Learning Trimodal Relation for Audio-Visual Question Answering with Missing Modality

Kyu Ri Park[1], Hong Joo Lee[2,3†], and Jung Uk Kim[1†]

[1] Kyung Hee University, Yong-in, South Korea
{kyuri0924, ju.kim}@khu.ac.kr
[2] Technical University of Munich, Munich, Germany,
[3] Munich Center for Machine Learning (MCML), Munich, Germany
hongjoo.lee@tum.de

Abstract. Recent Audio-Visual Question Answering (AVQA) methods rely on complete visual and audio input to answer questions accurately. However, in real-world scenarios, issues such as device malfunctions and data transmission errors frequently result in missing audio or visual modality. In such cases, existing AVQA methods suffer significant performance degradation. In this paper, we propose a framework that ensures robust AVQA performance even when a modality is missing. First, we propose a Relation-aware Missing Modal (RMM) generator with Relation-aware Missing Modal Recalling (RMMR) loss to enhance the ability of the generator to recall missing modal information by understanding the relationships and context among the available modalities. Second, we design an Audio-Visual Relation-aware (AVR) diffusion model with Audio-Visual Enhancing (AVE) loss to further enhance audio-visual features by leveraging the relationships and shared cues between the audio-visual modalities. As a result, our method can provide accurate answers by effectively utilizing available information even when input modalities are missing. We believe our method holds potential applications not only in AVQA research but also in various multi-modal scenarios. The code is available at https://github.com/VisualAIKHU/Missing-AVQA.

Keywords: Missing modality · Audio-Visual Question Answering · Diffusion Model

1 Introduction

In the era of artificial intelligence, research efforts aimed at understanding scenes by integrating multi-modal information have made significant progress. A notable example in this field is Audio-Visual Question Answering (AVQA), which integrates video, audio, and text inputs to comprehend complex situations and generate responses by assimilating relevant video and audio information based on the questions. It involves extracting salient information from inputs and training networks to identify correlated features for accurate prediction.

† Corresponding author

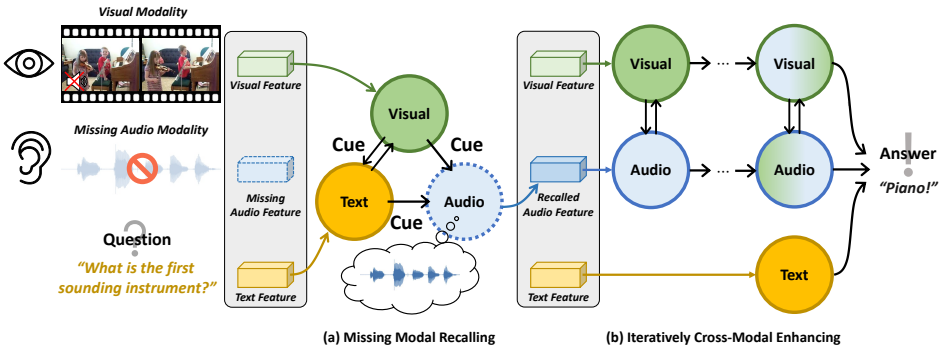

Fig. 1: Concept diagram of our methodology. Leveraging mutual cues in trimodal relations to recall and enhance missing information.

Due to the importance of AVQA, many recent studies have been proposed. Li et al. [21] introduced a large dataset named MUSIC-AVQA [21] and a spatio-temporal based audio-visual network tailored for the AVQA task. Yun et al. introduced Pano-AVQA [46] that focuses on AVQA within panoramic videos. Pano-AVQA established a new benchmark comprising spherical spatial relation QAs and audio-visual relation QAs. Transformer-based models trained on Pano-AVQA exhibit an enhanced semantic understanding of panoramic surroundings. Yang et al. proposed HAVF [44], a study focused on developing a hierarchical audio-visual fusing module to model semantic correlations among audio, visual, and text modalities. Li et al. proposed a Progressive Spatio-Temporal Perception Network (PSTP-Net) [20] that progressively identifies key spatio-temporal regions relevant to a question.

Although aforementioned methods have led to significant improvements, a critical issue remains unaddressed: the inherent reliance of existing AVQA approaches on complete input modalities. In real-world scenarios, it is common for certain modalities to be unavailable due to device malfunctions [8, 25] or environmental factors [1, 23, 24, 30], such as low-light conditions or noisy surroundings. Therefore, the limitations of current AVQA methods become evident when dealing with incomplete input modalities, leading to reduced performance and inaccurate responses. Some previous works tried to handle missing modality issues by generating pseudo features [19, 41, 42]. However, these approaches face challenges when applied to the AVQA task due to its unique complexities. Existing methods have primarily tackled the problem by handling modality in one-to-one pairs [23], neglecting the inter-dependencies among different modalities. Consequently, they can only generate pseudo features corresponding to the given modality, failing to consider the broader multi-modal context. However, the AVQA task poses a more intricate challenge, necessitating visual-audio reasoning [3, 26, 31, 48] across diverse contexts to answer complex questions effectively. AVQA requires a nuanced understanding of the context of question to generate relevant information for accurate responses, making it essential to flexibly generate pseudo features [41] for missing modalities depending on the specific question.

In this paper, we propose a new approach with a novel AVQA framework to address the issue of missing modalities. This approach is based on analyzing human cognitive psychology [22, 27, 39], specifically the ability to recall information through audio-visual integration [2, 11, 14, 34, 40]. Fig. 1 shows the conceptual diagram of our method. When only one modality is available (*e.g.,* visual), humans can retrieve associated information of another modality (*e.g.,* the sound of a piano) through the integration of visual-text cues. We propose a Relation-aware Missing Modal (RMM) generator. This generator takes visual-text or audio-text information as input. By associating the available modalities, the RMM generator recalls the missing modality and derives a pseudo feature for it. By determining the correlation between the two modalities, we can effectively identify the missing modality.

