalization, we formulate audio-visual saliency prediction as a conditional generation modeling of the saliency map, which treats the input video and audio as the conditions. Figure 2 illustrates the overview of the proposed DiffSal, which contains parts of Video and Audio Encoders as well as Saliency-UNet. The former is used to extract multi-scale spatio-temporal video features and audio features from image sequences and corresponding audio signals. By conditioning on these semantic video and audio features, the latter performs multi-modal attention modulation to progressively refine the ground-truth saliency map from the noisy map. Each part of DiffSal is elaborated on below.

4.1. Video and Audio Encoders

Video Encoder. Let $I = [I_1, \cdots, I_j, \cdots, I_{T_v}], I_j \in \mathbb{R}^{H_v \times W_v \times 3}$ denotes an RGB video clip of length T_v. This serves as the input to a video backbone network, which produces spatio-temporal feature maps. The backbone consists of 4 encoder stages and outputs 4 hierarchical video feature maps, illustrated in Figure 2. The generated feature maps are denoted as $\{\mathbf{f}_v^i\}_{i=1}^N \in \mathbb{R}^{T_v^i \times h_v^i \times w_v^i \times C_v^i}$, where $(h_v^i, w_v^i) = (H_v, W_v)/2^{i+1}, N = 4$. In practical implementation, we employ the off-the-shelf MViTv2 [33] as a video encoder to encode the spatial and temporal information of image sequences. More generally, MViTv2 can also be replaced with other general-purpose encoders, *e.g.*, S3D[57], Video Swin Transformer[34].

Audio Encoder. To temporally synchronize the audio features with the video features in a better way, initially, we transform the raw audio into a log-mel spectrogram through Short-Time Fourier Transform (STFT). Then, the spectrogram is partitioned into T_a slices of dimension $H_a \times W_a \times 1$ with a hop-window size of 11 ms. To extract per-frame audio feature $\bar{\mathbf{f}}_{a,i}$ where $i \in \{1, \cdots, T_a\}$, a pre-trained 2D fully convolutional VGGish network [24] is performed on AudioSet [19], resulting in a feature map of size $\mathbb{R}^{h_a \times w_a \times C_a}$. To improve the inter-frame consistency, we further introduce a temporal enhancement module consisting of a patch embedding layer as well as a transformer layer. Then, the audio features are rearranged in the spatio-temporal dimension and fed into the patch embedding layer to obtain $\bar{\mathbf{f}}_a \in \mathbb{R}^{T_a \times h_a \times w_a \times C_a}$. For the retaining of temporal position information, a learnable positional embedding $\mathbf{e}^{pos}$ is incorporated along the temporal dimension:

$$\bar{\mathbf{f}}_a = [\bar{\mathbf{f}}_{a,0} + \mathbf{e}_0^{pos}, \cdots, \bar{\mathbf{f}}_{a,T_a} + \mathbf{e}_{T_a}^{pos}], \tag{4}$$

where $[\cdot;\cdot]$ represents concatenation operation. The processed feature is finally fed into the Multi-head Self Attention (MSA), the layer normalization (LN) [5] and the MLP layer to produce the spatio-temporal audio features $\mathbf{f}_a \in \mathbb{R}^{T_a \times h_a \times w_a \times C_a}$:

$$\begin{aligned} \bar{\mathbf{f}}_a &= \texttt{MSA}(\texttt{LN}(\bar{\mathbf{f}}_a)) + \bar{\mathbf{f}}_a, \\ \mathbf{f}_a &= \texttt{MLP}(\texttt{LN}(\bar{\mathbf{f}}_a)) + \bar{\mathbf{f}}_a. \end{aligned} \tag{5}$$

4.2. Saliency-UNet

To learn the underlying distribution of saliency maps, we design a novel conditional denoising network g_ψ with multi-modal attention modulation, named Saliency-UNet. This network is designed to leverage both audio features $\mathbf{f}_a$ and video features $\{\mathbf{f}_v^i\}_{i=1}^N$ as the conditions, guiding the network in generating the saliency map $\hat{S}_0$ from the noisy map S_t:

$$\hat{S}_0 = g_\psi(S_t, t, \mathbf{f}_a, \mathbf{f}_v), \tag{6}$$

where $S_t = \sqrt{\bar{\alpha}_t} S_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, noise ϵ is from a Gaussian distribution, and $t \in_R \{1, 2, \ldots, T\}$ is a random diffusion step.

