Implicit Counterfactual Learning for Audio-Visual Segmentation

Mingfeng Zha[1], Tianyu Li[1]*, Guoqing Wang[1], Peng Wang[1], Yangyang Wu[2], Yang Yang[1], Heng Tao Shen[3]
[1]University of Electronic Science and Technology of China, [2]Zhejiang University, [3]Tongji University

Abstract

Audio-visual segmentation (AVS) aims to segment objects in videos based on audio cues. Existing AVS methods are primarily designed to enhance interaction efficiency but pay limited attention to modality representation discrepancies and imbalances. To overcome this, we propose the implicit counterfactual framework (ICF) to achieve unbiased cross-modal understanding. Due to the lack of semantics, heterogeneous representations may lead to erroneous matches, especially in complex scenes with ambiguous visual content or interference from multiple audio sources. We introduce the multi-granularity implicit text (MIT) involving video-, segment- and frame-level as the bridge to establish the modality-shared space, reducing modality gaps and providing prior guidance. Visual content carries more information and typically dominates, thereby marginalizing audio features in the decision-making. To mitigate knowledge preference, we propose the semantic counterfactual (SC) to learn orthogonal representations in the latent space, generating diverse counterfactual samples, thus avoiding biases introduced by complex functional designs and explicit modifications of text structures or attributes. We further formulate the collaborative distribution-aware contrastive learning (CDCL), incorporating factual-counterfactual and inter-modality contrasts to align representations, promoting cohesion and decoupling. Extensive experiments on three public datasets validate that the proposed method achieves state-of-the-art performance.

1. Introduction

Referring video segmentation [19, 75, 78, 85] which utilizes text or audio prompts to identify matching targets in visual content, has been applied to diverse tasks such as embodied intelligence [15] and autonomous driving [69]. Unlike highly structured text, the low density and fuzziness of audio align more closely with natural attributes, posing significant challenges for audio-visual segmentation (AVS).

In Figure 1, we categorize the AVS issues into four parts based on the complexity of audio and visual inputs. From left to right, the sounding targets transition from static to highly dynamic scenarios, ranging from cases where visual segmentation models alone suffice to those requiring robust temporal and cross-modal joint representations. From bottom to top, as the audio sources evolve from single-source to multi-source and multi-class, the need for decoupling modality representations and achieving accurate alignment becomes increasingly critical. These challenges can be summarized as: *a) Multi-source audio and complex dynamics*, where the simultaneous presence of multiple audio sources and temporal variations complicates effective association, resulting in mismatches, particularly in complex scenes within frames and rapid changes across frames, causing low intra-modal cohesion; *b) Learning preference*, where models favor high information-dense visual features and underutilize sparse audio, leading to weak inter-modal coupling. Current works [5, 13, 16, 36] primarily design problem-specific functional components and combining them. Although effective, learning optimal parameters in a fixed data space risks capturing spurious correlations. For instance, since guitars are typically played by humans to produce sound, models might incorrectly associate humans as sound producers based on statistical regularities. Inspired by the human brain's ability to self-project and simulate unreal scenarios [1, 4, 11], we formulate the ICF, which leverages text to bridge visual and audio modalities, constructs counterfactual text samples, and establishes the contrastive strategy based on representation distributions. This raises three key questions: *1) Why use language as a bridge instead of direct interaction? 2) Why construct implicit counterfactual samples? 3) Why construct contrast learning based on representation distribution?*

We answer the first question. In complex scenes, multiple visual objects may emit sounds synchronously or asynchronously (many-to-many), or a single sound may correspond to multiple visual objects (one-to-many), making direct matching prone to ambiguities. Furthermore, temporal delays between visual content and audio can exacerbate the issue. We aim to leverage the semantic information of text to capture the objects, actions, and their relationships within the scene, facilitating cross-modal alignment and modeling contextual continuity across frames (***challenge a***). Differ-

*Corresponding author.