Next, we introduce an Audio-Visual Relation-aware (AVR) diffusion model to enhance the features generated by the RMM generator with mutual cues. The real feature of the available modality (visual) and the pseudo feature of the missing modality (audio) are combined and passed into the AVR diffusion model. This process enhances the features by leveraging mutual cues from the different modalities. Then, the output of the diffusion becomes the input for the AVQA predictor in place of the missing modality. These two steps are designed with new insights to mimic the two steps of human trimodal interpretation. The RMM uses trimodal relations for missing modalities, and the AVR enhances features using complementary audio-visual relations. This approach ensures that the enhanced features from both modalities contribute to more accurate and robust predictions, improving the overall performance of the AVQA system.

The main contributions of this paper are summarized as follows:

- We introduce a novel AVQA framework that simultaneously addresses the issue of missing modalities.
- We propose the Relation-aware Missing Modal (RMM) generator, a new approach to recall the information of the missing modality by associating two existing modalities and deriving a pseudo feature.
- We present the Audio-Visual Relation-aware (AVR) diffusion model, which emphasizes the mutual enhancement of modalities by effectively utilizing and referencing information from each other.

2 Related Work

2.1 Audio-Visual Question Answering

Recently, several works have used audio, visual, and text modalities for multimodal scene understanding. Schwartz et al. [38] proposed a baseline for audio-visual scene understanding, consisting of feature extractors, a multi-modal attention module, and an answer generation module. Yun et al. introduced the Pano-AVQA network [46] for semantic scene understanding in panoramic videos, proposing spherical spatial embedding methods and using equirectangular and NFoV [10] projections to reduce visual distortion during feature extraction. Li

et al. [21] developed the MUSIC-AVQA dataset for AVQA tasks in musical performance scenes, offering more complex relations such as existential, location, counting, comparative, and temporal information, and proposed spatial and temporal grounding networks. Li et al. [20] also proposed PSTP-Net, which finds key spatio-temporal regions from video using a temporal segment selection module, spatial regions selection module, and audio-guided visual attention module.

Most of these works assumed the completeness of modality which may not considered missing modality. In real-world situations, some modalities can be missed due to device malfunctions or environmental constraints [1, 23, 24, 30]. Under these missing modalities situations, AVQA systems can become unstable and fail to work properly. Unlike previous works, this paper focuses on handling the problem of missing modalities.

2.2 Missing Modality in Multi-modal Learning

Recently, some works have addressed the missing modality problem in multi-modal learning [13, 17–19, 41, 42]. Woo et al. [42] investigated the effects of architecture, data augmentation, and regularization under missing modalities and proposed an Action Masked Auto Encoder (ActionMAE) that generates pseudo features of missing modalities for inference. Wang et al. proposed Shared-Specific Feature Modeling (ShaSpec) [41], which extracts shared and modality-specific features to enhance input data representation using shared and specific encoders. Lee et al. [19] introduced a missing-aware prompting method for transformer models that helps the model be aware of the missing modality by attaching missing-aware prompts at the input. Woo et al. [42] also proposed a transformer architecture to fuse features from all modalities into a comprehensive set using hybrid modality-specific encoders, intra-modal transformers, and inter-modal transformers. In the context of autonomous driving, Choi et al. [4] proposed a Shared Cross-modal Embedding method to encode features effectively, addressing missing modality issues. Additionally, Wu et al. [43] addressed the missing modality problem using a knowledge distillation method with a vision teacher, an auditory teacher, and an audio-visual student. These studies have demonstrated significant success in addressing the missing modality problem. However, their applicability to the AVQA task is limited due to its unique complexities. Existing methods primarily handle modality in one-to-one pairs, overlooking the interdependencies among different modalities. For example, [42] pairs input images, depth images, and IR images to generate estimated features. In contrast, the AVQA task demands a nuanced understanding of the question context to produce relevant information, necessitating the flexible generation of pseudo features for missing modalities based on the specific question. Consequently, this paper proposes a method to tackle the missing modality problem more effectively by comprehending the context of questions and generating missing modality information that corresponds to the context of the questions.

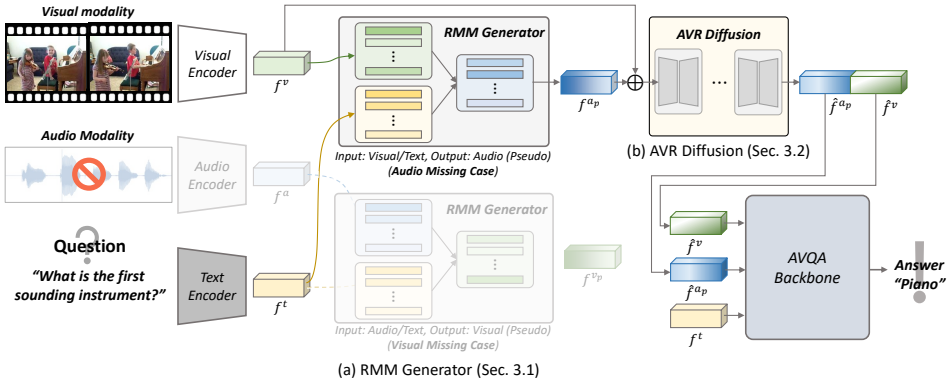

Fig. 2: Overall architecture of the proposed AVQA framework for missing modality (audio missing example). We introduce (a) Relation-aware Missing Modal (RMM) generator and (b) Audio-Visual Relation-aware (AVR) diffusion model. More details for learning RMM generator and AVR diffusion are in Sec. 3.1 and Sec. 3.2.

2.3 Diffusion Based Models

The diffusion model generates data through an iterative process of adding and removing noise. It consists of a forward process that adds noise at each time step and a reverse process that removes noise at each time step. Diffusion models represent a promising group of generative models that simplify the data generation process into a step-by-step noise reduction technique [7]. Diffusion models have been shown to perform well in several image generation tasks [5], including super-resolution [32, 37, 47], effective image restoration [9], image processing [12, 28], text conditioning [6, 29, 35], image inpainting [36], and more. Stable diffusion [35] built a DPM (Diffusion Probabilistic Model) in latent space to reduce the number of pixels. Most of these previous works on diffusion take only one modality as input to the diffusion process and derive the output of the same modality. In contrast, our AVR diffusion model integrates audio and visual modalities simultaneously, ensuring effective fusion and thorough learning of the diffusion process for enhanced feature extraction.