Our Saliency-UNet can be functionally divided into two parts: feature encoding and feature decoding, as shown in Figure 2. The first part encodes multi-scale noise feature maps $\{\mathbf{f}_s^i\}_{i=1}^N \in \mathbb{R}^{h_v^i \times w_v^i \times C_v^i}$ from the noisy map S_t using multiple ResNet stages. The second part utilizes our designed multi-modal attention modulation (MAM) across multiple scales for the interaction of noise features, audio features, and video features. The MAM stage comprises an upsampling layer, a multi-modal attention block, and a 3D temporal convolution. This stage not only computes the global spatio-temporal correlation between multi-modal features but also progressively enhances the spatial resolution of the feature maps. At last, a prediction head is employed to produce the predicted saliency map $\hat{S}_0$. The entire network incorporates 4 layers of the ResNet stage for feature encoding and 4 layers of the MAM stage for feature decoding.

For more robust multi-modal feature generation, two techniques in MAM are introduced: efficient spatio-temporal cross-attention and multi-modal interaction module.

Efficient Spatio-Temporal Cross-Attention. Given video features $\mathbf{f}_v^i \in \mathbb{R}^{T_v^i \times h_v^i \times w_v^i \times C_v^i}$, audio features $\mathbf{f}_a \in \mathbb{R}^{T_a \times h_a \times w_a \times C_a}$ and noise features $\mathbf{f}_s^i \in \mathbb{R}^{h_v^i \times w_v^i \times C_v^i}$, the video features and noise features are concatenated as a spatio-temporal noise feature map $[\mathbf{f}_v^i; \mathbf{f}_s^i] \in \mathbb{R}^{(T_v^i+1) \times h_v^i \times w_v^i \times C_v^i}$. The processed feature map is then fed into the multi-modal interaction module along with the

audio features to obtain $\mathbf{f}_{avs} \in \mathbb{R}^{(T_v^i+1)\times h_v^i \times w_v^i \times C_v^i}$. To introduce audio information in the saliency prediction, we design an efficient spatio-temporal cross-attention (ECA), which not only effectively reduces the computational overhead of the standard cross-attention [51], but also possesses spatio-temporal interactions between features.

For all processed features, they are first converted to 2D feature sequences in the spatio-temporal domain and obtained $Q = V = [\mathbf{f}_v^i; \mathbf{f}_s^i] \in \mathbb{R}^{(T_v^i+1)h_v^i w_v^i \times C_v^i}; K = \mathbf{f}_{avs}^i \in \mathbb{R}^{(T_v^i+1)h_v^i w_v^i \times C_v^i}$. We find that directly following the standard cross-attention would take a prohibitively large amount of memory due to the high spatio-temporal resolution of the feature map. Therefore, a spatio-temporal compression technique is employed that can effectively reduce the computational overhead without compromising performance, as:

$$[\mathbf{f}_v^i; \mathbf{f}_s^i] = \texttt{ECA}(QW_Q, \texttt{STC}(K)W_K, \texttt{STC}(V)W_V), \tag{7}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C_i \times C_i}$ are parameters of linear projections, $\texttt{STC}(\cdot)$ is the spatio-temporal compression, which is defined as:

$$\texttt{STC}(x) = \texttt{LN}(\texttt{Conv3d}(x)), \tag{8}$$

Here, the dimension of features is reduced by controlling the kernel size as well as the stride size of 3D convolution.

