AV-Unified: A Unified Framework for Audio-visual Scene Understanding

Guangyao Li, Xin Wang, *IEEE Member*, Wenwu Zhu, *IEEE Fellow*

***Abstract*—When humans perceive the world, they naturally integrate multiple audio-visual tasks within dynamic, real-world scenes. However, current works such as event localization, parsing, segmentation and question answering are mostly explored individually, making it challenging to comprehensively understand complex audio-visual scenes and explore inter-task relationships. Hence, we propose AV-Unified, a unified framework that enables joint learning across a wide range of audio-visual scene understanding tasks. AV-Unified standardizes the diverse input-output formats of each task and incorporates a multi-scale spatiotemporal perception network to effectively capture audio-visual associations. Specifically, we unify the inputs and outputs of all supported tasks by converting them into sequences of discrete tokens, establishing a shared representation that allows a single architecture to be trained jointly across heterogeneous varied datasets. Considering the varying temporal granularity of audio-visual events, a multi-scale temporal perception module is designed to capture key cues. Meanwhile, to overcome the lack of auditory supervision in the visual domain, we design a cross-modal guidance-based spatial perception module that models spatial audio-visual associations. Furthermore, task-specific text prompts are employed to enhance the model's adaptability and task-awareness. Extensive experiments on benchmark datasets (*e.g.*, AVE, LLP, MUSIC-AVQA, VGG-SS and AVS) demonstrate the effectiveness of AV-Unified across temporal, spatial, and spatiotemporal tasks.**

***Index Terms*—Audio-visual Scene Understanding, Unified Framework.**

I. Introduction

Cognitive neuroscience suggests that when multiple sensory modalities such as auditory and visual systems work together, the brain forms a highly consistent and enriched perception of the environment [1], [2]. This integration enables cross-modal information transfer and understanding, allowing the brain to transcend individual sensory boundaries. Such multimodal perception and reasoning are fundamental aspects of cognitive intelligence and are essential for human comprehension of the world. Vision and hearing, as the two most important senses for humans to perceive the world, describe objects of interest from different perspectives and are often complementary. By utilizing the audio-visual elements within these multimodal scenes, it is possible to explore more comprehensive scene information [3], thereby overcoming the limitations of perception restricted to a single modality.

Guangyao Li, Xin Wang and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Beijing National Research Center for Information Science and Technology, Beijing 100084, China. E-mail: {guangyaoli, xin_wang, wwzhu}@tsinghua.edu.cn. Corresponding authors: Xin Wang and Wenwu Zhu. The project supported by the National Natural Science Foundation of China (No. 62502268, 62222209), the China Postdoctoral Science Foundation (No.2024M761681), and Beijing National Research Center for Information Science and Technology under Grant No.BNR2023TD03006.

In recent years, significant progress has been made in leveraging audio and visual modalities jointly to enhance the understanding of multimodal scenes, such as event localization (AVE) [4], [5], video parsing (AVVP) [6], [7], sound source localization (SSL) [8], [9], segmentation (AVS) [10]–[12], and question answering (AVQA) [13]–[15], *etc*. Among these, AVE [4] and AVVP [6] focus on localizing the temporal boundaries of visual and auditory events; SSL [8] localizes sound sources by learning the co-occurrence of audio and visual cues; AVS [10] seeks to accurately segment the complete appearance of objects making sound within video frames; and AVQA [14] aim to answer questions related to different visual objects, sounds, and their associations in videos. While these studies have made significant progress, they mostly have been explored based on individual tasks, making it challenging to comprehensively understand complex audio-visual scenes. Considering that humans do not dissect scenes into multiple subtasks when perceiving the world, but instead have a unified understanding of multiple tasks within a scene, it's evident that these tasks are often interrelated and can mutually assist each other. For instance, when a baby and a dog are playing, a caregiver can simultaneously perceive the baby's laughter and the spatial location of the laughter in the field of view, thus better understanding the baby's emotions. Hence, unifying multiple audio-visual tasks in a single framework, allowing a shared-parameter model to solve them, is a valuable topic.

Recently, successful explorations in task unification in the fields of Computer Vision (CV) and Natural Language Processing (NLP) have shown the potential for unifying audio-visual scene understanding tasks. Existing research has attempted to unify tasks in audio-visual learning, such as ONEAVM [16], which addresses tasks like sound source separation and localization but is essentially a multi-task learning paradigm. UniAV [17] has only unified temporal localization tasks, neglecting the importance of spatial localization, highlighting the limitations in the comprehensiveness of existing research. More recently, Crab [18] has achieved initial progress in multi-task joint training, but it relies on fine-tuning with externally constructed data. Thus, developing a unified framework that integrates temporal localization tasks (*e.g.*, AVE [4] and AVVP [6]) with spatial localization tasks (*e.g.*, SSL [8] and AVS [10]) and extends to spatiotemporal reasoning tasks (*e.g.*, AVQA [14]) is crucial for advancing comprehensive audio-visual scene understanding.

To achieve the goal of a unified framework for audio-visual scene understanding tasks, several challenges need to be addressed. **Firstly**, these tasks encompasses various aspects, including temporal event localization, spatial segmentation, and spatiotemporal reasoning. It involves multimodal data such

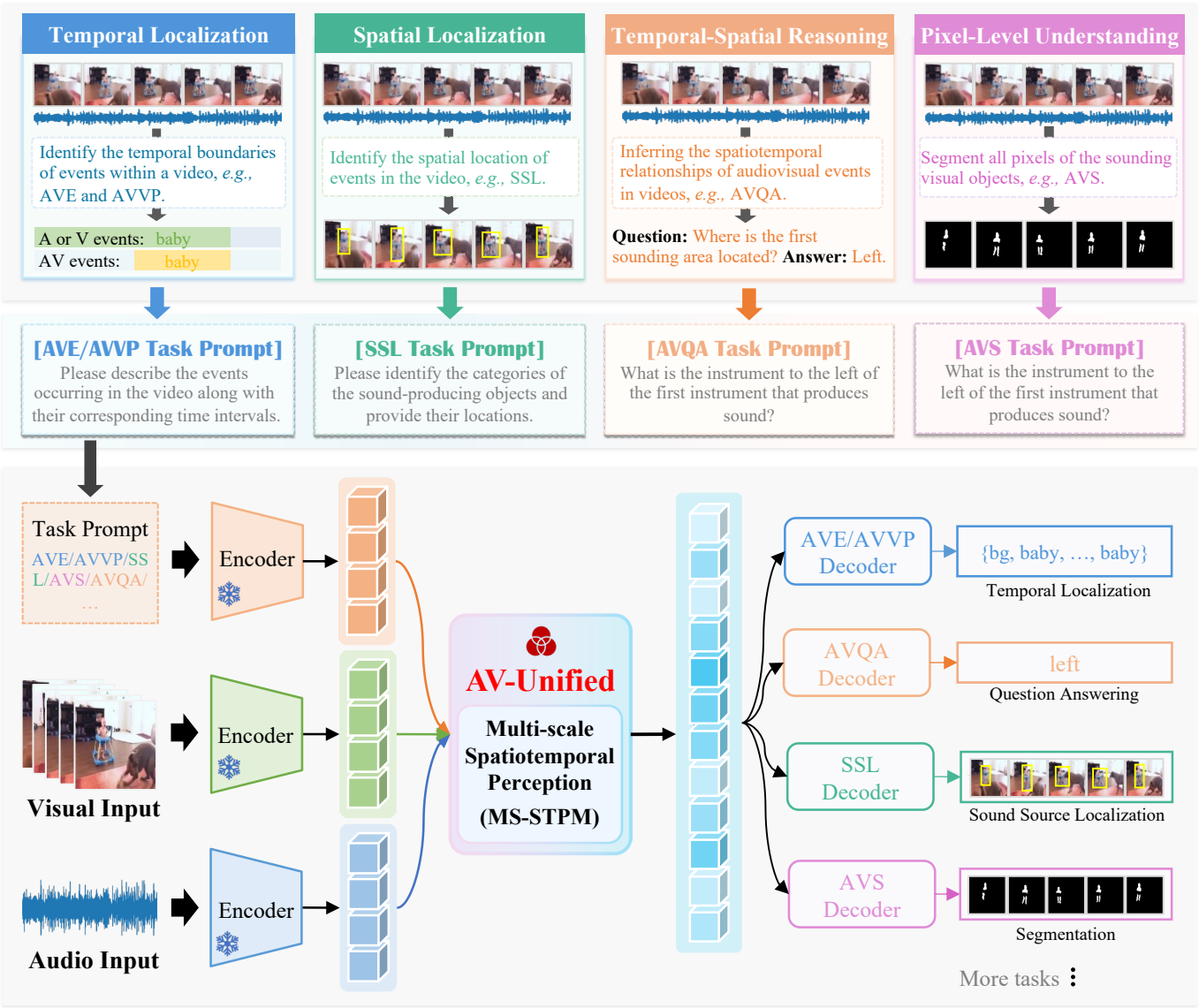

Fig. 1: AV-Unified is a single sequence-to-sequence model that performs a variety of audio-visual tasks using a unified architecture without a need for either task or modality specific branches. A schematic of the model with multiple demonstrative audio-visual tasks: event localization, video parsing, sound source localization, segmentation and question answering.