3 Methodology

Fig. 2 shows the overall architecture of the proposed AVQA framework addressing missing modalities during inference. The visual modal input, audio input, and question pass through their corresponding encoders to obtain f^v, f^a, f^t. Our method consists of three major components: (1) Relation-aware Missing Modal (RMM) generator (see Fig. 2 (a)), (2) Audio-Visual Relation-aware (AVR) diffusion model (see Fig. 2 (b)), and (3) AVQA backbone. For example, in the case of audio being missing, the RMM generator generates a pseudo audio feature for the missing modality (*e.g.,* audio) using the existing modalities (*e.g.,* visual and question). Since our work addresses missing scenarios during the inference phase, the RMM generator is trained to learn pseudo audio features that closely

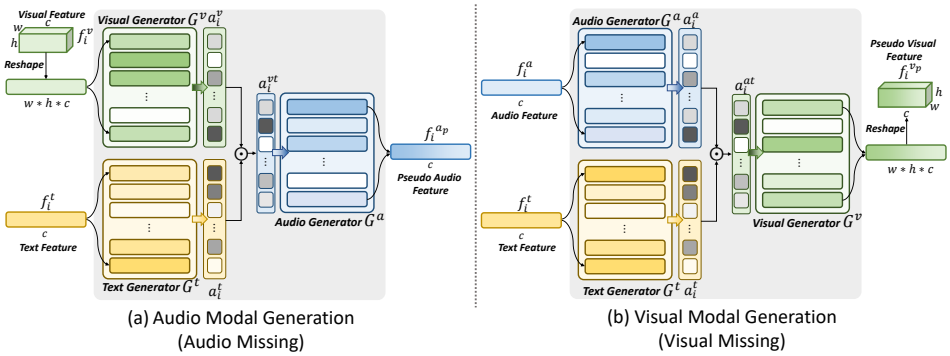

Fig. 3: RMM generator operates in two scenarios: (a) generating pseudo audio when audio input is missing, and (b) generating pseudo visuals when visual input is missing. Based on the addressing vector (*i.e.*, a_i^{vt}, a_i^{at}), pseudo modality feature is obtained by a weight summation with the corresponding generator. Each generator in the three modalities shares weights.

resemble the real audio features. Then, with the pseudo audio feature and the visual modal feature, the AVR diffusion model enhances the feature representation of each modality by referring to cross-modal knowledge. To train the AVR diffusion model, real features f^v and f^a are utilized during training. Finally, the features of the two modalities obtained through the AVR diffusion model, along with the question feature, are passed through the AVQA backbone to estimate the answer. A similar process is employed to account for the missing visual modality.

Our novelty consists of two key steps: (1) using trimodal relations to handle missing modalities and (2) enhancing features with complementary audio-visual relations. The RMM generator leverages trimodal relations to generate pseudo features for the missing modality. Then, the AVR diffusion model uses cross-modal knowledge to enhance the feature representation of each modality, combining the pseudo audio feature with the visual feature. By doing so, even when a single-modal input is missing during the inference phase, our RMM generator and AVR diffusion model can recall the missing modal information and improve feature representation, showing robust performance for AVQA.

3.1 Relation-aware Missing Modal Generator

The Relation-aware Missing Modal (RMM) generator is inspired by the remarkable ability of human brain to integrate and process different types of information from various sensory modalities. When humans encounter content, they can often infer missing audio information by analyzing visual cues and context, and similarly, infer visual information from audio cues. This capability arises from shared cues between different modalities that come from the same object or scene. For instance, when only auditory information of piano sound is provided, humans can perceive the sound and recall the image of a piano (visual information). Conversely, when watching a muted video of a piano performance, humans can recall the piano sound (auditory information).

The RMM generator mimics this cognitive mechanism by leveraging the relationships between visual, auditory, and textual modality to effectively handle multi-modal inputs. It uses a slot-based architecture, where each modality is represented by L learnable parameter vectors that capture essential features. These slots are crucial for recalling and reconstructing missing information. By analyzing spatio-temporal patterns in the input modality, the RMM generator identifies correlations between modalities, enabling it to predict and restore missing information precisely. This approach mirrors the interconnected cortical regions of brain specialized in visual, auditory, and language processing, enhancing the ability of model to synthesize and interpret diverse modality [16].

Fig. 3 (a) illustrates the scenario when the audio modality is missing [30]. The visual feature f_i^v and the text feature f_i^t are passed through RMM generator. Particularly, f_i^v is flattened to create a vector $\bar{f}_i^v \ 2 \ \mathsf{R}^{1\times whc}$. RMM generator is composed of three modal generators. Each generators consists of L slots denoted as $\mathsf{G} = f\mathsf{G}_j^v, \mathsf{G}_j^a, \mathsf{G}_j^t g_{j=1}^L$ ($\mathsf{G}_j^v \ 2 \ \mathsf{R}^{1\times whc}, \mathsf{G}_j^a, \mathsf{G}_j^t \ 2 \ \mathsf{R}^{1\times c}$). These slots represent the number of learnable parameter vectors for each modality in the RMM generator, facilitating the generation of missing features by establishing relationships between modalities. Each slot contains data that serves as a reminder of the corresponding modality. When the audio modality is missing, the visual generator G^v and text generator G^t use $\bar{f}_i^v$ and f_i^t to determine how much of the L slots from the auditory generator G^a are needed to recall the audio information. That is, $\bar{f}_i^v$ and L slots of G^v are calculated to obtain visual addressing vector $a_i^v = f a_{i1}^v, \ldots, a_{iL}^v g \ 2 \ \mathsf{R}^{1\times L}$. Each element of a_i^v is calculated as:

$$a_{ij}^v = \frac{\exp\left(s_{ij}^v\right)}{\sum_{m=1}^{L} \exp\left(s_{im}^v\right)}, \ s_{ij}^v = \frac{\bar{f}_i^v \, p \, \mathsf{G}_j^{v\top}}{\overline{c}}. \tag{1}$$