Multi-modal Interaction Module. To take full advantage of different modal features, we model the interaction between audio features $\mathbf{f}_a$, video features $\mathbf{f}_v^i$ and noise features $\mathbf{f}_s^i$ at each scale i to obtain a robust multi-modal feature representation. A straightforward approach would be to directly concatenate and aggregate $\mathbf{f}_a$, $\mathbf{f}_v^i$ and $\mathbf{f}_s^i$, but this fusion method does not acquire global correlations between various modalities. Therefore, an effective multi-modal interaction strategy is proposed to capture crucial audio-visual activity changes in the spatio-temporal domain. In specific, this process starts with convolution and upsampling on the audio features, which is to construct spatially size-matched feature triples $(\widetilde{\mathbf{f}}_a, \mathbf{f}_v^i, \mathbf{f}_s^i)$. Subsequently, the video features and noise features are concatenated to obtain spatio-temporal noise features $[\mathbf{f}_v^i; \mathbf{f}_s^i]$. This processed feature undergoes an element-wise product operation with the audio features, resulting in $\overline{\mathbf{f}}_{avs}^i$. Following this, we perform average activation along the temporal dimension over $\overline{\mathbf{f}}_{avs}^i$ to pool global temporal information into a temporal descriptor. For the indication of critical motion regions, a softmax function is applied to obtain a mask by highlighting the segments of the corresponding spatio-temporal audio features, which exhibit key audio-visual activity changes:

$$\begin{aligned} \widetilde{\mathbf{f}}_a &= \texttt{Conv}(\texttt{UpSample}(\mathbf{f}_a)), \\ \mathbf{f}_{avs}^i &= \texttt{softmax}(\texttt{Pool}([\mathbf{f}_v^i; \mathbf{f}_s^i] * \widetilde{\mathbf{f}}_a)) * \widetilde{\mathbf{f}}_a. \end{aligned} \tag{9}$$

where $\texttt{Conv}(\cdot)$, $*$, $\texttt{softmax}$ and $\texttt{Pool}$ denote the operations of convolution, element-wise product, softmax and average pooling, respectively.

4.3. Overall Training and Inference Algorithms

Training. A diffusion process is initiated to create noisy maps by introducing corruption to ground-truth saliency maps. To reverse this process, the Saliency-UNet is trained for saliency map denoising. The overall training procedure of DiffSal is outlined in Algorithm 1 in the Appendix. In detail, Gaussian noises are sampled following α_t in Eq. 1 and added to the ground-truth saliency maps, resulting in noisy samples. At each sampling step t, the parameter α_t is pre-defined by a monotonically decreasing cosine scheme, as employed in [26]. The standard mean square error serves as the optimization function to supervise the model training:

$$\mathcal{L} = \|S_0 - g_\psi(S_t, t, \mathbf{f}_a, \mathbf{f}_v)\|^2. \tag{10}$$

where S_0 and $g_\psi(S_t, t, \mathbf{f}_a, \mathbf{f}_v)$ denote the ground-truth and predicted saliency maps, respectively.

Inference. The proposed DiffSal engages in denoising noisy saliency maps sampled from a Gaussian distribution, and progressively refines the corresponding predictions across multiple sampling steps. In each sampling step, the Saliency-UNet processes random noisy saliency maps or the predicted saliency maps from the previous sampling step as input and generates the estimated saliency maps for the current step. For the next step, DDIM [47] is applied to update the saliency maps. The detailed inference procedure is outlined in Algorithm 2.

5. Experiments

Experiments are conducted on six audio-visual datasets. The following subsections introduce the implementation details and evaluation metrics. The experimental results are represented with analysis through ablation studies and comparison with state-of-the-art works.

5.1. Setup

Audio-Visual Datasets: Six audio-visual datasets in saliency prediction have been employed for the evaluation, which are: AVAD [38], Coutrot1 [13], Coutrot2 [14], DIEM [40], ETMD [31], and SumMe [22]. The significant characteristics of these datasets are elaborated below. (i) The AVAD dataset comprises 45 video clips with durations ranging from 5 to 10 seconds. These clips cover various audio-visual activities, such as playing the piano, playing basketball, conducting interviews, *etc*. This dataset contains eye-tracking data from 16 participants. (ii) The Coutrot1 and Coutrot2 datasets are derived from the Coutrot dataset. Coutrot1 consists of 60 video clips covering four visual categories: one moving object, several moving objects, landscapes, and faces. The corresponding eye-tracking data are

Table 1. Ablation of different components in DiffSal.

Method	Components		AVAD		ETMD	
	ECA	MIM	CC ↑	SIM ↑	CC ↑	SIM ↑
baseline			0.701	0.547	0.632	0.498
	✓		0.716	0.556	0.644	0.503
		✓	0.714	0.551	0.638	0.502
	✓	✓	**0.738**	**0.571**	**0.652**	**0.506**

Table 2. Ablation of video and audio modalities.