as images, audio, text, and masks. Thus, a key challenge lies in integrating these diverse modalities into a unified input format while ensuring the model generates consistent outputs adaptable to various downstream tasks. **Secondly**, unifying audio-visual scene understanding tasks is inherently more complex than handling single-modal visual or language tasks, introducing unique difficulties: 1) Temporal perception: Previous works [4], [6], [10], [14] typically samples events uniformly per second, which can disrupt the continuity of events. In real-world scenes, events in natural scenes may span multiple seconds and occur at varying time scales. Hence, it is essential to design models capable of capturing multi-scale audio-visual events to accurately reflect their completeness and continuity. 2) Spatial perception: The lack of supervised information for spatial visual objects and sounds makes it difficult for models to associate sounds with corresponding visual regions in videos. While some existing studies leverage pre-trained object detectors to identify salient regions, these models are typically trained on datasets that lack rare or domain-specific object categories in audio-visual scene (*e.g.*, suona, erhu in AVQA task). This limitation reduces their effectiveness in locating relevant targets, thereby hindering accurate sound-region association and overall spatial perception.

To track these challenges, we propose a Unified Audio-Visual Perception Framework (**AV-Unified**) that enables joint learning across multiple audio-visual scene understanding tasks, as illustrated in Fig 1. **In terms of task paradigm**, AV-Unified unifies the diverse input and output formats of various tasks into a unified sequence-to-sequence paradigm, allowing all tasks to be trained using a single model with shared parameters. To effectively support both temporal and spatial perception within this multi-task framework, we design a **M**ulti-scale **S**patio**t**emporal **P**erception **M**odel (**MS-STPM**) that extracts unified representations adaptable to all

tasks. Additionally, we employ a task-prompt guided learning module that enables the model to automatically attend to task-specific representational cues. Specifically, **For temporal perception**, considering the varying durations of audio-visual events, we designed a multi-scale temporal perception module to integrate audio and visual events at different temporal scales, capturing multimodal information over various time spans. **For spatial perception**, to effectively model the spatial association between audio and visual modalities, a cross-modal guidance-based spatial perception module is developed where audio and visual modalities guide each other, facilitating alignment between visual patch sequences and audio signals, and supporting the learning of fine-grained spatial associations in complex scenes. **For task-specific prompt**, given that different tasks rely on distinct spatiotemporal information, the model needs to selectively attend to relevant features. To this end, we design a prompt-guided feature selection module, which enhances the model's ability to adaptively extract and emphasize the most informative representations for each task. Extensive experiments on benchmark datasets including AVE, LLP, MUSIC-AVQA, VGG-SS and AVS validate the effectiveness of AV-Unified across temporal, spatial, and spatiotemporal tasks.

Our contributions can be summarized as follows:

- Unified the inputs and outputs of classic audio-visual scene understanding tasks (*temporal localization: AVE, AVVP; spatial localization: SSL; pixel-level understanding: AVS; spatiotemporal reasoning: AVQA*) by converting all tasks into a sequence-to-sequence format and training them through a shared parameter network.
- A multi-scale spatiotemporal perception model is proposed to capture events at varying scales and effectively establish audio-visual associations within spatial contexts, addressing the challenges of diverse event durations and the lack of supervisory signals for spatial audio-visual alignment.
- The proposed task-prompt guided learning module is introduced to steer the model toward learning representations that are specifically relevant to each task.
- The proposed AV-Unified achieves well performance across multiple tasks on several benchmark datasets, thoroughly demonstrating its effectiveness and versatility in comprehensive audio-visual scene understanding.

II. Related Works

A. Audio-visual Scene Understanding

Inspired by the multisensory perception of humans, the community has paid more and more attention to audio-visual scene understanding in recent years [3], [19]–[36]. Visual and auditory modalities have distinct features [37]–[42], including cognitive grounding, semantic and spatiotemporal consistency, and strong support from real-world data. It includes various interesting tasks such as event localization (AVE) [4], [5], video parsing (AVVP) [6], [43], [44], sound source localization (SSL) [8], [9], segmentation (AVS) [10]–[12], [45], question answering (AVQA) [14], [46], [47], *etc*. AVE [4] aims to identify auditory and sound events within a video. Addressing the limitations of audiovisual event localization, Tian [6] further introduced the AVVP task, which aims to localize the temporal boundaries of events in a video and categorize them as audible, visible, or both, for a more granular scene understanding. The SSL [8], [9] aims to semantically associate sounds with corresponding visual regions without relying on category annotations. It requires both localizing the sounding object in the visual scene and identifying its category. AVS [10] aims to accurately segment the complete appearance of objects producing sound in video frames, using audio as a guiding signal to determine which object to segment and obtain its complete pixel mask. AVQA [14] aims to answer questions related to different visual objects, sounds, and their associations in the video. These studies integrate rich audiovisual cues within multimodal scenes to overcome limitations in perception inherent to single modalities, thereby utilizing both auditory and visual modalities to explore finer-grained scene comprehension.

Apart from the above methods that facilitate scene understanding by excavating and analyzing different modalities, a unified model should be able to perception their spatio-temporal correlation. Hence, the AV-Unified framework is proposed, which achieves joint learning of the AVE, AVVP, SSL, AVS and AVQA tasks.

B. Audio-Visual Unified Framework

Multimedia scene understanding tasks often involve diverse input-output formats, including images, video frames, and pixel-level masks. In recent years, there has been growing interest in developing unified model architectures to enable more generalized and scalable scene understanding. A widely adopted technical approach is to standardize the representation of task inputs and outputs, typically through tokenization [48]–[51]. This enables a consistent modeling interface across different tasks, as illustrated by UniTab [52], Pix2Seq [53], and Unified-IO [54]. Recently, some studies have begun incorporating the audio modality into unified models. For instance, ONE-AVM [16] unifies localization, separation, and recognition within a single methodological framework, though it fundamentally follows a multi-task learning paradigm. Unified-IO2 [55] includes the audio modality but lacks explicit modeling of cross-modal audiovisual associations. UniAV [17] focuses solely on temporal tasks, while Meerkat [56] considers both temporal and spatial tasks but simplifies audio-visual segmentation by converting pixel-level masks into bounding boxes, thereby lacking fine-grained pixel-level understanding. The recent Crab [18] has made initial progress in jointly training multiple tasks, but it requires fine-tuning with externally constructed data. These efforts highlight the progress made toward unified learning frameworks. However, most existing approaches still underutilize the audio modality and, more critically, fail to model the intrinsic relationships between audio and visual modalities in a unified manner.

In contrast to previous work, the AV-Unified framework proposed in this paper not only unifies the input-output formats of multiple audio-visual tasks but also explicitly and systematically models the deep correlations between audio and visual modalities, enabling a more comprehensive understanding of complex audio-visual scenes.