A high value of a_{ij}^v means a strong correlation between $\bar{f}_i^v$ and G_j^v. Else, they are weakly correlated. Similar to Eq. (1), text addressing vector $a_i^t = f a_{i1}^t, \ldots, a_{iL}^t g \ 2 \ \mathsf{R}^{1\times L}$ is obtained using f_i^t and G^t. Next, visual-text addressing vector a_i^{vt} is computed as $a_i^{vt} = softmax(a_i^v \quad a_i^t)$ (is element-wise multiplication). We designed a_i^{vt} to highlight the strongly correlated audio slots among L slots by associating visual and textual modalities. Given a_i^v and a_i^t (each vector represents the association between each modal feature and L slots), we perform element-wise multiplication and softmax to generate a_i^{vt}. Then, a_i^{vt} and G^a are aggregated to generate pseudo audio feature $f_i^{a_p} \ 2 \ \mathsf{R}^{1\times c}$, which can be represented as:

$$f_i^{a_p} = \sum_{j=1}^{L} a_{ij}^{vt} \quad \mathsf{G}_j^a. \tag{2}$$

If the j-th element of a_{ij}^{vt} is high, the j-th slot of G_j^a will play a more important role in recalling the audio modal information. Likewise, Fig. 3 (b) illustrates the case where the visual modality is missing. The RMM generator accepts an audio feature f_i^a and a text feature f_i^t to produce a pseudo visual feature $f_i^{v_p}$ in $\mathsf{R}^{1\times w\times h\times c}$. This process is similar to that shown in Fig. 3 (a).

It is important to note that both scenarios, audio missing (Fig. 3(a)) and visual missing (Fig. 3 (b)), are considered during the training phase. Therefore,

the weight parameters of each G^v, G^a, and G^t in Fig. 3 (a) and (b) are shared. Through the aforementioned process, the RMM generator is able to generate the missing modality feature by associating the text with a remaining modality. The generated pseudo feature can takes the place of the missing modality.

Relation-aware Missing Modal Recalling Loss. The main purpose of designing our RMM generator is to effectively recall the missing modal information during inference. Therefore, we propose Relation-aware Missing Modal Recalling (RMMR) loss $\mathcal{L}_{rmmr}$ to guide the outputs of the RMM generator to closely resemble the actual modal features. $\mathcal{L}_{rmmr}$ consists of the two losses: audio recalling loss $\mathcal{L}_a$ for audio modal missing, visual recalling loss $\mathcal{L}_v$ for visual modal missing. $\mathcal{L}_a$ guides the pseudo audio feature $f_i^{a_p}$ from RMM generator to link the semantic knowledge with the real audio feature f_i^a. Likewise, $\mathcal{L}_v$ guides the pseudo visual feature $f_i^{v_p}$ to be similar to the real visual feature f_i^v. The $\mathcal{L}_a$ and $\mathcal{L}_v$ are formulated as follows:

$$\mathcal{L}_a = \frac{1}{N}\sum_{i=1}^{N} \left\| f_i^a \quad f_i^{a_p} \right\|_2^2, \;\; \mathcal{L}_v = \frac{1}{N}\sum_{i=1}^{N} \left\| f_i^v \quad f_i^{v_p} \right\|_2^2, \tag{3}$$

where N indicates the number of the batch size.

Finally, we propose Relation-aware Missing Modal Recalling (RMMR) loss $\mathcal{L}_{rmmr}$. It can be represented as:

$$\mathcal{L}_{rmmr} = \mathcal{L}_a + \mathcal{L}_v. \tag{4}$$

Through Eq. (4), our AVQA framework with RMM generator can effectively recall the missing modal information. As a result, our AVQA method can provide accurate answers to questions, even in situations where one of the modalities is missing during the inference phase.

3.2 Audio-Visual Relation-aware Diffusion Model

In this section, we introduce the proposed Audio-Visual Relation-aware (AVR) diffusion model. The goal of this model is to enhance the feature representation of both the missing modality (*e.g.,* audio) as well as the original counterpart modality (*e.g.,* visual). As illustrated in Fig. 4 (a), the process begins by combining the real audio feature (*i.e.,* f^a) and the visual feature (*i.e.,* f^v) through concatenation, resulting in the combined feature (*i.e.,* f^{av}). This combined feature is then passed through a diffusion process, which includes a forward process q (adding noise) and a reverse process p_θ. In the reverse process, p_θ estimates the steps of the forward process in reverse, using the weight parameter θ of the autoencoder [7]. This allows the AVR diffusion model to learn how to recover the original data from the noise effectively. The forward process q and reverse process p_θ at time step $t \in [0, T]$ is defined as:

$$q(f_t^{av} | f_{t-1}^{av}) = \mathcal{N}(f_t^{av}; \sqrt{1 \quad \beta_t} f_{t-1}^{av}, \beta_t I), \;\; t \in [1, T], \tag{5}$$

$$p_\theta(f_{t-1}^{av} | f_t^{av}) = \mathcal{N}(f_{t-1}^{av}; \mu_\theta(f_t^{av}, t), \Sigma_\theta(f_t^{av}, t)), \tag{6}$$

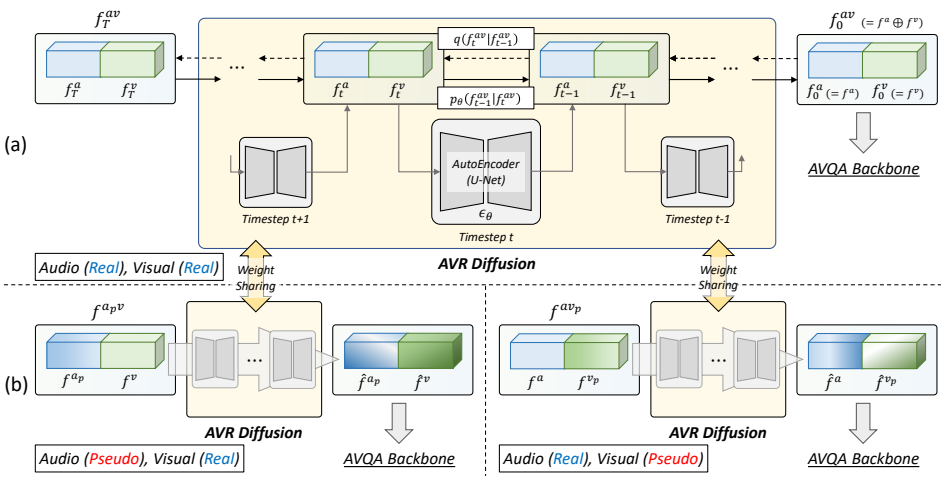

Fig. 4: AVR diffusion process illustrates how the model learns to generate enhanced features for both audio and visual modalities by leveraging cross-modal knowledge. (a) depicts the diffusion and reverse process of concatenated features between real features, while (b) represents the reverse process of features where the pseudo feature and the real feature are concatenated.

where β_t indicates hyper-parameters to control amount of noise at time step t, $\mu_\theta(f_t^{av}, t)$ and $\Sigma_\theta(f_t^{av}, t))$ denote mean and variance, respectively. In this process, combining features of audio and visual modalities enables mutual information utilization, resulting in enhanced feature representations. This process is repeated T timesteps. As a result, our AVR diffusion has learned the ability to enhance feature representations for a given input.