Model	AVAD		ETMD	
	CC ↑	SIM ↑	CC ↑	SIM ↑
Audio-Only	0.343	0.283	0.365	0.295
Video-Only	0.716	0.556	0.644	0.503
Ours	**0.738**	**0.571**	**0.652**	**0.506**

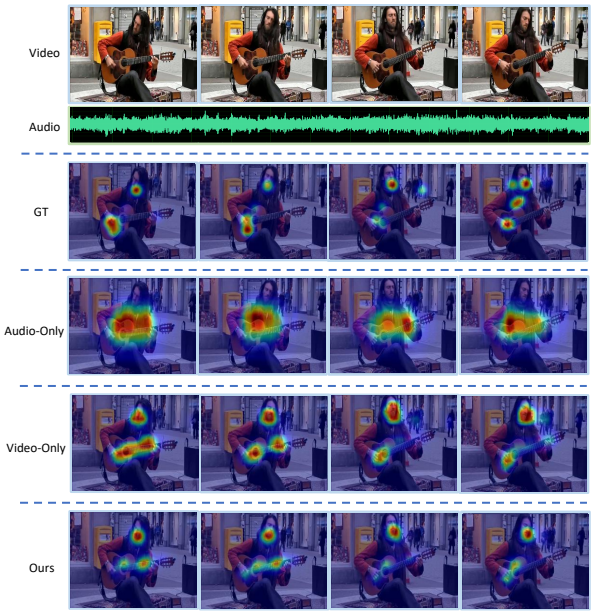

Figure 3. Visualizing the saliency results when different modalities are used. The audio-only approach can localize the sound source coming from the performer's guitar, while the video-only approach focuses on both the performer's face as well as the guitar.

obtained from 72 participants. Coutrot2 includes 15 video clips recording four individuals having a meeting, with eye-tracking data from 40 participants. (iii) The DIEM dataset contains 84 video clips, including game trailers, music videos, advertisements, *etc*., captured from 42 participants. Notably, the audio and visual tracks in these videos do not naturally correspond. (iv) The ETMD dataset comprises 12 video clips extracted from various Hollywood movies, with eye-tracking data annotated by 10 different persons. (v) The SumMe dataset consists of 25 video clips covering diverse topics, such as playing ball, cooking, travelling, *etc*.

Implementation Details: To facilitate implementation, a pre-trained MViTv2 [33] model on Kinetics [9] and a pre-trained VGGish [24] on AudioSet [19] are adopted. The input samples of the network consist of 16-frame video clips of size $224 \times 384 \times 3$ with the corresponding audio, which is transformed into 9 slices of 112×192 log-Mel spectrograms. For the spatio-temporal compression in efficient spatio-temporal cross-attention, the kernel size and stride size of the 3D convolution in the ith MAM stage are set to 2^i and 2^i, respectively. For a fair comparison, the video branch of DiffSal is pre-trained using the DHF1k dataset [53] following [58], and the entire model is fine-tuned on these audio-visual datasets using this pre-trained weight.

The training process chooses Adam as the optimizer with the started learning rate of $1e{-}4$. The computation platform is configured by four NVIDIA GeForce RTX 4090 GPUs in a distributed fashion, using Pytorch. The total sampling step T is defined as 1000 and the entire training is terminated within 5 epochs. The batch size is set to 20 across all experiments. During inference, the iterative denoising step is set to 4.

Evaluation Metrics: Following previous works, four widely-used evaluation metrics are adopted [8]: CC, NSS, AUC-Judd (AUC-J), and SIM. The same evaluation codes are used as in previous works [50, 58].

5.2. Ablation Studies

Extensive ablation studies are performed to validate the design choices in the method. The AVAD and ETMD datasets are selected for ablation studies, following the approach in [58].

Effect of Components of DiffSal. To validate the effectiveness of each module in the proposed framework, a baseline model is initially defined as the video-only version of DiffSal and replaces the multi-modal attention modulation in Saliency-UNet with a pure convolution operation. As shown in Table 1, the baseline model demonstrates a good performance that indicates a potential capability of the diffusion-based framework in the AVSP task. With the incorporation of the designed efficient spatio-temporal cross-attention (ECA), and the multi-modal interaction module (MIM) components, the overall performance of the model has been enhanced continually. As the core module, ECA presents a significant improvement in the CC metric by 0.015 on the AVAD dataset, and 0.012 on the ETMD dataset for the whole DiffSal framework. With the addition of the audio features and the MIM, the model also has another improvement of 0.022 in the CC metric on the AVAD dataset. All of these have demonstrated the effectiveness of ECA and MIM in the proposed DiffSal.