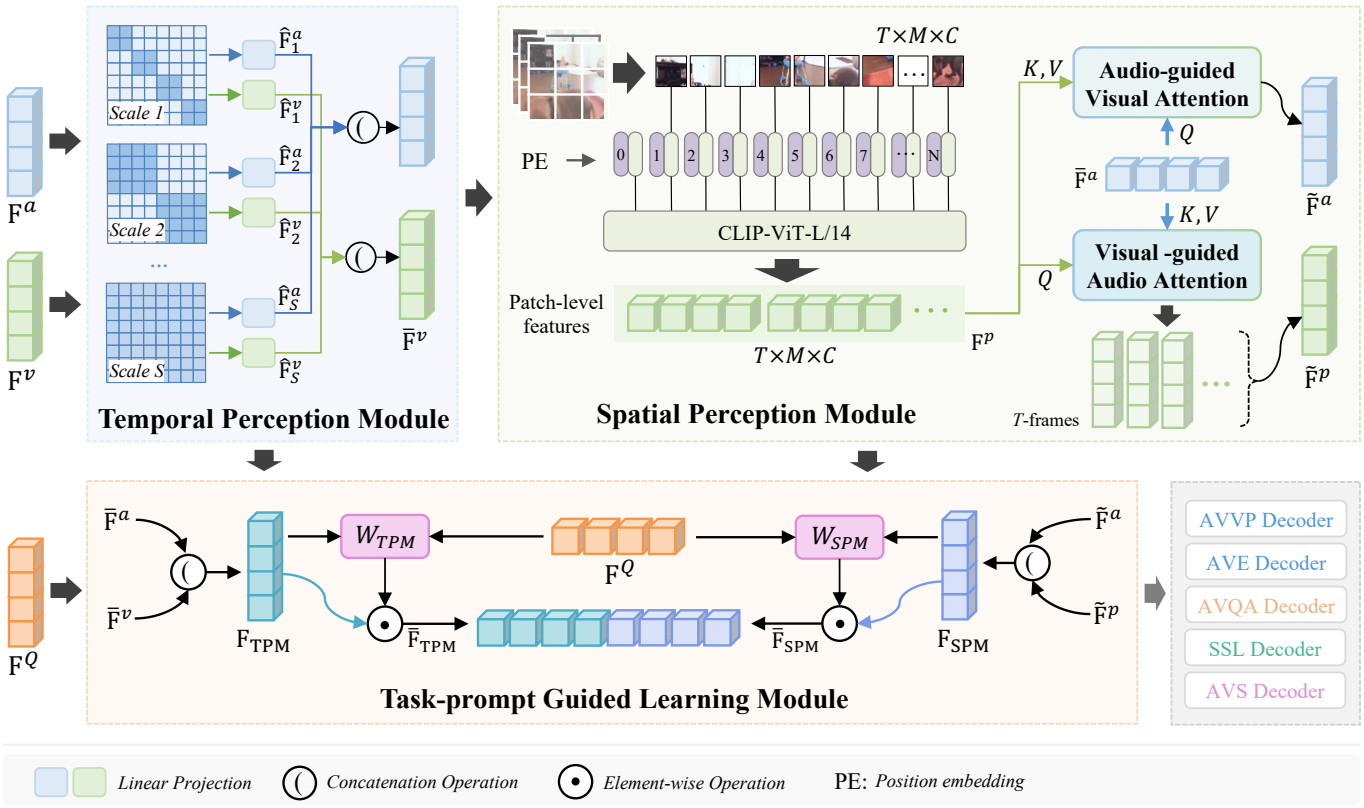

Fig. 2: The proposed Multi-scale Temporal-Spatial Perception Framework. First, the visual and audio features extracted by the encoder are fed into a temporal perception module to capture key audio-visual temporal cues. Then, a spatial perception module performs cross-modal guidance and interaction based on these temporal cues, uncovering spatial associations between the audio and visual modalities. Next, carefully designed task-specific textual prompts guide the model to focus on features that are most relevant to the current task. Finally, the learned representations are serialized and passed to task-specific decoders to address different downstream audiovisual scene understanding tasks.

III. AV-Unified Framework

We proposes a unified framework (AV-Unified) that achieves joint learning for multiple audiovisual scene understanding tasks. AV-Unified standardizes the diverse input and output formats of each task and integrates a **M**ulti-**s**cale **S**patio**t**emporal **P**erception **M**odel (**MS-STPM**) for modeling audio-visual associations, as illustrated in Fig. 2.

A. Unified Task Representations

Given an input video sequence containing both visual and audio tracks, we divide it into T non-overlapping audio and visual segments pairs $\{a_t, v_t\}_{t=1}^{T}$, where each segment spans one second. Subsequently, each video frame is further divided into M patches, and a special [CLS] token is prepended to the first patch. For the given task-specific text prompt Q, we tokenize it into N individual words, denoted as $\{q_n\}_{n=1}^{N}$.

Audio representation. For each audio segment a_t, using the pre-trained VGGish [57] model to extract its feature, denoted as $f_t^a \in \mathbb{R}^D$, where D is the feature dimension. The VGGish model is a VGG-like 2D CNN pre-trained on the large-scale AudioSet [57] dataset, operating on transformed audio spectrograms. The resulting second-level spectrogram features over time can be represented as $F^a = \{f_1^a, f_2^a, \ldots, f_T^a\}$.

Visual representation. A fixed number of frames are sampled from each visual segment v_t. Then we apply pre-trained CLIP [58], with frozen parameters, extract both frame-level and token-level features as f_t^v and f_t^p on video frames, respectively, where $f_t^v \in \mathbb{R}^D$, $f_t^p \in \mathbb{R}^{M \times D}$ and M are token numbers of one frame. Finally, the visual frame-level and token-level features can be denoted as $F_v = \{f_1^v, f_2^v, ..., f_T^v\}$, $F_p = \{f_1^p, f_2^p, ..., f_T^p\}$, respectively.

Text representation. Given the task-specific text prompt Q, we represent each word q_n in a fixed length vector with word embeddings, and then feed it into the pre-trained CLIP [58] model to get the text feature F^Q, where $F^Q \in \mathbb{R}^D$. And the first $[CLS]$ token is used for extracting text features. Note that for the AVQA task, the text prompt is the asked question.

B. Temporal Perception Module

To effectively model audio-visual events with varying durations and temporal scales, we propose a Multi-Scale Temporal Perception Module (**TPM**) designed to capture both fine-grained and coarse-grained temporal dependencies across audio and visual modalities. Specifically, we introduce a multi-scale window attention mechanism with varying window sizes, such as a stacked shifted-window Transformer, where the window size increases progressively with network depth. To

emphasize the importance of local context, our attention strategy employs multi-scale window attention centered around each token. Given a fixed window size S each token attends to $S/2$ neighboring tokens on both sides. In each scale, given a set of audio-visual features $\mathbf{F}^a = \{\mathbf{f}_t^a\}_{t=1}^T$, $\mathbf{F}^v = \{\mathbf{f}_t^v\}_{t=1}^T$ in T segments, HAN [6] applied self-attention and cross-attention layers to aggregate the unimodal and cross-modal information at each timestamp:

$$\hat{\mathbf{F}}_i^a = \phi_{sa}(\mathbf{f}_{i,t}^a, \mathbf{F}_i^a), + \phi_{ca}(\mathbf{f}_{i,t}^a, \mathbf{F}_i^v), i = 2, 4, 6, ..., S, \quad (1)$$

$$\hat{\mathbf{F}}_i^v = \phi_{sa}(\mathbf{f}_{i,t}^v, \mathbf{F}_i^v), + \phi_{ca}(\mathbf{f}_{i,t}^v, \mathbf{F}_i^a), i = 2, 4, 6, ..., S, \quad (2)$$

where S is the size of the sliding window. Then the transformer encoder is employed to aggregate both within-modality and cross-modality information using multi-head attention blocks:

$$\phi_{sa}(\mathbf{f}_{i,t}^a, \mathbf{F}_i^a) = \mathcal{G}(\frac{\mathbf{f}_{i,t}^a \mathbf{F}_i^{a\top}}{\sqrt{d}})\mathbf{F}_i^a, \phi_{ca}(\mathbf{f}_{i,t}^a, \mathbf{F}_i^v) = \mathcal{G}(\frac{\mathbf{f}_{i,t}^a \mathbf{F}_i^{v\top}}{\sqrt{d}})\mathbf{F}_i^v, \quad (3)$$

where $\mathcal{G}(\cdot)$ denotes the *Softmax* function, and $\phi_{sa}(\cdot)$ and $\phi_{ca}(\cdot)$ represent the self-attention and cross-attention operations, respectively. These operations apply dot-product attention over features across temporal steps using non-shared MLPs. Then, we aggregate $\hat{\mathbf{f}}_{s,t}^a$ and $\hat{\mathbf{f}}_{s,t}^v$ across all stages:

$$\overline{\mathbf{F}}^a = \mathbf{\Phi}([\hat{\mathbf{F}}^a, \hat{\mathbf{F}}_1^a, \hat{\mathbf{F}}_2^a, ..., \hat{\mathbf{F}}_S^a]), \quad \overline{\mathbf{F}}^v = \mathbf{\Phi}([\hat{\mathbf{F}}^v, \hat{\mathbf{F}}_1^v, \hat{\mathbf{F}}_2^v, ..., \hat{\mathbf{F}}_S^v]), \quad (4)$$

where $\hat{\mathbf{F}}_S^a = \{\hat{\mathbf{f}}_{s,1}^a, \hat{\mathbf{f}}_{s,2}^a, ..., \hat{\mathbf{f}}_{s,t}^a\}$, $\hat{\mathbf{F}}_S^v = \{\hat{\mathbf{f}}_{s,1}^v, \hat{\mathbf{f}}_{s,2}^v, ..., \hat{\mathbf{f}}_{s,t}^v\}$, and $\mathbf{\Phi}$ is concatenate operation. With the proposed TPM, the model is able to capture critical audio-visual cues along the temporal dimension for improved contextual understanding.