Since we aim to address the missing modalities (audio or visual) in the inference phase, we also combine (f^{a_p}, f^v) to generate $f^{a_p v}$ for audio missing (see Fig. 4 (b) left) and (f^a, f^{v_p}) to generate f^{av_p} for visual missing (see Fig. 4 (b) right). $f^{a_p v}$ and f^{av_p} go through the reverse process of AVR diffusion to produce $\hat{f}^{a_p v}$ and $\hat{f}^{av_p}$. Finally, after leveraging cross-modal knowledge through AVR diffusion, the audio and visual features are separated, *i.e.*, ($\hat{f}^a$, $\hat{f}^{v_p}$) and ($\hat{f}^{a_p}$, $\hat{f}^v$) to use each individual feature as an input to the AVQA backbone. Note that the existing AVQA networks [20, 21, 38, 46] require individual inputs (audio, visual, question (text)) to answer the questions.

The role of our AVR diffusion can be highlighted as follows: (1) Through the diffusion process, AVR diffusion learns the ability to generate the enhanced features for both audio-visual modalities by jointly leveraging cross-modal knowledge of f^{av}. (2) Also, the features of the pseudo modality and the counterpart original modality are combined and passed through AVR diffusion to further enhance their representations, considering the missing case in the inference phase.

Audio-Visual Enhancing Loss. To guide AVR diffusion can perform the aforementioned roles, we introduce the Audio-Visual Enhancing (AVE) loss L_{ave},

which can be represented as:

$$L_{ave} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}[\|\epsilon - \hat{\epsilon}_\theta(f_t^{av}, t)\|_2^2], \tag{7}$$

where ϵ denote noise in normal distribution $\mathcal{N}(0, I)$ and $\hat{\epsilon}_\theta(f_t^{av}, t)$ denotes the prediction of the autoencoder of AVR diffusion at the t-th time step.

3.3 Total Loss

To learn the AVQA network to perform robustly in missing modality scenarios, the total loss function is defined as follows:

$$\begin{aligned} L_{avqa} = L_{ce}(f^a, f^v, f^t) + L_{ce}(\hat{f}^{a_p}, \hat{f}^v, f^t) + L_{ce}(\hat{f}^a, \hat{f}^{v_p}, f^t), \\ L_{Total} = L_{avqa} + \lambda_1 L_{rmmr} + \lambda_2 L_{ave}, \end{aligned} \tag{8}$$

where λ_1,λ_2 denote the balancing hyper-parameters, L_{avqa} denotes the loss function of the AVQA predictors [20, 21, 38, 46], and L_{ce} indicates the cross-entropy loss when (f^a, f^v, f^t), $(\hat{f}^{a_p}, \hat{f}^v, f^t)$ and $(\hat{f}^a, \hat{f}^{v_p}, f^t)$ pairs are input to AVQA, respectively. With L_{avqa} and the proposed two loss functions (L_{rmmr} and L_{ave}), our AVQA framework can provide more accurate answers to questions even in the missing modality situations.

4 Experimental Results

4.1 Experimental Settings

Dataset. In the experiments, two publicly available open-source datasets, namely MUSIC-AVQA [21] and AVQA [44], were utilized. The MUSIC-AVQA dataset serves as an AVQA benchmark for comprehensive scene understanding in musical performance. It encompasses 45,867 question-answer pairs derived from 9,288 videos. To elaborate further, 32,087, 4,595, and 9,185 question-answer pairs were allocated for training, validation, and testing, respectively. Similarly, the AVQA dataset is a large-scale AVQA dataset designed for reasoning about multiple audio-visual relationships in real-life scenarios. It comprises 57,335 question-answer pairs sourced from 57,015 videos. In accordance with the specifications provided in [44], 34,401, 5,734, and 17,200 question-answer pairs were earmarked for training, validation, and testing, respectively.

AVQA Network. To verify the capability of the proposed method, we adopted our method to four recently introduced AVQA networks (AVST [21][1], AVSD [38][2], Pano-AVQA [46][3], and PSTP-Net [20][4]) for which official code is available. Note that, all our implements are conducted by referring to the official codes. In

[1] https://github.com/GeWu-Lab/MUSIC-AVQA
[2] https://github.com/idansc/simple-avsd
[3] https://github.com/HS-YN/PanoAVQA
[4] https://github.com/GeWu-Lab/PSTP-Net

Table 1: Results on MUSIC-AVQA dataset under missing modalities (visual missing (upper), audio missing (lower)) (V: visual, A: audio, Q: question). * denotes a network that needs additional information (*i.e.,* CLIP) for AVQA task.