Effect of Video and Audio Modalities. Table 2 shows the contribution of spatio-temporal information from each modality in the Video and Audio Encoders to the overall performance. The experimental observations reveal that the video-only model exhibits significantly greater strength

Table 3. Ablation of different multi-modal interaction methods.

Method	AVAD		ETMD	
	CC ↑	SIM ↑	CC ↑	SIM ↑
DiffSal(w/ MIM)	**0.738**	**0.571**	**0.652**	**0.506**
w/ Bilinear	0.716	0.556	0.644	0.503
w/ Addition	0.706	0.543	0.606	0.464
w/ Concatenation	0.716	0.556	0.644	0.503

Table 4. Ablation of different cross-attention strategies. The computational cost is evaluated based on input audio of size $1 \times 9 \times 112 \times 192$ and video of size $3 \times 16 \times 224 \times 384$. # Params and #Mem denote the number of parameters and memory footprint of the model, respectively.

Attention	#Params	#Mem	AVAD		ETMD	
			CC ↑	SIM ↑	CC ↑	SIM ↑
SCA	76.43M	5.32G	0.713	0.531	0.628	0.476
ECA	76.57M	1.20G	**0.737**	**0.571**	**0.652**	**0.509**

than the spatio-temporal audio-only version, which verifies the essential role of the video modality. Figure 3 also visualizes the model predictions utilizing the two modal encoders separately. It is clear that either the audio-only or video-only approach can predict saliency areas in the scene, and the combination of the two modalities leads to more accurate predictions. This demonstrates the generalization of the DiffSal framework to audio-only, video-only, and audio-visual scenarios as well.

Effect of Different Cross-Attention Strategies. The design of efficient spatio-temporal cross-attention mechanism is further evaluated. As shown in Table 4, using efficient spatio-temporal cross-attention (ECA) not only leads to better performance but also greatly reduces the memory footprint of the model compared to using the standard cross-attention (SCA) strategy. This shows that the designed ECA can compress the effective spatio-temporal cues in the features and reduce the interference of irrelevant noises.

Effect of Different Multi-modal Interaction Methods. The effects of using different multi-modal interaction methods, such as Bilinear [49], Addition, and Concatenation, are compared in Table 3. These multi-modal interaction methods can be found in recent state-of-the-art works [30, 50], and the video features and noise features are firstly concatenated before their interaction with audio features. Experimental results show that the proposed MIM can outperform all the other three interaction methods, and obtain more robust multi-modal features. In contrast, the performance degradation of the other three methods suffers from the noise information embedded in the features.

Effect of Different Training Losses. Table 5 compares the impact on DiffSal using different loss functions from

Table 5. Ablation of different training losses.

Model	Losses			AVAD		ETMD	
	$\mathcal{L}_{CE}$	$\mathcal{L}_{KL}$	$\mathcal{L}_{MSE}$	CC ↑	SIM ↑	CC ↑	SIM ↑
Ours	✓			0.690	0.490	0.617	0.422
		✓		0.720	0.552	0.644	0.496
			✓	**0.738**	**0.571**	**0.652**	**0.506**

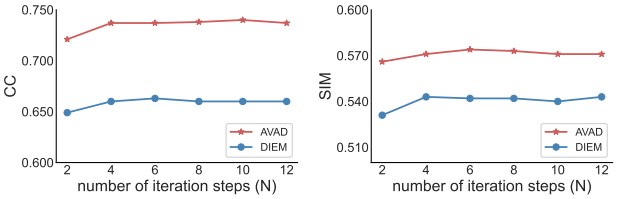

Figure 4. Performance analysis of denoising steps on AVAD and DIEM datasets.

previous state-of-the-art approaches [50, 58], such as the cross entropy (CE) loss $\mathcal{L}_{CE}$ and the Kullback-Leibler divergence (KL) loss $\mathcal{L}_{KL}$. During the training process, it is observed that the model with the $\mathcal{L}_{CE}$ converges slowly and yields sub-optimal performance. Compared to the $\mathcal{L}_{MSE}$, employing the $\mathcal{L}_{KL}$ can achieve acceptable results, but there is still a gap in the performance of training with the $\mathcal{L}_{MSE}$. This suggests that simple MSE loss can be used in the AVSP task as an alternative to these task-tailored loss functions.