C. Spatial Perception Module

Considering that the position of sound sources and their corresponding visual objects often reflects the spatial correlation between audio and visual modalities, we propose a cross-modal guidance-based Spatial Perception Module (**SPM**). This module leverages strong cross-modal perception capabilities to effectively model spatial associations between audio-visual semantic signals. Specifically, at each time step, there are significant spatial correlations among image patches within a video frame. Given the visual patch-level features $\mathbf{F}^p$ and audio embedding $\mathbf{F}^a$, we first apply a self-attention mechanism to model the intra-frame relationships among visual patches, thereby enhancing the patch-level representation of each video frame. This process can be formally expressed as:

$$\hat{\mathbf{f}}_{t,m}^p = \mathbf{f}_{t,m}^p + \phi_{sa}(\mathbf{f}_{t,m}^p, \mathbf{F}_t^p), \quad (5)$$

where $m = \{1, 2, ..., M\}$. To obtain semantically aligned cross-modal representations, we subsequently perform bi-directional cross-modal attention, where audio features guide the refinement of visual representations and vice versa. Specifically, we apply audio-guided visual attention and visual-guided audio attention to produce enhanced modality-specific features. This process is formally defined as:

$$\widetilde{\mathbf{f}}_{t,m}^p = \hat{\mathbf{f}}_{t,m}^p + \phi_{ca}(\hat{\mathbf{f}}_{t,m}^p, \mathbf{f}_t^a), \quad \widetilde{\mathbf{f}}_t^a = \overline{\mathbf{f}}_t^a + \phi_{ca}(\overline{\mathbf{f}}_t^a, \hat{\mathbf{F}}_t^p), \quad (6)$$

where $\hat{\mathbf{f}}_{t,m}^p \in \mathbb{R}^{M\times C}$, and $\hat{\mathbf{F}}_t^p = \{\hat{\mathbf{f}}_{t,1}^p, \hat{\mathbf{f}}_{t,2}^p, ..., \hat{\mathbf{f}}_{t,M}^p\}$, $\hat{\mathbf{F}}_t^p \in \mathbb{R}^{M\times C}$. Then, we aggregegate $\widetilde{\mathbf{f}}_t^a$, $\widetilde{\mathbf{f}}_t^p = \{\widetilde{\mathbf{f}}_{t,m}^p\}_{m=1}^M$ in all temporal segments as: $\widetilde{\mathbf{F}}^a = \{\widetilde{\mathbf{f}}_1^a, \widetilde{\mathbf{f}}_2^a, ..., \widetilde{\mathbf{f}}_T^a\}$, $\widetilde{\mathbf{F}}^p = \{\widetilde{\mathbf{f}}_1^p, \widetilde{\mathbf{f}}_2^p, ..., \widetilde{\mathbf{f}}_T^p\}$, where $\widetilde{\mathbf{F}}^a \in \mathbb{R}^{T\times C}$ and $\widetilde{\mathbf{F}}^p \in \mathbb{R}^{T\times M\times C}$. $\widetilde{\mathbf{F}}^a$ and $\widetilde{\mathbf{F}}^p$ are the representations obtained from the SMP. With the proposed TPM and SPM, the model is able to establish robust and effective associations between audio-visual cues, thereby enabling it to accurately capture and localize potential sounding regions within video frames.

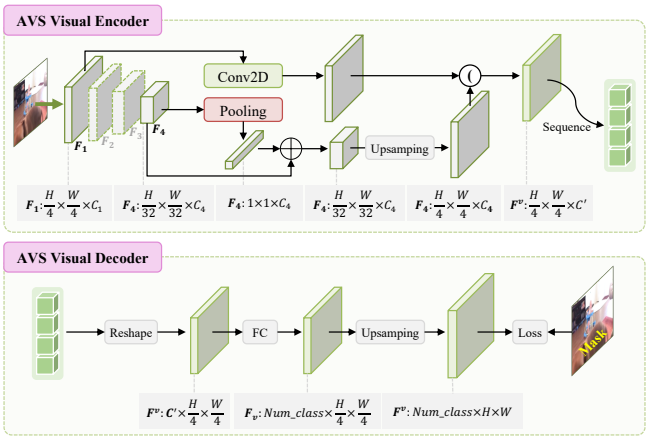

Fig. 3: Encoder and decoder for AVS task.

D. Task-prompt Guided Learning Module

Considering the preferences in training for multiple tasks, such as AVE and AVVP tasks leaning more towards temporal perception, AVS tasks favoring spatial perception, and AVQA task leaning towards spatiotemporal perception. Hence, a Task-prompt Guided Learning Module (**TPGL**) is designed. Its purpose is to manually set a textual prompt for each task to guide the preferences during training, with the prompt for the AVQA task is input question.

For the given audiovisual representations $\overline{\mathbf{F}}^a, \overline{\mathbf{F}}^p$, and $\widetilde{\mathbf{F}}^a, \widetilde{\mathbf{F}}^p$ obtained through TPM and SPM, we employ a linear projection layer to map their dimensions to the same size and apply a $ReLU$ activation function. Then, for the $\widetilde{\mathbf{F}}^a$ obtained by SPM, we transform its dimension from $T \times C$ to $T \times 1 \times C$ and concatenate it with $\widetilde{\mathbf{F}}^p$ at the patch level. Simultaneously, for $\overline{\mathbf{F}}^a, \overline{\mathbf{F}}^p$ obtained through TPM, we also concatenate them along the temporal dimension using a concatenation operation. This process can be represented as:

$$\mathbf{F}_{\texttt{TPM}} = \mathbf{\Phi}([\overline{\mathbf{F}}^a, \overline{\mathbf{F}}^p]), \quad \mathbf{F}_{\texttt{SPM}} = \mathbf{\Phi}([\widetilde{\mathbf{F}}^a, \widetilde{\mathbf{F}}^p]), \quad (7)$$

where $\mathbf{F}_{\texttt{TPM}} \in \mathbb{R}^{2\times T\times C}$, $\mathbf{F}_{\texttt{SPM}} \in \mathbb{R}^{T\times(M+1)\times C}$. Then, for the given task-prompt feature $\mathbf{F}^Q \in \mathbb{R}^{1\times C}$, we use it as the *Query* to calculate similarities separately along the temporal dimension of $\mathbf{F}_{\texttt{TPM}}$ and the patch dimension of $\mathbf{F}_{\texttt{SPM}}$:

$$\mathbf{W}_{\texttt{TPM}} = \mathcal{G}(\frac{\mathbf{F}^Q \mathbf{F}_{\texttt{TPM}}^\top}{\sqrt{d}})\mathbf{F}_{\texttt{TPM}}, \mathbf{W}_{\texttt{SPM}} = \mathcal{G}(\frac{\mathbf{F}^Q \mathbf{F}_{\texttt{SPM}}^\top}{\sqrt{d}})\mathbf{F}_{\texttt{SPM}}, \quad (8)$$

where $\mathbf{W}_{\texttt{TPM}} \in \mathbb{R}^{2T}$, $\mathbf{W}_{\texttt{SPM}} \in \mathbb{R}^{T\times(M+1)}$. The dimensions of $\mathbf{W}_{\texttt{TPM}}$ and $\mathbf{W}_{\texttt{SPM}}$ are transformed into $2T \times 1$ and $T \times (M+1)$ respectively, resulting in updated features $\overline{\mathbf{F}}_{\texttt{TPM}}$, $\overline{\mathbf{F}}_{\texttt{SPM}}$:

$$\overline{\mathbf{F}}_{\texttt{TPM}} = \mathbf{W}_{\texttt{TPM}} \odot \mathbf{F}_{\texttt{TPM}}, \quad \overline{\mathbf{F}}_{\texttt{SPM}} = \mathbf{W}_{\texttt{SPM}} \odot \mathbf{F}_{\texttt{SPM}}, \quad (9)$$

where $\overline{\mathbf{F}}_{\texttt{TPM}} \in \mathbb{R}^{2T\times C}$, $\overline{\mathbf{F}}_{\texttt{SPM}} \in \mathbb{R}^{T\times C\times(M+1)}$. Then, they are concatenated into a single sequence to be utilized for various downstream audiovisual tasks.