Method	Scenario	Audio Question			Visual Question			Audio-Visual Question						All
		Cnt.	Comp	Avg	Cnt.	Loc	Avg	Exist	Loc	Cnt.	Comp	Temp	Avg	Avg
AVSD [38]	Input: $A+Q$ (V: Missing)	68.10	61.56	65.68	59.22	55.54	57.35	72.36	56.51	52.10	57.76	50.61	58.11	59.25
AVSD + Ours		80.33	64.31	74.43	74.10	64.41	69.20	79.25	70.91	57.83	65.40	58.88	67.03	68.91
Pano-AVQA [46]		65.66	48.92	59.46	51.79	51.47	51.63	49.95	53.14	46.93	52.65	34.34	48.28	51.14
Pano-AVQA + Ours		79.15	64.14	73.62	72.51	62.20	67.30	79.15	70.51	59.57	63.49	55.84	66.33	67.87
AVST [21]		68.10	61.23	65.56	61.80	51.22	56.45	80.70	55.10	48.00	60.09	45.87	58.38	59.14
AVST + Ours		78.27	67.17	74.18	72.10	64.08	68.04	77.94	67.43	58.48	65.21	58.88	65.99	67.98
PSTP-Net* [20]		70.40	62.79	67.60	58.06	51.59	54.79	80.16	55.81	46.74	58.58	51.34	58.77	59.27
PSTP-Net* + Ours		76.60	65.49	72.50	68.50	62.86	65.65	82.79	63.95	57.50	61.13	57.66	64.82	66.39
AVSD [38]	Input: $V+Q$ (A: Missing)	38.45	55.74	44.86	42.20	37.30	39.72	58.09	44.74	27.08	49.78	15.41	40.53	41.08
AVSD + Ours		79.84	64.81	74.30	74.94	69.39	72.13	79.25	71.15	59.02	65.03	60.22	67.45	69.90
Pano-AVQA [46]		39.24	57.57	46.03	43.45	35.50	39.43	80.10	39.17	26.43	49.87	16.99	43.56	42.91
Pano-AVQA + Ours		79.45	64.48	73.93	74.85	69.80	72.30	77.02	71.46	58.59	64.21	60.95	69.95	69.90
AVST [21]		29.55	54.41	38.76	35.03	27.61	31.27	58.29	41.37	21.90	49.60	13.47	38.44	36.60
AVST + Ours		79.84	64.81	74.30	74.77	72.24	73.49	78.74	70.59	58.15	62.67	60.46	66.44	69.71
PSTP-Net* [20]		78.76	55.72	70.27	75.52	71.35	73.41	77.73	68.54	54.67	59.85	58.03	64.25	67.74
PSTP-Net* + Ours		81.02	60.94	73.62	78.20	77.47	77.83	80.16	71.38	61.63	62.03	62.77	67.92	71.55

our study, we selected AVST [21] as our baseline, given its status as the latest network to exclusively utilize audio, visual, and text inputs. Given that PSTP-Net [20] integrates additional clip [33] features alongside these three modalities, it was not considered for our baseline. As such, AVST was used for conducting our ablation study experiments, being acknowledged as the most contemporary approach, except PSTP-Net.

Implementation Details. We train our AVQA framework on a single RTX 4090 GPU with a batch size of 4, utilizing the Adam optimizer [15] with an initial learning rate of 10^{-4}. For the RMM Generator, we use $L = 75$, and the default number of timesteps for the forward and reverse process of diffusion is 10. In our experiments, we set $\lambda_1 = \lambda_2 = 1$.

4.2 Evaluation Under Missing Modality.

Results on MUSIC-AVQA Dataset. We adopt the four state-of-the-art AVQA networks, *i.e.*, AVSD [38], Pano-AVQA [46], AVST [21], and PSTP-Net [20] on MUSIC-AVQA dataset [21] to demonstrate the ability of our method in handling missing modalities. As shown in Table 1, when the visual modality is missing, existing AVQA methods struggle to estimate answers, achieving around 51 59%. On the other hand, applying our approach to the existing AVQA networks significantly improves accuracy. Furthermore, our proposed method exhibits even more substantial improvements when the audio modality is missing. The results verify the effectiveness of our method in the missing modalities.

Results on AVQA Dataset. Furthermore, we extended our experiments to the AVQA dataset [44]. Table 2 shows the results on the AVQA dataset. Notably, our method remains effective even when one modality is missed. These results demonstrate the efficacy of our proposed approach in compensating for missing modalities. Furthermore, our method exhibits flexibility enough, as it can seamlessly integrate into various existing AVQA network architectures.

Table 2: Results on AVQA dataset under missing modalities (visual missing (upper), audio missing (lower)) (V: visual, A: audio, Q: question).

Method	Scenario	Question Type								All Avg
		Used	When	Before Next	Why	Where	Happen	Come From	Which	
AVSD [38]	Input: $A+Q$ (V: Missing)	52.94	42.86	56.00	56.98	42.39	59.53	44.65	40.36	45.80
AVSD + Ours		39.63	45.45	59.41	47.11	58.14	68.00	42.86	64.71	46.35
Pano-AVQA [46]		41.18	28.57	58.00	52.33	27.52	45.12	34.40	30.35	34.31
Pano-AVQA + Ours		58.82	42.86	66.00	52.33	45.65	59.82	44.55	39.58	45.94
AVST [21]		35.29	23.81	56.00	54.65	27.95	41.07	33.98	28.83	32.84
AVST + Ours		58.82	42.86	64.00	47.67	46.09	59.11	46.16	40.56	46.65
AVSD [38]	Input: $V+Q$ (A: Missing)	52.94	42.86	56.00	55.81	42.49	59.50	44.63	40.30	45.77
AVSD + Ours		39.63	45.45	59.41	47.11	58.14	68.00	42.86	64.71	46.38
Pano-AVQA [46]		70.59	52.38	68.00	58.14	53.77	64.64	57.60	51.71	56.32
Pano-AVQA + Ours		64.71	61.90	74.00	55.81	72.44	71.78	72.35	70.08	71.28
AVST [21]		52.94	47.62	62.00	56.98	55.08	65.06	58.58	52.39	57.03
AVST + Ours		70.59	57.14	70.00	59.30	72.34	72.96	73.77	65.86	70.28

Table 3: Comparison of our method with recent approaches for handling missing modalities in the MUSIC-AVQA dataset (visual missing (upper), audio missing (lower)) (V: visual, A: audio, Q: question). We adopt AVST, denoted as B, for the baseline of AVQA task. **Bold**/underlined fonts indicate the best/second-best results.