Effect of Denoising Steps. The impact of the number of iterative denoising steps on the final performance is studied in Figure 4, which shows that more iteration steps result in better performance. With diminishing marginal benefits as the step number increases, a steady increase in performance is observed. For a linearly increasing of the computational cost with the step number, $N = 4$ is used to maintain a good balance between performance and computational cost.

5.3. Method Comparison

Comparisons with State-of-the-art Methods. As shown in Table 6, the experimental results of our DiffSal are compared with recent state-of-the-art works on six audio-visual saliency datasets. The table highlights the superiority of DiffSal, as it outperforms the other comparable works on all datasets by defined metrics. Notably, DiffSal significantly surpasses the previous top-performing methods, such as CASP-Net [58] and ViNet [30], and becomes the new state-of-the-art on these six benchmarks. The performance boost is very encouraging: DiffSal can achieve an average relative performance improvement of up to 6.3% compared to the second-place performer. Such substantial improvements validate the effectiveness of the diffusion-based approach as an effective audio-visual saliency prediction framework.

Qualitative Results. The ability of the model to handle

Table 6. Comparison with state-of-the-art methods on six audio-visual saliency datasets. **Bold** text in the table indicates the best result, and underlined text indicates the second best result. Our DiffSal significantly outperforms the previous state-of-the-arts by a large margin.

Method	#Params	#FLOPs	DIEM				Coutrot1				Coutrot2			
			CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J↑	SIM ↑	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet$_{\text{TPAMI'2019}}$ [54]	-	-	0.522	2.02	0.869	0.427	0.425	1.92	0.850	0.361	0.448	3.16	0.926	0.322
TASED-Net$_{\text{ICCV'2019}}$ [37]	21.26M	91.80G	0.557	2.16	0.881	0.461	0.479	2.18	0.867	0.388	0.437	3.17	0.921	0.314
STAViS$_{\text{CVPR'2020}}$ [50]	20.76M	15.31G	0.579	2.26	0.883	0.482	0.472	2.11	0.868	0.393	0.734	5.28	0.958	0.511
ViNet$_{\text{IROS'2021}}$ [30]	33.97M	115.31G	0.632	2.53	0.899	0.498	0.56	2.73	0.889	0.425	0.754	5.95	0.951	0.493
TSFP-Net$_{\text{arXiv'2021}}$ [10]	-	-	0.651	2.62	0.906	0.527	0.571	2.73	0.895	0.447	0.743	5.31	0.959	0.528
CASP-Net$_{\text{CVPR'2023}}$ [58]	51.62M	283.35G	0.655	2.61	0.906	0.543	0.561	2.65	0.889	0.456	0.788	6.34	0.963	0.585
Ours(DiffSal)	76.57M	187.31G	**0.660**	**2.65**	**0.907**	**0.543**	**0.638**	**3.20**	**0.901**	**0.515**	**0.835**	**6.61**	**0.964**	**0.625**
Method	**#Params**	**#FLOPs**	**AVAD**				**ETMD**				**SumMe**			
			CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J↑	SIM ↑	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet$_{\text{TPAMI'2019}}$ [54]	-	-	0.580	3.17	0.905	0.446	0.477	2.36	0.915	0.329	0.379	1.79	0.868	0.296
TASED-Net$_{\text{ICCV'2019}}$ [37]	21.26M	91.80G	0.601	3.16	0.914	0.439	0.509	2.63	0.916	0.366	0.428	2.1	0.884	0.333
STAViS$_{\text{CVPR'2020}}$ [50]	20.76M	15.31G	0.608	3.18	0.919	0.457	0.569	2.94	0.931	0.425	0.422	2.04	0.888	0.337
ViNet$_{\text{IROS'2020}}$ [30]	33.97M	115.31G	0.674	3.77	0.927	0.491	0.571	3.08	0.928	0.406	0.463	2.41	0.897	0.343
TSFP-Net$_{\text{arXiv'2021}}$ [10]	-	-	0.704	3.77	0.932	0.521	0.576	3.07	0.932	0.428	0.464	2.30	0.894	0.360
CASP-Net$_{\text{CVPR'2023}}$ [58]	51.62M	283.35G	0.691	3.81	0.933	0.528	0.620	3.34	0.940	0.478	0.499	2.60	0.907	0.387
Ours(DiffSal)	76.57M	187.31G	**0.737**	**4.22**	**0.935**	**0.571**	**0.652**	**3.66**	**0.943**	**0.506**	**0.572**	**3.14**	**0.921**	**0.447**