TABLE I: The performance of the AV-Unified Framework on the LLP Dataset for the AVVP Task.

Method	Segment-level					Event-level				
	Audio	Visual	Audio-Visual	Type	Event	Audio	Visual	Audio-Visual	Type	Event
AVEL [4]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5	36.5
AVSDN [59]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [6]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MA [60]	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
CVCMS [61]	60.8	63.5	57.0	60.5	59.5	53.8	58.9	49.5	54.0	52.1
DHHN [62]	61.4	63.4	56.8	60.5	59.5	54.6	60.8	51.1	55.5	53.3
MM-Pyramid [63]	61.1	60.3	55.8	59.7	59.1	53.8	56.7	49.4	54.1	51.2
MGN [64]	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
AV-Unified (Ours)	**61.6**	**66.9**	**59.8**	**62.8**	**61.5**	**54.8**	**63.0**	**52.2**	**56.7**	**53.9**

TABLE II: The performance of the AV-Unified on AVE.

Method	Fully-Supervised	Weakly-Supervised
AVEL [4]	72.7	66.7
AVIN [65]	75.2	69.4
AVT [66]	75.8	70.2
CMRAN [67]	77.4	73.0
PSP [68]	77.8	73.5
MM-Pyramid [63]	77.8	73.2
AV-Unified (Ours)	**78.7**	**74.2**

TABLE III: The performance of the AV-Unified on VGG-SS.

Method	CIoU(%)	AUC(%)
Attention10K [8]	18.50	30.20
CoarsetoFine [69]	29.10	34.80
AVObject [70]	29.70	35.70
LVS [71]	34.40	38.20
HardPos [72]	34.60	38.00
EZ-VSL [73]	38.85	39.54
AV-Unified (Ours)	**39.16**	**41.24**

With the integration of the proposed TPM, SPM, and TPGL, the model is able to effectively capture audio-visual associations in dynamic scenes, enabling a comprehensive understanding of audio-visual environments.

E. Training Objective

A strong multimodal model has to be exposed to solving diverse sets of problems during pre-training. To train the AV-Unified multi-task unified framework, we combined several benchmark datasets, including AVE, LLP, VGG-SS, AVS and MUSIC-AVQA. Each training batch is composed of samples from a single task. To mitigate the problem of catastrophic forgetting, we randomly sample a batch from one task in each iteration and update the model parameters using the loss computed for that specific task. Specifically, for a batch of data, assuming that the batch comes from the task *target*, then:

$$LOSS = \sum_{i=1}^{N} w_i \cdot loss_i, \tag{10}$$

where *N* is the total number of tasks, $loss_i$ represents the loss of the i-th task, $w_i = \mathbb{1}_{[i=target]}$. It is worth noting that, since the AVS task involves pixel-level masks, both its encoder and decoder adopt the architecture illustrated in Fig. 3.

IV. Experiment and Analysis

A. Datasets and Evaluation Metrics

Audio-Visual Event (AVE) [4]. It contains 4,143 videos covering 28 event categories, *e.g.*, human and animal activities, and vehicle sounds. Videos in AVE dataset are temporally labeled with audio-visual event boundaries. Each video lasts 10 seconds, and each event lasts at least 2 seconds.

Look, Listen and Parse (LLP) [6]. The LLP consists of 11,849 10-seconds video clips annotated with 25 event categories. It covers various real-life scenes such as speech, music performances, car, cheering, dog, etc. We use the 10,000 video clips with only video-level event annotations for model training. The detailed annotations are available for the remaining 1,849 validation and test videos.

Audio-visual Segmentation (AVS), including **Semisupervised Single-sound Source Segmentation (S4)** [10]. It contains a total of 4, 932 videos, with 3,452 videos for training, 740 for validation, and 740 for testing. The target objects cover 23 different categories, including humans, animals, vehicles, and musical instruments. Besides, Each video contains five frames, but only the first frame is annotated. **Fully-supervised Multiple-sound Source Segmentation (MS3)** [10]. The MS3 contains 424 videos and each video has multiple sounding sources and the sounding objects are visible in the frames. Each video was trimmed to 5 seconds, covering the same categories as the S4 subset. **Fully-supervised Audio-Visual Semantic Segmentation (AVSS)** [74]. The AVSS containing a semantic-labels subset that provides pixel-wise semantic labels, as a significant complement of S4 and MS3. And it includes 12,356 videos covering 70 categories.

VGG Sound Source (VGG-SS) [71] is designed for evaluating sound source localization. The dataset includes over 200 categories and 5,000 videos, featuring annotated labels based on the VGGSound dataset. Each visible sound source in a video clip is explicitly annotated with bounding boxes.

MUSIC-AVQA [14], it contains 9,288 videos covering 22 different musical instruments, with a total duration of over 150 hours and 45,867 QA pairs. The questions are designed under multi-modal scenes containing 33 question templates covering nine types, depending on which modalities are used to discover question-related clues for answer prediction.

B. Implementation Details

For the visual stream, videos are divided into 1-second segments, with frames sampled at $1fps$. We employ the CLIP-ViT-L-14 [58] model pre-trained on ImageNet to extract 512-dimensional feature representations for each visual segment, where the [CLS] token is used as the visual frame-level feature. For the audio stream, signals are sampled at 16 kHz, a standard sampling rate for audio processing. We use the VGGish

TABLE IV: The performance of the AVQA task trained jointly with other multiple audio-visual scene understanding tasks.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
FCNLSTM [75]	70.80	65.66	68.90	64.58	48.08	56.23	82.29	59.92	46.20	62.94	47.45	60.42	60.81
GRU [76]	71.29	63.13	68.28	66.08	68.08	67.09	80.67	61.03	51.74	62.85	57.79	63.03	65.03
Hco_Att [77]	70.80	54.71	64.87	63.49	67.10	65.32	79.48	59.84	48.80	56.31	56.33	60.32	62.45
MCAN [78]	78.07	57.74	70.58	71.76	71.76	71.76	80.77	65.22	54.57	56.77	46.84	61.52	65.83
PSAC [79]	75.02	66.84	72.00	68.00	70.78	69.41	79.76	61.66	55.22	61.13	59.85	63.60	66.62
HME [80]	73.65	63.74	69.89	67.42	70.20	68.83	80.87	63.64	54.89	63.03	60.58	64.78	66.75
HCRN [81]	71.29	50.67	63.69	65.33	64.98	65.15	54.15	53.28	41.74	51.04	46.72	49.82	56.34
AVSD [82]	72.47	62.46	68.78	66.00	74.53	70.31	80.77	64.03	57.93	62.85	61.07	65.44	67.32
PanoAVQA [13]	75.71	65.99	72.13	70.51	75.76	73.16	82.09	65.38	61.30	63.67	62.04	66.97	69.53
ST-AVQA [14]	77.78	67.17	73.87	73.52	75.27	74.40	82.49	69.88	64.24	64.67	65.82	69.53	71.59
COCA [83]	79.35	67.68	75.42	75.10	75.43	75.23	**83.50**	66.63	69.72	64.12	65.57	69.96	72.33
PSTP-Net [84]	73.97	65.59	70.91	77.15	77.36	77.26	76.18	72.23	71.80	71.79	69.00	72.57	73.52
TASS [85]	**83.38**	63.13	**75.92**	**80.37**	**79.51**	**79.93**	82.39	68.91	75.89	64.40	69.22	72.33	74.98
AV-Unified (Ours)	72.23	**78.26**	72.60	75.93	73.38	75.61	79.00	**77.07**	**76.27**	**76.89**	**75.65**	**76.96**	**76.42**

TABLE V: The performance of the S4, MS3 and AVSS task trained individually and jointly with other multiple audio-visual scene understanding tasks in AV-Unified framework.