Method	Scenario	Audio Question			Visual Question			Audio-Visual Question						All Avg
		Cnt.	Comp	Avg	Cnt.	Loc	Avg	Exist	Loc	Cnt.	Comp	Temp	Avg	
AVST (B) [21]	Input: $V+Q$ (A: Missing)	29.55	54.41	38.76	35.03	27.61	31.27	58.29	41.37	21.90	49.60	13.47	38.44	36.60
B+Lee et al. [19] (CVPR'23)		69.91	63.47	67.54	58.40	55.35	56.85	80.67	57.23	47.93	62.03	47.81	59.62	60.28
B+ShaSpec [41](CVPR'23)		76.99	59.76	70.64	72.35	66.69	69.49	79.15	66.09	53.26	61.22	56.20	63.66	66.44
B+Woo et al. [42] (AAAI'23)		77.29	63.64	72.25	72.26	67.59	69.90	79.05	68.06	55.76	61.13	56.57	64.62	67.37
B+Yao et al. [45] (AAAI'24)		77.09	59.43	70.58	72.43	67.10	69.74	79.15	65.69	53.04	61.94	56.20	63.68	66.50
B + Ours		79.84	64.81	74.30	74.77	72.24	73.49	78.74	70.59	58.15	62.67	60.46	66.44	69.71
AVST (B) [21]	Input: $A+Q$ (V: Missing)	68.10	61.23	65.56	61.80	51.22	56.45	80.70	55.10	48.00	60.09	45.87	58.38	59.14
B+Lee et al. [19] (CVPR'23)		71.39	64.14	68.72	63.32	59.18	61.23	81.38	60.32	53.59	61.04	54.62	62.42	63.22
B+ShaSpec [41] (CVPR'23)		77.58	67.51	73.87	72.43	64.00	68.17	78.04	66.80	57.83	60.67	55.96	64.29	67.01
B+Woo et al. [42] (AAAI'23)		78.66	65.82	73.93	71.60	65.31	68.41	78.54	65.77	57.39	61.58	56.20	64.29	67.08
B+Yao et al. [45] (AAAI'24)		77.78	67.17	73.87	72.51	64.73	68.58	77.83	66.96	58.59	59.49	57.42	64.40	67.18
B + Ours		78.27	67.17	74.18	72.10	64.08	68.04	77.94	67.43	58.48	65.21	58.88	65.99	67.98

4.3 Comparison with Existing Missing Modality Handling Methods.

We compare our method with the state-of-the-art methods [19, 41, 42, 45] that handle the missing modality on the MUSIC-AVQA dataset. We adopt AVST for the base AVQA backbone. As shown in Table 3, even in the visual modality missing and the audio modality missing, our method achieves the highest performance in overall accuracy (*i.e.,* 'All Avg' metric). The results show that even in the missing scenario, our RMM generator effectively recall the missing modal information. Also our AVR diffusion further enhances the feature representation of the audio-visual modality by leveraging the cross-modal relation.

4.4 Ablation Study

Effect of the RMM Generator and AVR Diffusion Model. We conducted experiments to evaluate the effectiveness of pseudo features generated by our proposed RMM generator and AVR diffusion model. Table 4 presents the results when features for the missing modality are generated by each module. The absence of the visual modality leads to a decline in accuracy, particularly for questions related to visual information. However, leveraging pseudo features generated by the RMM generator helps alleviate the impact of the missing modality,

Table 4: Effect of pseudo features generated from each component with missing modalities, by adopting AVST model on MUSIC-AVQA dataset. RMM denotes RMM Generator and AVR refers to AVR diffusion model (V: visual, A: audio, Q: question).

Scenario	Components		Question Type			All
	RMM	AVR	Audio Question	Visual Question	Audio-Visual Question	Avg
Input: $A+Q$ (V: Missing)	✗	✗	65.56	56.45	58.38	59.14
	✓	✗	69.50	59.74	62.22	62.85
	✗	✓	73.00	66.23	62.99	65.62
	✓	✓	74.18	68.04	65.99	67.98
Input: $V+Q$ (A: Missing)	✗	✗	38.76	31.27	38.44	36.60
	✓	✗	68.64	57.93	58.93	60.38
	✗	✓	69.71	70.02	62.97	66.03
	✓	✓	74.30	73.49	66.44	69.71

Table 5: AVQA results on MUSIC-AVQA by changing the number of slots.

# of Slot	Missing Modality	
	V	A
-	59.14	36.60
25	67.41	68.75
50	67.23	68.85
75	67.98	69.71
100	67.17	68.85

Table 6: AVQA results on MUSIC-AVQA by changing time steps of diffusion process.

# of Time Step	Missing Modality	
	V	A
-	62.85	60.38
5	67.51	67.04
10	67.98	69.71
20	67.37	68.20

leading to improved accuracy. Furthermore, employing the diffusion model for pseudo feature generation yields even better results compared to the RMM. While the RMM generator relies on similarity-based feature generation, the diffusion model surpasses it by learning the generating more enhanced representations. We also utilized the features generated from the RMM generator as the initial input for the AVR diffusion model, enabling it to be effectively enhanced. Consequently, by integrating the RMM generator and AVR diffusion model, we generated enhanced features learned with cross-modal knowledge.

Slot Number of RMM Generator Network. To investigate the effect of varying the number of slots in the RMM network, we ran experiments varying the slot size over a range of values: {25, 50, 75, 100}. Table 5 shows the results of the ablation study corresponding to each slot number. As shown in the table, performance peaks when the slot number is set to 75, regardless of the missing modality scenarios. Furthermore, even with variations in the slot number, our method consistently outperforms the baseline approach across all settings.

Effect of the Number of Time Steps in Diffusion Process. In this experiment, we investigate the impact of varying the number of time steps in the diffusion process. Specifically, we examine three different settings: 5, 10, and 20 time steps. As depicted in the Table 6, the AVQA performance is optimal when the number of time steps is set to 10. Consequently, we adopt this value for all subsequent experiments.

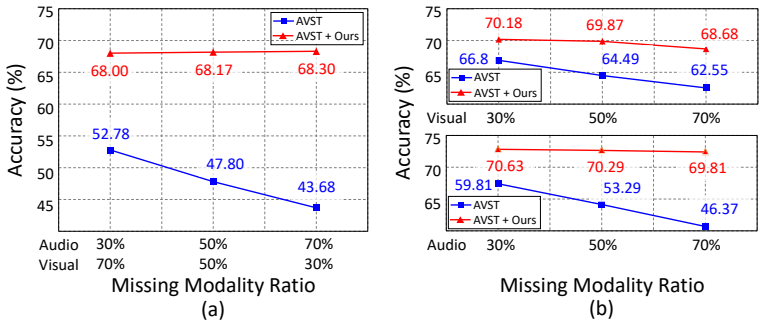

Fig. 5: AVQA results on the MUSIC-AVQA dataset vary based on (a) the missing ratio of visual and audio modalities, and (b) the varying missing ratios for visual (upper) or audio (lower) modalities.