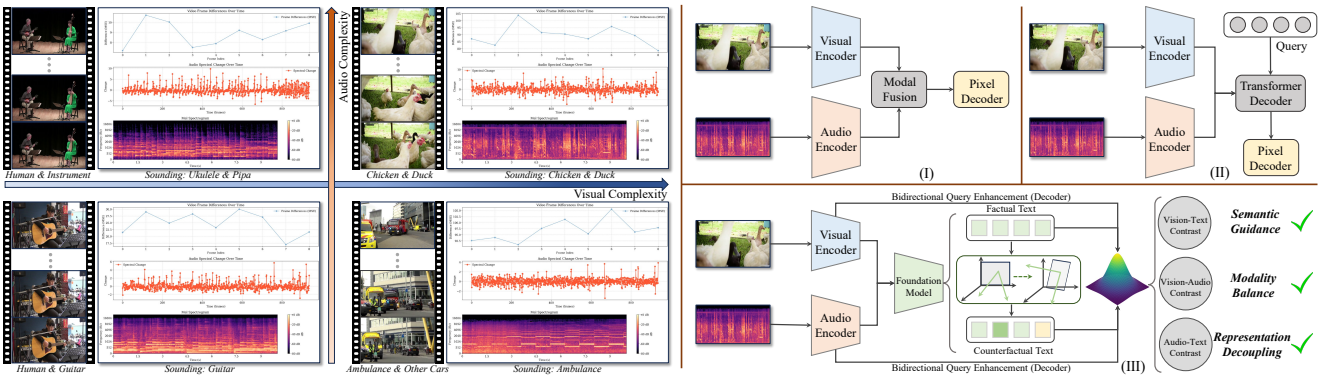

Figure 1. The problem decomposition and method paradigm of AVS task. In the left subfigure, we divide the task into four quadrants based on two metrics: Mean Squared Error (MSE) to measure inter-frame content changes (reflecting visual complexity) and spectral changes (Mel spectrogram) to characterize audio complexity. The dependence of visual complexity on audio exhibits a certain positive correlation. In the right subfigure, existing works can be categorized into (I) Modal fusion; (II) Query learning (with numerous variants prompt-based strategies that are challenging to unify). Our approach (III) involves counterfactual implicit text generation and Gaussian spatial modality contrast, which can be adapted to paradigm (I) and paradigm (II).

ing from [37, 72], which generates explicit texts/categories for visual end, we match the best implicit texts at the feature level for both vision and audio. This approach is based on: a) Implicit texts, derived from higher-order spaces, are less susceptible to low-level noise and more suitable for constructing counterfactual samples; b) Unlike visual semantics, which often encompass silent objects, audio semantics are inherently localized and focused.

We answer the second question. To alleviate the knowledge discrepancies in data structure, some works [31, 77] modify explicit text attributes, such as nouns, colors, sizes, or spatial orientations, or employ generative models (*e.g.,* DALL-E [52]) to create counterfactual images or adjust the audio spectrum (***challenge b***). However, explicit strategies typically require selecting elements for modification from the predefined pool of assumptions, with the upper bound of possibilities potentially smaller than implicit strategies. Our approach also ensures that representations of factual and counterfactual texts progressively diverge, facilitating curriculum-based model training. Additionally, explicit strategies incur significant storage and training costs and are challenging to implement in an end-to-end manner.

We answer the third question. Previous audio-visual works [6, 7] construct contrastive learning at the feature level, which is sensitive to internal structural variations. In contrast, our strategy focuses on: a) Leveraging statistical distributions to model contextual features, enhancing robustness to scene changes, audio noises/mixtures, and inaccuracies in implicit texts, thereby mitigating incorrect matches caused by hard contrastive learning; b) Ensuring stable alignment between visual and audio features with factual texts while maintaining internal compactness, thus preventing degradation from counterfactual texts.

Technically, we propose the multi-granularity implicit text (MIT), which takes video-, segment-, and frame-level visual features as inputs to the foundation models to retrieve best matching implicit textual representations. The similar principles are applied to audio features. We further introduce the semantic counterfactual (SC), which leverages the latent diffusion model to establish continuous and controllable intra-sample and inter-sample orthogonality in the noise space during the forward process, generating counterfactual samples through the denoising process. Finally, we formulate the collaborative distribution-aware contrastive learning (CDCL), which transforms modality features into Gaussian distributions and constructs contrasts in the joint embedding space based on an auxiliary entropy (uncertainty) metric. Our main contributions are as follows:

- To the best of our knowledge, we are the first to model learning biases in the AVS task based on causal inference theory and introduce implicit text to generate counterfactual samples, achieving semantic guidance and balance.
- We propose the MIT and the SC to establish semantic correlations and variances in the continuous space, and further formulate the CDCL to decouple and cohere modality feature distributions.
- Extensive experiments on three public datasets demonstrate that our method achieves state-of-the-art results and can be seamlessly integrated into other AVS approaches, improving performance by 3%-4%.