Method	S4		MS3		AVSS	
	mIoU	F-Score	mIoU	F-Score	mIoU	F-Score
MSSL [69]	44.9	66.3	26.1	36.3	-	-
SST [86]	66.3	80.1	42.6	57.2	-	-
iGAN [87]	61.6	77.8	42.9	54.4	-	
LGVT [88]	74.9	87.3	40.7	59.3	-	-
TPAVI [10]	78.7	87.9	54.0	64.5	29.8	35.2
CATR [12]	81.4	89.6	59.0	67.0	-	-
BAVS [89]	82.0	88.6	58.6	65.5	32.6	36.4
AVSeg [90]	82.1	**89.9**	58.4	**69.3**	36.7	**42.0**
AV-Unified (Ours)	**83.2**	87.1	**59.5**	**69.3**	**37.4**	41.9

network pre-trained on AudioSet to extract 128-dimensional audio features. For each input task-prompt, we adopt the same visual frame-level encoder to extract a 512-dimensional feature vector. All experiments are conducted using the Adam optimizer with an initial learning rate of 1×10^{-4}, which is decayed by a factor of 0.1 every 10 epochs. The batch size and the total number of training epochs are set to 64 and 200, respectively. The proposed AV-Unified framework is trained on 6×NVIDIA A100-40G GPUs.

C. Quantitative Results and Analysis

To validate the effectiveness of the proposed AV-Unified framework, we compare it against recent existing methods across multiple tasks. For temporal localization tasks, we conduct evaluations on the AVE and LLP datasets. For spatial and pixel-level localization tasks, we use the VGG-SS and AVS datasets, respectively. For spatiotemporal reasoning tasks, we perform evaluation on the MUSIC-AVQA dataset.

Comparison with other related models.

As shown in Tab. I, II, III, IV, V, and VI, the proposed AV-Unified framework consistently improves performance across all tasks. Moreover, the varying difficulty levels of these subtasks place different demands on the model's spatio-temporal perception capabilities. In joint training, more challenging subtasks tend to enhance the model's capacity to capture complex spatio-temporal patterns, whereas easier subtasks may dilute this capacity. As a result, the model may perform well on complex tasks but struggle with more simpler ones. For instance, as shown in Tab. IV and Tab. VI, joint training significantly benefits AVQA, particularly in complex reasoning types such as *Counting*, *Localization*, *Comparative*, and *Temporal*, indicating that joint learning helps improve reasoning accuracy on these more complex tasks. However, we also observe slight performance drops under unimodal settings (audio-only or visual-only), suggesting that the current joint optimization strategy leaves room for improvement. Taking the AVS task as an example: AVSS is the most challenging subtask, MS3 represents moderate difficulty, and S4 is the easiest. A comparison of single-task and joint training results in Tab. V and Tab. VI reveals that joint training leads to a drop in F-score for S4, while MS3 shows a 0.9% improvement. This phenomenon is observed not only across different levels of task complexity but also within the same category of audio-visual scene understanding tasks.

Limitation Analysis. Comparing the results of unified training with existing methods, we observe a notable performance gap in certain tasks, particularly the S4, AVSS, and MUSIC-AVQA datasets. This discrepancy can be attributed to two main factors: This discrepancy can be attributed to two primary factors. First, many existing methods are task-specific, designed exclusively for a single task without considering cross-task interactions. Such specialization allows for targeted optimization of both model architecture and learning objectives, thereby favoring performance on individual tasks. In contrast, our multi-task unified framework must balance multiple objectives simultaneously, making such task-specific optimization difficult. While multi-task joint learning facilitates the acquisition of shared audiovisual representations, its main advantage lies in adaptability across diverse tasks rather than maximizing performance for any single one. Second, due to computational resource constraints, we processed the video data for these tasks at a lower sampling rate, which introduced a significant discrepancy in data preprocessing. This difference substantially affects the model's ability to learn temporal representations, leading to a performance drop compared with existing single-task methods.

D. Ablation Studies

Effectiveness of AV-Unified. By comparing single-task and unified training results, we observe that training with the

TABLE VI: The performance of multi-task trained individually (AV-Unified *w/o. jt*, *jt*: joint training) and jointly (AV-Unified) with other multiple audio-visual scene understanding tasks in AV-Unified framework.

Method	AVE [4]		LLP [6]				VGG-SS [71]		S4 [10]		MS3 [10]		AVSS [74]		MUSIC-AVQA [14]	
	Fully	Weakly	Segment-level Type	Segment-level Event	Event-level Type	Event-level Event	CIoU	AUC	mIoU	F-Score	mIoU	F-Score	mIoU	F-Score	AV	All
AV-Unified *w/o. jt*	77.4	73.5	62.2	60.9	56.4	53.6	38.94	40.80	82.40	**89.00**	59.30	68.40	37.10	37.40	71.94	75.32
AV-Unified (Ours)	**78.7**	**74.2**	**62.8**	**61.5**	**56.7**	**53.9**	**39.16**	**41.24**	**83.20**	87.10	**59.50**	**69.30**	**43.10**	**41.90**	**76.96**	**76.42**

TABLE VII: MS-TSPM's module configuration results.

Method	Audio	Visual	Audio-Visual	All
w/o MS-STPM	73.93	79.23	70.37	73.35
w/o TPM	**77.72**	79.81	71.21	74.64
w/o SPM	75.85	**80.64**	71.84	74.88
w/o TPGL	75.54	79.93	71.74	74.59
AV-Unified (Ours)	72.60	75.61	**76.96**	**76.42**

unified AV-Unified framework generally yields better performance across most tasks. As shown in Tab. I and Tab. II, for temporal tasks, unified training improves performance on the AVE task by 1.3% and 0.7% across two evaluation metrics compared to single-task training. Similarly, for the AVVP task, multiple sub-metrics show performance gains. We also observe improvements in spatial perception tasks. In the multi-source sound segmentation (MS3) subtask shown in Tab. V and Tab. VI, joint training increases mIoU by 0.2% and F-score by 0.9%, achieving the best results for this subtask. For more challenging spatio-temporal reasoning tasks, AV-Unified improves the average performance metric by 1.10%. These results demonstrate that incorporating more tasks and training data in a unified framework enhances the model's ability to learn consistent audiovisual representations, thereby benefiting the learning of individual subtasks and validating the effectiveness of the proposed AV-Unified framework.

Effectiveness of MS-TSPM. To evaluate the impact of MS-TSPM on the AV-Unified framework, we conducted a series of ablation experiments. As shown in Tab. VII, on the MUSIC-AVQA dataset, the model performance drops significantly when the MS-TSPM structure is removed (73.35% *vs.* 76.42%). In addition, by comparing the results of AV-Unified without unified training in Tab. VI with those presented in Tab. I, IV, II, III, and V, we observe that even when using MS-TSPM alone, the model still achieves strong performance on specific tasks. This indicates that the model can adapt to different task types and effectively capture the spatiotemporal correlations in audiovisual data, making it well-suited for tasks such as temporal localization, spatial segmentation, and spatiotemporal reasoning. We further analyzed the contribution of each component within MS-TSPM, using the MUSIC-AVQA as a case study. As shown in Tab. VII, when the carefully designed TPM, SPM, and TPGL modules are introduced together, the model achieves the best performance regardless of whether multi-task joint training is applied. Each module also brings a clear performance gain when added individually. In contrast, removing all modules leads to a significant performance drop. These results demonstrate the effectiveness of the proposed components. Working in combination, they help the model better capture the spatiotemporal associations in audiovisual scenes and enhance overall performance.