4.5 Discussions

Varying Missing Ratios. We also experimented with various missing rates to test aspects similar to real-world applications. In Fig. 5 (a), we can see that visual and audio perform consistently well even when one of them is missing with a higher ratio. In Fig. 5 (b), we can see that the probability of missing each modality is not significantly affected, and performs consistently well. Fig. 5 results prove that our method can work robustly in terms of real-world applications such as various missing ratio situations.

Limitation. We addresses the problem of missing modalities in only inference situations that are executed after training has taken. However, in terms of further real-world applications, it is possible that missing modalities may occur during learning. So in future work, this consideration will lead to the study of AVQA networks that can robustly cope with missing modalities in training situations.

5 Conclusion

In this work, we introduced a novel Audio-Visual Question Answering (AVQA) framework designed to tackle the challenge of missing modalities in real-world scenarios. Our framework incorporates the Relation-aware Missing Modal (RMM) generator and the Audio-Visual Relation-aware (AVR) diffusion model. The RMM generator generates the pseudo feature of the missing modality, while the AVR diffusion model enhances audio-visual representations. It effectively handles situations where audio or visual information is missing. Through our experiments and comparisons with state-of-the-art AVQA methods, we demonstrated the superior performance of our approach, even in scenarios where one modality is missing. This contributes to enhancing the robustness and accuracy of AVQA networks in real-world environments.

Acknowledgements

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by the IITP grant funded by the Korea government (MSIT) (No. 2022-0-00124: Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities, No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program), and conducted by CARAI grant funded by DAPA and ADD (UD230017TD).

References

1. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Int. Conf. Knowledge Discovery and Data Mining (2018)
2. Calvert, G.A., Hansen, P.C., Iversen, S.D., Brammer, M.J.: Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect. Neuroimage (2001)
3. Chen, Y., Xian, Y., Koepke, A., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
4. Choi, C., Choi, J.H., Li, J., Malla, S.: Shared cross-modal trajectory prediction for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural Inform. Process. Syst. (2021)
6. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst. (2020)
8. Jin, T., Cheng, X., Li, L., Lin, W., Wang, Y., Zhao, Z.: Rethinking missing modality learning from a decoding perspective. In: ACM Int. Conf. Multimedia (2023)
9. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Adv. Neural Inform. Process. Syst. (2022)
10. Kim, D., Park, K., Lee, G.: Oddeyecam: A sensing technique for body-centric peephole interaction using wfov rgb and nfov depth cameras. In: ACM Symp. User Interface Software Technology (2020)
11. Kim, D., Um, S.J., Lee, S., Kim, J.U.: Learning to visually localize sound sources from mixtures without prior source knowledge. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
12. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
13. Kim, J.U., Park, S., Ro, Y.M.: Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory. In: AAAI (2022)
14. Kim, J.U., Ro, Y.M.: Enabling visual object detection with object sounds via visual modality recalling memory. IEEE Trans. Neural Netw. Learn. Syst. (2023)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

16. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE (2015)
17. Lee, S., Kim, H.I., Ro, Y.M.: Weakly paired associative learning for sound and image representations via bimodal associative memory. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
18. Lee, S., Park, S., Ro, Y.M.: Audio-visual mismatch-aware video retrieval via association and adjustment. In: Eur. Conf. Comput. Vis. (2022)
19. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
20. Li, G., Hou, W., Hu, D.: Progressive spatio-temporal perception for audio-visual question answering. In: ACM Int. Conf. Multimedia (2023)
21. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
22. Lindenberger, U.: Human cognitive aging: corriger la fortune? science (2014)
23. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
24. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: AAAI (2021)
25. Maheshwari, H., Liu, Y.C., Kira, Z.: Missing modality robustness in semi-supervised multi-modal semantic segmentation. In: IEEE Winter Conf. Appl. Comput. Vis. (2024)
26. Majumder, S., Chen, C., Al-Halah, Z., Grauman, K.: Few-shot audio-visual learning of environment acoustics. Adv. Neural Inform. Process. Syst. (2022)
27. McGrew, K.S.: Chc theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research (2009)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
29. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
30. Parthasarathy, S., Sundaram, S.: Training strategies to handle missing modalities for audio-visual expression recognition. In: Proc. ACM Int. Conf. Multimodal Interact. (2020)
31. Pian, W., Mo, S., Guo, Y., Tian, Y.: Audio-visual class-incremental learning. In: Int. Conf. Comput. Vis. (2023)
32. Qiu, Z., Yang, H., Fu, J., Fu, D.: Learning spatiotemporal frequency-transformer for compressed video super-resolution. In: Eur. Conf. Comput. Vis. (2022)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
34. Raij, T., Uutela, K., Hari, R.: Audiovisual integration of letters in the human brain. Neuron (2000)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
36. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH (2022)

37. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
38. Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
39. Sweller, J.: Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. Instructional science (2004)
40. Um, S.J., Kim, D., Kim, J.U.: Audio-visual spatial integration and recursive attention for robust sound source localization. In: ACM Int. Conf. Multimedia (2023)
41. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
42. Woo, S., Lee, S., Park, Y., Nugroho, M.A., Kim, C.: Towards good practices for missing modality robust action recognition. In: AAAI (2023)
43. Wu, R., Wang, H., Dayoub, F., Chen, H.T.: Segment beyond view: Handling partially missing modality for audio-visual semantic segmentation. In: AAAI (2024)
44. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: ACM Int. Conf. Multimedia (2022)
45. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: AAAI (2024)
46. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: Int. Conf. Comput. Vis. (2021)
47. Zeng, Y., Yang, H., Chao, H., Wang, J., Fu, J.: Improving visual quality of image synthesis by a token-based generator with transformers. Adv. Neural Inform. Process. Syst. (2021)
48. Zhang, J., Xu, X., Shen, F., Lu, H., Liu, X., Shen, H.T.: Enhancing audio-visual association with self-supervised curriculum learning. In: AAAI (2021)