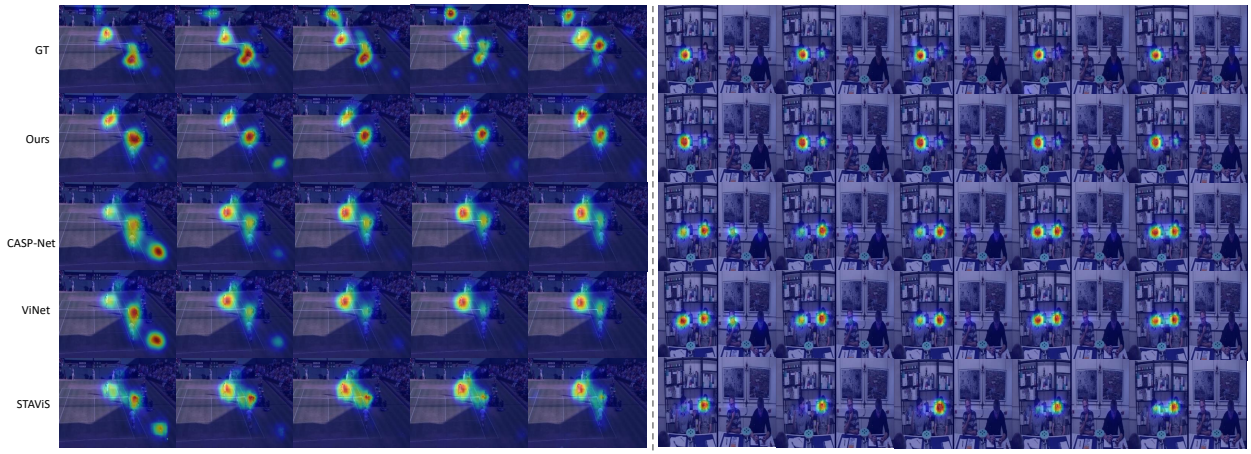

Figure 5. Qualitative results of our method compared with other state-of-the-art works. Challenging scenarios involving fast movement on the tennis court and multiple speakers indoors.

challenging scenarios, such as fast movement on the tennis court and multiple speakers indoors, is further examined. Figure 5 compares the DiffSal against other state-of-the-art approaches, such as CASP-Net [58], ViNet [30] and STAViS [50]. It is observed that DiffSal produces saliency maps much closer to the ground-truth for various challenging scenes. In contrast, CASP-Net focuses mainly on audio-visual consistency and lacks adopting an advanced network structure, leading to sub-optimal results. STAViS is only able to generate unsurprisingly saliency maps by employing sound source localization. More visualization results can be found in the supplementary.

Efficiency Analysis. Table 6 compares the number of parameters and computational costs of the DiffSal with previous state-of-the-art works. Compared to CASP-Net, the computational complexity of DiffSal is at a moderate level, even though incorporating Saliency-UNet in DiffSal leads to an increase in the number of model parameters. From a performance perspective, the DiffSal model achieves the best performance with the second-highest computational complexity.