2. Related Work

Audio-Visual Segmentation. Audio-visual source localization (AVL) [23, 28, 45, 49, 61] determines the approximate location of sound-emitting objects in videos by leveraging audio cues at the regional level, primarily through unsupervised learning to establish cross-modal representation correlations. Recently, Zhou *et al.* [83] advanced the

AVL task by exploring pixel-level audio-visual understanding and introducing the AVS task. Existing fully supervised AVS methods can be broadly categorized into feature decoupling and fusion-based [6, 16, 32, 35, 38, 41, 44, 47, 76], query generation-based [13, 25, 33, 36], and prompt injection-based [7, 37, 40, 60, 71]. Additionally, some studies further explored efficient label learning [2, 14, 48], 3D spatial perception [59], and medical scenario [10]. However, the above works generally overlook biases in learning caused by differences in modality representations and distributions. Unlike [60], which mitigates model preferences via active queries and biased branches, we aim to achieve semantic guidance and representation balance through implicit texts and counterfactual samples.

Contrastive Learning. Contrastive learning [80, 81] aims to minimize the distance between paired samples/categories while maximizing the distance between unmatched samples/different categories. [6, 7, 46, 70] constructed positive and negative sample pairs using semantic categories (prompts) as anchors to facilitate representation disentanglement. In contrast, [60] focused exclusively on audio features. Our strategy differs in two aspects: 1) Measuring modality differences from representation distributions; 2) Establishing heterogeneous tri-modal contrastive learning.

Counterfactual Learning. Based on the ladder theory [50] of causal inference, previous AVS works focused on constructing biased modality associations (*i.e., if..., then...*). Due to the interference of confounding factors, statistically strong correlations between two variables do not necessarily imply causality. Intervention and counterfactual serve as effective methods to achieve impartiality and have been applied in various scenarios, *e.g.,* visual question answering [42, 79], image segmentation [9, 82], and vision-language navigation [65, 66]. Exploring unexperienced situations through intervention poses significant challenges. We instead leverage counterfactual reasoning (*i.e., if not..., then...*) to construct the continuous hypothesis space, thereby expanding the potential boundaries of samples and establishing unbiased and accurate correlations.

3. Methodology

3.1. Overview

In Figure 2, we utilize the visual encoder to generate multi-scale features $\mathbf{F} \in \mathbb{R}^{T\times C_i\times H_i\times W_i}$ (T, C, H, W denote the number of frames, channels, width, and height, respectively). For audio, we perform short-time Fourier transform to obtain Mel-spectrograms, which are then processed by the audio encoder to generate audio features $\mathbf{A} \in \mathbb{R}^{T\times D}$ (D denotes feature dimension). For visual features, we leverage foundation models to match text features corresponding to video-, segment- and frame-level features, *i.e.,* implicit text ℓ^v. Similar operations are conducted on audio features to generate ℓ^a, further forming the composite factual text z. Subsequently, we add noise to z based on the latent diffusion model and orthogonalize partial representations in the noise space, generating counterfactual samples ℓ^{cf} via the denoising process. Finally, we model the probability space of tri-modal features based on Gaussian distributions and conduct contrastive learning across paired modalities.

3.2. Multi-granularity Implicit Text

Inspired by [20], humans associate visual content with auditory cues, typically relying on: 1) Global scene perception (video-level); 2) Inter-frame object motion and semantic change (segment-level); 3) Current moment context (frame-level). For video-level features, we construct long-range pixel-level dependencies between frames. Specifically, we utilize three matrices to map $\mathbf{F}^v$ into query $\mathbf{Q}^v$, key $\mathbf{K}^v$, and value $\mathbf{V}^v$, thereby obtaining the correlation matrix $\mathcal{A}^v$,

$$\mathcal{A}^v = \texttt{Softmax}\left(\frac{\mathbf{Q}^v(\mathbf{K}^v)^\top}{\tau}\right) \in \mathbb{R}^{TC\times TC} \quad (1)$$