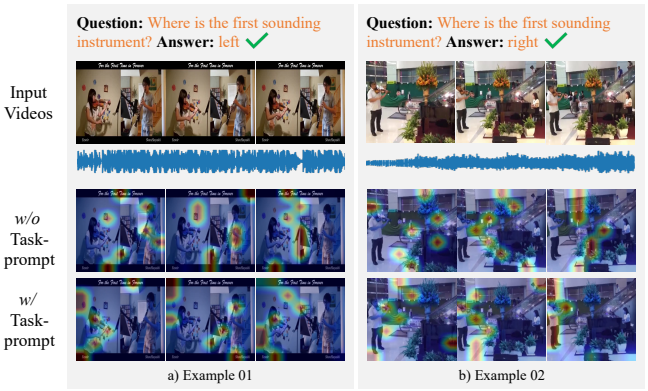

Fig. 4: Visualization of the Temporal-Spatial Task (AVQA). The audio-visual representation of the input video is first processed by the AV-Unified framework to obtain a unified representation, corresponding to the *w/o task-specific prompt* setting. Then, a task-specific prompt is applied to guide the model in selecting the most relevant features required for the task, corresponding to the *w/ task-specific prompt* setting.

E. Visualization Results

To further analyze the spatiotemporal representations learned by the AV-Unified framework, we visualize the results of the AVQA task using heatmaps. Fig. 4 presents two representative examples. These visualizations clearly demonstrate that the task prompt effectively guides the model to focus on task-relevant information within the unified audiovisual representation. Without the task prompt, the model's attention is scattered and often directed toward irrelevant regions, making it difficult to accurately localize sounding instruments. In contrast, when the prompt is provided, the model attends precisely to the instrument areas, which contain the critical audiovisual cues needed to answer the question correctly. These results highlight the importance of explicit task guidance in improving the model's ability to extract meaningful spatiotemporal information from complex audio visual scenes.

F. Discussion with Related Works

Crab and the AV-Unified are proposed around the same period, with Crab capable of jointly handling multiple audio-visual tasks and achieving results of notable reference value. However, the core design philosophies of the two frameworks differ significantly. Crab relies on constructing a large-scale audiovisual dataset to first train a general audiovisual scene understanding model, followed by task-specific fine-tuning. This approach requires substantial additional data and incurs high computational costs. In contrast, AV-Unified is designed based on the intrinsic characteristics of each task. By leveraging the proposed MS-TSPM module, it effectively captures the

spatiotemporal correlations in audiovisual signals, achieving strong performance across multiple tasks without extra data or fine-tuning. Furthermore, while Crab's performance gains primarily depend on the generalization capabilities of the pretrained large model, AV-Unified explicitly models audiovisual spatiotemporal relationships, directly providing essential support for its performance in multi-task scenarios.

V. Conclustion

In this paper, we propose AV-Unified, a unified framework that integrates event localization, video parsing, spatial localization, segmentation, and question answering within the context of audiovisual scene understanding. AV-Unified reformulates all tasks into a unified sequence-to-sequence format and trains them using a shared-parameter network for joint learning. To address challenges such as multi-scale temporal events and the lack of supervision linking spatial visual objects with corresponding sounds, we introduce a Multi-scale Spatiotemporal Perception Model (MS-TSPM), which effectively captures events across different temporal scales and models spatial audiovisual associations. Extensive experiments on multiple benchmark datasets validate the effectiveness and robustness of the proposed framework, demonstrating its strong potential for comprehensive audiovisual scene understanding.

Nevertheless, we observe that AV-Unified's performance on certain subtasks remains suboptimal. This is primarily due to the complexity of achieving effective cross-task collaboration, which often requires extensive experimentation and larger-scale data support. Future research may address this by incorporating larger and more diverse audiovisual datasets, exploring more advanced architectures, and designing better training strategies. While the current framework represents an initial step toward unified audiovisual modeling, future work could extend it to broader task sets. Overall, we believe this study lays a solid foundation and offers a promising direction for unified audiovisual scene understanding.

References

[1] N. P. Holmes and C. Spence, "Multisensory integration: space, time and superadditivity," *Current Biology*, vol. 15, no. 18, pp. R762–R764, 2005.

[2] X. Wang, Y. Zhou, B. Huang, H. Chen, and W. Zhu, "Multi-modal generative ai: Multi-modal llms, diffusions and the unification," *arXiv e-prints*, pp. arXiv–2409, 2024.

[3] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.

[4] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 247–263.

[5] M. Brousmiche, J. Rouat, and S. Dupont, "Multi-level attention fusion network for audio-visual event recognition," *arXiv preprint arXiv:2106.06736*, 2021.

[6] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *European Conference on Computer Vision*. Springer, 2020, pp. 436–454.

[7] K. K. Rachavarapu *et al.*, "Boosting positive segments for weakly-supervised audio-visual video parsing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 192–10 202.

[8] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4358–4366.

[9] D. Hu, Y. Wei, R. Qian, W. Lin, R. Song, and J.-R. Wen, "Class-aware sounding objects localization via audiovisual correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9844–9859, 2021.

[10] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio–visual segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 2022, pp. 386–403.

[11] J. Liu, Y. Wang, C. Ju, Y. Zhang, and W. Xie, "Annotation-free audio-visual segmentation," *arXiv preprint arXiv:2305.11019*, 2023.

[12] K. Li, Z. Yang, L. Chen, Y. Yang, and J. Xiao, "Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1485–1494.

[13] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2031–2041.

[14] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 108–19 118.

[15] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, "Avqa: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3480–3491.

[16] S. Mo and P. Morgado, "A unified audio-visual learning framework for localization, separation, and recognition," in *International Conference on Machine Learning*. PMLR, 2023, pp. 25 006–25 017.

[17] T. Geng, T. Wang, Y. Zhang, J. Duan, W. Guan, and F. Zheng, "Uniav: Unified audio-visual perception for multi-task video localization," *CoRR*, 2024.

[18] H. Du, G. Li, C. Zhou, C. Zhang, A. Zhao, and D. Hu, "Crab: A unified audio-visual scene understanding model with explicit cooperation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 804–18 814.

[19] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard, "Real acoustic fields: An audio-visual room acoustics dataset and benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 886–21 896.

[20] T. Mahmud, S. Mo, Y. Tian, and D. Marculescu, "Ma-avt: Modality alignment for parameter-efficient audio-visual transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7996–8005.

[21] H. Duan, Y. Xia, M. Zhou, L. Tang, J. Zhu, and Z. Zhao, "Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[22] Y. Cheng, Y. Li, J. He, and R. Feng, "Mixtures of experts for audio-visual learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 219–243, 2024.

[23] Y. Wang, P. Sun, D. Zhou, G. Li, H. Zhang, and D. Hu, "Ref-avs: Refer and segment objects in audio-visual scenes," *arXiv preprint arXiv:2407.10957*, 2024.

[24] S. Gong, Y. Zhuge, L. Zhang, Y. Wang, P. Zhang, L. Wang, and H. Lu, "Avs-mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation," *IEEE Transactions on Multimedia*, 2025.

[25] S. Gong, Y. Zhuge, L. Zhang, P. Zhang, and H. Lu, "Complementary and contrastive learning for audio-visual segmentation," *IEEE Transactions on Multimedia*, 2025.

[26] S. Yan, R. Zhang, Z. Guo, W. Chen, W. Zhang, H. Li, Y. Qiao, H. Dong, Z. He, and P. Gao, "Referred by multi-modality: A unified temporal transformer for video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6449–6457.

[27] G. Li, Y. Xu, and D. Hu, "Multi-scale attention for audio question answering," *arXiv preprint arXiv:2305.17993*, 2023.

[28] R. Guo, X. Ying, Y. Chen, D. Niu, G. Li, L. Qu, Y. Qi, J. Zhou, B. Xing, W. Yue *et al.*, "Audio-visual instance segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 550–13 560.

[29] J. Zhou, Z. Li, Y. Yu, Y. Zhou, R. Guo, G. Li, Y. Mao, M. Han, X. Chang, and M. Wang, "Mettle: Meta-token learning for memory-efficient audio-visual adaptation," *arXiv preprint arXiv:2506.23271*, 2025.

[30] G. Li, H. Du, and D. Hu, "Avqa-cot: When cot meets question answering in audio-visual scenarios," 2024.

[31] J. Nie, X. Wang, R. Hou, G. Li, H. Chen, and W. Zhu, "Dynamic spatio-temporal graph reasoning for videoqa with self-supervised event recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 4145–4158, 2024.

[32] H. Chen, X. Wang, H. Chen, Z. Zhang, W. Feng, B. Huang, J. Jia, and W. Zhu, "Verified: A video corpus moment retrieval benchmark for fine-grained video understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40 393–40 406, 2024.