6. Conclusion

In this work, we introduce a novel Diffusion architecture for generalized audio-visual Saliency prediction (DiffSal), formulating the prediction problem as a conditional generative task of the saliency map by utilizing input video and audio as conditions. The framework involves extracting spatio-temporal video and audio features from image sequences and corresponding audio signals. A Saliency-UNet is designed to perform multi-modal attention modulation, progressively refining the ground-truth saliency map from the noisy map. Extensive experiments have proven that DiffSal achieves superior performance compared to previous state-of-the-art methods in six challenging audio-visual benchmarks.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613*, 2023. 4

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[6] Yaniv Benny and Lior Wolf. Dynamic dual-output diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11482–11491, 2022. 2

[7] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *European Conference on Computer Vision*, pages 170–188. Springer, 2022. 2

[8] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 6

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[10] Qinyao Chang and Shiping Zhu. Temporal-spatial feature pyramid for video saliency detection. *arXiv preprint arXiv:2105.04213*, 2021. 1, 2, 8

[11] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 2, 3

[12] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2

[13] Antoine Coutrot and Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision*, 14(8):5–5, 2014. 5

[14] Antoine Coutrot and Nathalie Guyader. Multimodal saliency models for videos. In *From Human Attention to Computational Attention*, pages 291–304. Springer, 2016. 5

[15] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. Look around you: Saliency maps for omnidirectional images in vr applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017. 1

[16] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021. 2

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 4

[18] Richard Droste, Jianbo Jiao, and J Alison 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*, pages 419–435. Springer, 2020. 3

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 4, 6

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2, 4

[21] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. Diffusioninst: Diffusion model for instance segmentation. *arXiv preprint arXiv:2212.02773*, 2022. 3

[22] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 5

[23] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 2

[24] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 4, 6

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 5

[27] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 2

[28] Feiyan Hu, Simone Palazzo, Federica Proietto Salanitri, Giovanni Bellitto, Morteza Moradi, Concetto Spampinato, and Kevin McGuinness. Tinyhd: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation. In *Proceedings of the IEEE/CVF Winter Confer-*

ence on Applications of Computer Vision, pages 2051–2060, 2023. 1, 2, 3

[29] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3460, 2017. 1

[30] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet 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)*, pages 3520–3527. IEEE, 2021. 1, 2, 7, 8, 3

[31] Petros Koutras and Petros Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38: 15–31, 2015. 5

[32] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Vqbb: Image-to-image translation with vector quantized brownian bridge. *arXiv preprint arXiv:2205.07680*, 2022. 2

[33] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 4, 6

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 4

[35] Cheng Ma, Haowen Sun, Yongming Rao, Jie Zhou, and Jiwen Lu. Video saliency forecasting transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6850–6862, 2022. 1, 3

[36] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009. 1

[37] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2394–2403, 2019. 8, 3

[38] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(1):1–23, 2016. 1, 2, 5

[39] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020. 1, 2

[40] Parag K Mital, Tim J Smith, Robin L Hill, and John M Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, 3(1):5–24, 2011. 5

[41] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):773–786, 2018. 1

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[45] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 2

[46] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 1

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5

[48] Hamed R Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019. 2

[49] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 7

[50] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4766–4776, 2020. 1, 2, 6, 7, 8

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[52] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[53] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4894–4903, 2018. 6, 1

[54] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):220–237, 2019. 8

[55] Ziqiang Wang, Zhi Liu, Gongyang Li, Yang Wang, Tianhong Zhang, Lihua Xu, and Jijun Wang. Spatio-temporal self-attention network for video saliency prediction. *IEEE Transactions on Multimedia*, 2021. 2, 3

[56] Julia Wolleb, Robin Sandkühler, Florentin Bieder, and Philippe C Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022. 2

[57] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 4

[58] Junwen Xiong, Ganglai Wang, Peng Zhang, Wei Huang, Yufei Zha, and Guangtao 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*, pages 6441–6450, 2023. 1, 2, 6, 7, 8

[59] Hao Xue, Minghui Sun, and Yanhua Liang. Ecanet: Explicit cyclic attention-based network for video saliency prediction. *Neurocomputing*, 468:233–244, 2022. 3

[60] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 2

[61] Shunyu Yao, Xiongkuo Min, and Guangtao Zhai. Deep audio-visual fusion neural network for saliency estimation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1604–1608. IEEE, 2021. 1

[62] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2

[63] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022. 2

[64] Xiaofei Zhou, Songhe Wu, Ran Shi, Bolun Zheng, Shuai Wang, Haibing Yin, Jiyong Zhang, and Chenggang Yan. Transformer-based multi-scale feature integration network for video saliency prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2, 3

[65] Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018. 1