where τ and $\top$ represent the temperature coefficient and transpose, respectively. To reduce computational costs, we establish spatiotemporal correlations from the channel dimension rather than the spatial dimension (complexity is reduced from $\mathcal{O}(H^2W^2)$ to $\mathcal{O}(C^2)$). We further update,

$$\mathbf{F}^v := \mathbf{F}^v + \mathbf{V}^v\mathcal{A}^v \quad (2)$$

For segment-level features, we employ similar strategy for sampling. The difference lies in setting the time window to $\lceil\frac{T}{2}\rceil$ or $\lceil\frac{T}{4}\rceil$ and sliding it backward. When the number of frames within the window is insufficient, we utilize preceding frames for supplementation. For frame-level features, the distinction lies in the transition from inter-frame temporal to intra-frame context correlation. Based on textual inversion [12, 62], we utilize VideoCLIP [74] to generate implicit text l^v corresponding to $\mathbf{F}^v$,

$$l^v \in \arg\max_{l_1,l_2,\ldots,l_k} \frac{1}{k^t}\sum_{i=1}^{k^t} \texttt{sim}(\mathbf{F}^v, \texttt{VideoCLIP}(l_i)) \quad (3)$$

where we introduce k^t learnable parameters l to capture multiple visual concepts to avoid focusing exclusively on the most salient semantics. For $\mathbf{F}^s$, likewise. For each frame feature $\mathbf{F}^f$, we obtain T factual texts using CLIP [51]. Thus, we can obtain semantic representations ranging from coarse to fine granularity. To measure the contribution of each element, we introduce a weight query w to dynamically adjust the weights, forming the fused text features ℓ^v,

$$\ell^v = \sum_{p\in\{v,s,f\}}\sum_{n=1}^{N_p}\left(\frac{\exp(w_n^p)}{\sum_{q\in\{v,s,f\}}\sum_{m=1}^{N_q}\exp(w_m^q)}\cdot l_n^p\right) \quad (4)$$

where N_p represents the number of factual texts corresponding to each scale. For audio features, we use CLAP [73] to match the audio factual text ℓ^a. We introduce gate