[33] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 271–14 280.

[34] Y. Zhou, X. Wang, H. Chen, X. Duan, and W. Zhu, "Intra-and inter-modal curriculum for multimodal learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3724–3735.

[35] X. Wang, Z. Wu, H. Chen, X. Lan, and W. Zhu, "Mixup-augmented temporally debiased video grounding with content-location disentanglement," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4450–4459.

[36] P. Li, X. Wang, Z. Zhang, Y. Meng, F. Shen, Y. Li, J. Wang, Y. Li, and W. Zhu, "Realtcd: Temporal causal discovery from interventional data with large language model," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 4669–4677.

[37] X. Wang, H. Chen, and W. Zhu, "Disentangled representation learning for multimedia," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9702–9704.

[38] X. Wang, B. Meng, H. Chen, Y. Meng, K. Lv, and W. Zhu, "Tiva-kg: A multimodal knowledge graph with text, image, video and audio," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 2391–2399.

[39] H. Chen, X. Wang, X. Lan, H. Chen, X. Duan, J. Jia, and W. Zhu, "Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3117–3128.

[40] Z. Qian, X. Wang, X. Duan, H. Chen, and W. Zhu, "Dynamic spatio-temporal modular network for video question answering," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4466–4477.

[41] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

[42] N. Ding, C. Deng, M. Tan, Q. Du, Z. Ge, and Q. Wu, "Image captioning with controllable and adaptive length levels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 764–779, 2024.

[43] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=zfo2LqFEVY

[44] W. Hou, G. Li, Y. Tian, and D. Hu, "Towards long form audio-visual video understanding," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[45] Y. Wang, W. Liu, G. Li, J. Ding, D. Hu, and X. Li, "Prompting segmentation with sound is generalizable audio-visual source localizer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5669–5677.

[46] G. Li, H. Du, and D. Hu, "Boosting audio visual question answering via key semantic-aware cues," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5997–6005.

[47] Z. Li, J. Zhou, J. Zhang, S. Tang, K. Li, and D. Guo, "Patch-level sounding object tracking for audio-visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 5075–5083.

[48] Y. Guo, Y. Zheng, M. Tan, Q. Chen, Z. Li, J. Chen, P. Zhao, and J. Huang, "Towards accurate and compact architectures via neural architecture transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6501–6516, 2022.

[49] M. Tan, G. Ni, X. Liu, S. Zhang, X. Wu, Y. Wang, and R. Zeng, "Bidirectional posture-appearance interaction network for driver behavior recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 242–13 254, 2022.

[50] Y. Guo, Y. Zheng, M. Tan, Q. Chen, J. Chen, P. Zhao, and J. Huang, "Nat: Neural architecture transformer for accurate and compact architectures," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[51] X. Zhang, Y. Chen, G. Li, and B. Liang, "Pede: Enhance multi-modal sarcasm detection in videos via prompted emotion distributions," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[52] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, "Unitab: Unifying text and box outputs for grounded vision-language modeling," in *European Conference on Computer Vision*. Springer, 2022, pp. 521–539.

[53] T. Chen, S. Saxena, L. Li, T.-Y. Lin, D. J. Fleet, and G. E. Hinton, "A unified sequence interface for vision tasks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 333–31 346, 2022.

[54] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022.

[55] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, "Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 439–26 455.

[56] S. Chowdhury, S. Nag, S. Dasgupta, J. Chen, M. Elhoseiny, R. Gao, and D. Manocha, "Meerkat: Audio-visual large language model for grounding in space and time," in *European Conference on Computer Vision*. Springer, 2024, pp. 52–70.

[57] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[58] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[59] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2002–2006.

[60] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1326–1335.

[61] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang, "Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 449–11 461, 2021.

[62] X. Jiang, X. Xu, Z. Chen, J. Zhang, J. Song, F. Shen, H. Lu, and H. T. Shen, "Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 719–727.

[63] J. Yu, Y. Cheng, R.-W. Zhao, R. Feng, and Y. Zhang, "Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 6241–6249.

[64] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 722–34 733, 2022.

[65] J. Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4372–4376.

[66] Y.-B. Lin and Y.-C. F. Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[67] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3893–3901.

[68] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8436–8444.

[69] R. Qian, H. D. Di Hu, M. Wu, N. Xu, and W. Lin, "Multiple sound sources localization from coarse to fine," *arXiv preprint arXiv:2007.06355*, 2020.

[70] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Computer Vision–ECCV*

2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, 2020, pp. 208–224.

[71] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing visual sounds the hard way," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 867–16 876.

[72] A. Senocak, H. Ryu, J. Kim, and I. S. Kweon, "Learning sound localization better from semantically similar samples," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4863–4867.

[73] S. Mo and P. Morgado, "Localizing visual sounds the easy way," in *European Conference on Computer Vision*. Springer, 2022, pp. 218–234.

[74] J. Zhou, X. Shen, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang *et al.*, "Audio-visual segmentation with semantics," *International Journal of Computer Vision*, pp. 1–21, 2024.

[75] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2283–2294, 2020.

[76] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[77] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *arXiv preprint arXiv:1606.00061*, 2016.

[78] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[79] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8658–8665.

[80] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1999–2007.

[81] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9972–9981.

[82] I. Schwartz, A. G. Schwing, and T. Hazan, "A simple baseline for audio-visual scene-aware dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 548–12 558.

[83] M. Lao, N. Pu, Y. Liu, K. He, E. M. Bakker, and M. S. Lew, "Coca: Collaborative causal regularization for audio-visual question answering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12 995–13 003, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26527

[84] G. Li, W. Hou, and D. Hu, "Progressive spatio-temporal perception for audio-visual question answering," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 7808–7816. [Online]. Available: https://doi.org/10.1145/3581783.3612293

[85] Y. Jiang and J. Yin, "Clip-powered tass: Target-aware single-stream network for audio-visual question answering," *International Journal of Computer Vision*, vol. 133, no. 5, pp. 2581–2598, 2025.

[86] B. Duke, A. Ahmed, C. Wolf, P. Aarabi, and G. W. Taylor, "Sstvos: Sparse spatiotemporal transformers for video object segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5912–5921.

[87] Y. Mao, J. Zhang, Z. Wan, Y. Dai, A. Li, Y. Lv, X. Tian, D.-P. Fan, and N. Barnes, "Transformer transforms salient object detection and camouflaged object detection," *arXiv preprint arXiv:2104.10127*, vol. 1, no. 2, p. 5, 2021.

[88] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 448–15 463, 2021.

[89] C. Liu, P. Li, H. Zhang, L. Li, Z. Huang, D. Wang, and X. Yu, "Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge," *arXiv preprint arXiv:2308.10175*, 2023.

[90] S. Gao, Z. Chen, G. Chen, W. Wang, and T. Lu, "Avsegformer: Audio-visual segmentation with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 155–12 163.

VI. Biography Section

Guangyao Li is currently a postdoctoral researcher in the Department of Computer Science and Technology at Tsinghua University. He received his Ph.D. degree from the Gaoling School of Artificial Intelligence, Renmin University of China, in 2024. His research interests include multimodal learning and audio-visual scene understanding.

Xin Wang is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications. He has published over 200 high-quality research papers in ICML, NeurIPS, IEEE TPAMI, IEEE TKDE, ACM KDD, WWW, ACM SIGIR, ACM Multimedia etc., winning three best paper awards including ACM Multimedia Asia. He is the recipient of ACM China Rising Star Award, IEEE TCMC Rising Star Award and DAMO Academy Young Fellow.

Wenwu Zhu is currently a Professor in the Department of Computer Science and Technology at Tsinghua University. He also serves as the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs, New Jersey as Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996. His research interests are in the area of data-driven multimedia networking and Crossmedia big data computing. He has published over 400 referred papers and is the inventor or co-inventor of over 100 patents. He received eight Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019.

He served as EiC for IEEE Transactions on Multimedia (2017-2019) and IEEE Transactions on Circuits and Systems for Video Technology (2024-2025). He served in the steering committee for IEEE Transactions on Multimedia (2015-2016) and IEEE Transactions on Mobile Computing (2007-2010), respectively. He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019, respectively. He is an AAAS Fellow, IEEE Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).