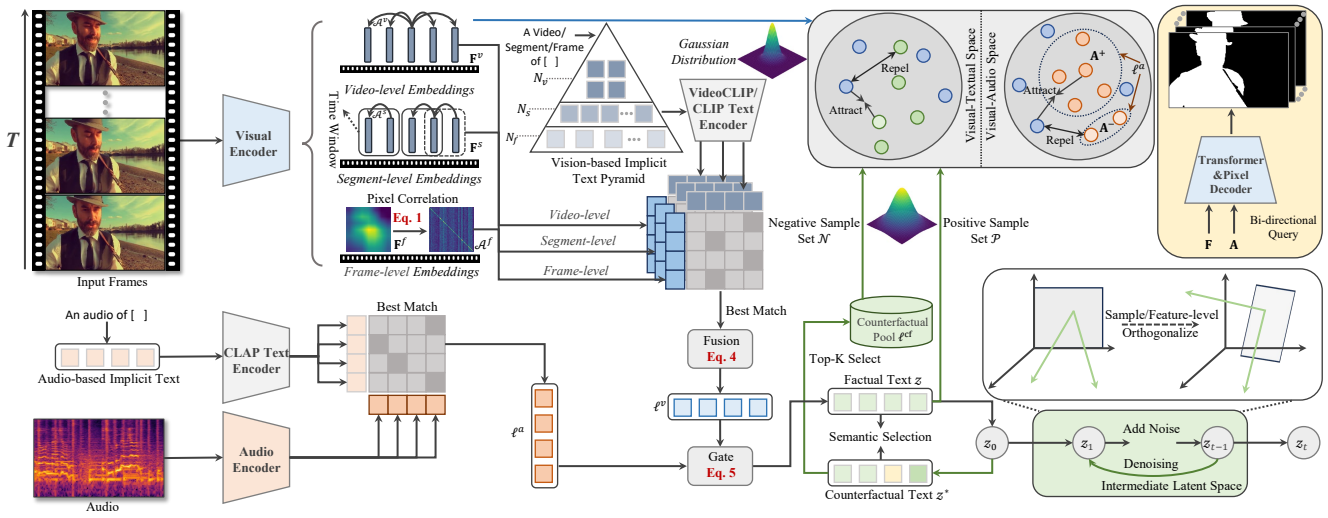

Figure 2. The framework of ICF. For any given audio-video pair, we obtain high-dimensional features using visual and audio encoders separately. For visual features, we establish global temporal correlations, clip semantic partitioning, and intra-frame context alignment at video-segment-frame levels. Subsequently, we initialize a visual-conditioned implicit text pyramid and search for the best elements that match the semantics and adaptively fuse them. For audio features, we perform similar operations using CLAP. Through gate filtering, we generate counterfactual text pools by orthogonalizing factual texts in the diffusion latent space. We transform modality features into distributions and form contrastive learning. Bi-directional query learning [5, 13] serves as the foundation for subsequent decoding.

mechanism (three MLP [39] layers) to alleviate semantics conflicts, and further generate composite factual text z,

$$z = \mathcal{F}(\texttt{Concat}(\texttt{Gate}(\ell^v), \texttt{Gate}(\ell^a))) \tag{5}$$

where $\texttt{Concat}$ denotes dimension-wise concatenation.

3.3. Semantic Counterfactual

At the feature-level, counterfactual samples can be generated directly by adding noise or applying transformations. This strategy may disrupt the original semantic distribution and be uncontrollable (Table 2). We utilize the stable sample construction capability of diffusion model to impose orthogonality in the latent space. To prevent learning direction confusion, we condition on visual features v and gradually add Gaussian noise to z in the forward process,

$$q(z_t|z_{t-1}, v) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I) \tag{6}$$

where β, I denote variance schedule parameter and identity matrix, respectively. We normalize z and then orthogonalize based on the Gram-Schmidt strategy [3],

$$r_{(i)} = \|r_{(i)} - (r_{(i)} \cdot z_{(i)}) z_{(i)}\|, \; r \sim \mathcal{N}(0, I) \tag{7}$$

where r is a random matrix with the same dimension as z. To control the degree of counterfactual, we introduce α to adjust the ratio. However, controlling all samples within the same batch to maintain the same degree does not align with reality. We further introduce $m, s \in (0, 1]$ to ensure the inter-sample and intra-sample feature variations,

$$\begin{aligned} z'_{(i)} &= \sqrt{1-\alpha \cdot m_{(i)}} \cdot z_{(i)} + \sqrt{\alpha \cdot m_{(i)}} \cdot r_{(i)}, \\ z'_{(j)} &= \sqrt{1-\alpha \cdot s_{(j)}} \cdot z_{(j)} + \sqrt{\alpha \cdot s_{(j)}} \cdot r_{(j)} \end{aligned} \tag{8}$$

when $\alpha \cdot m \; (\texttt{or} \; s) = 1$, completely orthogonality is achieved; when $\alpha \cdot m \; (\texttt{or} \; s) = 0$, z remains unchanged. In this way, we can customize the heterogeneity for each sample. During the reverse process, we gradually denoise to generate counterfactual text z^* conditioned on z'.

$$p_\theta(z'_{t-1} \mid z'_t, v) = \mathcal{N}(z'_{t-1}; \mu_\theta(z'_t, t, v), \Sigma_\theta(z'_t, t, v)) \tag{9}$$

where t, μ_θ Σ_θ respectively represent time step, the mean, and covariance prediction functions based on the model parameters θ. For control parameters, we analyze two options: *1) Hyperparameters*, fixed values that require manual design and are difficult to dynamically adjust based on video content; *2) Learnable parameters*, although capable of change through learning, may converge to local optima. Combining the two, we establish the boundary for learning. We aim to constrain the optimization direction through loss,

$$\mathcal{L}_{\text{ortho}} = \|z'_t - z_t\|^2 + \lambda_z \|z'_t \cdot z_t\|^2 \tag{10}$$

where λ_z is the balancing parameter. The first and second terms respectively ensure that the features maintain a certain level of similarity and orthogonality. z' is derived from z through the orthogonal transformation, preserving the internal noise structure. In other words, the noises added and removed come from the same distribution, *i.e.,* noise symmetry. We leverage the objective function in [53] to estimate the noise. The counterfactual text generation loss $\mathcal{L}_{\text{cf}}$ can be formulated as,

$$\mathcal{L}_{\text{cf}} = \mathbb{E}_{t, z_t, \epsilon}\left[\|\epsilon - \epsilon_\theta(z'_t, t)\|^2\right] + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} \tag{11}$$

where ϵ and ϵ_θ represent the actual added noise and the predicted noise, respectively. We select the Top-K samples that

are semantically close in a continuous manner to form the counterfactual text pool,

$$\ell^{\text{cf}} = \arg \underset{z_1^*, z_2^*, \ldots, z_k^*}{\texttt{TopK}} \left(\texttt{sim}(z^*, z)\right) \tag{12}$$

where $\texttt{sim}$ denotes cosine similarity. For samples with significant semantic differences, *i.e.,* simple ones, they occupy much storage space but exhibit the marginal effect for subsequent contrasts.

Discussion. The SC operates based on the MIT, and both are only applied during the training phase. *In the fully latent space (pure noise), orthogonality is ineffective, thus counterfactual generation intervenes in the intermediate state.*

3.4. Collaborative Distribution-aware Contrastive Learning

Feature-level contrastive learning is sensitive to anomalies, *e.g.,* blank video frames, mixed scenes, and drastic changes in audio spectra, which may lead to erroneous sample attraction and repulsion. This issue is further exacerbated when implicit textual semantic descriptions of video or audio content are inaccurate. To overcome this, we transform features into statistical probabilities to enhance tolerance towards anomalies. Specifically, we model the T frames using Gaussian distribution to potentially aggregate sequential information, generating mean (μ^v) and covariance (Σ^v),

$$\mu^v = \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^v,\ \Sigma^v = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{F}_t^v - \mu^v)(\mathbf{F}_t^v - \mu^v)^\top \tag{13}$$

The introduction of low-quality data [6] increases aleatoric uncertainty [22]. Prior works pay little attention to this and treat features from different modalities equally. Although we alleviate through distribution modeling, accurate quantification remains essential. We utilize continuous information entropy [57] $\mathcal{H}$ to reflect the average uncertainty,

$$\mathcal{H}(v) = \frac{1}{2}\log\left((2\pi e)^d \,\texttt{det}(\Sigma^v)\right) \tag{14}$$

where d and $\texttt{det}$ denote the dimension of v and the determinant of the matrix, respectively. Similarly, we incorporate audio features a into Eq. 13 and Eq. 14 to obtain the corresponding statistical metrics. We adopt Wasserstein Distance [55] along with modality entropy to measure the distance between distributions,

$$\begin{aligned}\mathcal{D}(v,a) = \|\mu^v - \mu^a\|_2^2 + \texttt{Tr}\left(\Sigma^v + \Sigma^a - 2(\Sigma^v\Sigma^a)^{1/2}\right) \\ + \gamma(\mathcal{H}(v) + \mathcal{H}(a))\end{aligned} \tag{15}$$

where $\texttt{Tr}$ denotes the trace of the matrix. The latter term measures the looseness (distance) within the distribution, where the low entropy (*i.e.,* low uncertainty) of the joint modality distribution indicates high cohesion. To facilitate visual-audio contrast, the most straightforward approach, *i.e.,* InfoNCE loss [18], is to consider the same $\langle \mathbf{F}^v, \mathbf{A}\rangle$ pairs as the positive sample set $\mathcal{P}$ and the rest as the negative sample set $\mathcal{N}$. However, this hard contrast strategy may cause videos with the same audio semantics to repel each other. We partition the sample space using ℓ^a as the anchor (visual content variations make z unreliable as the reference) to generate $\mathbf{A}^+$ and $\mathbf{A}^-$. We formulate $\mathcal{L}_{\text{v}\leftrightarrow\text{a}}$ as,

$$\begin{aligned}\mathcal{L}_{\text{v}\leftrightarrow\text{a}} = -\frac{1}{B}\sum_{i=1}^{B}\log\Bigg[\sum_{j\in\mathcal{P}(i)}\mathcal{D}'(\mathbf{F}_i^v, \mathbf{A}_j^+) \\ \frac{1}{\sum_{j\in\mathcal{P}(i)}\mathcal{D}'(\mathbf{F}_i^v, \mathbf{A}_j^+) + \sum_{k\in\mathcal{N}(i)}\mathcal{D}'(\mathbf{F}_i^v, \mathbf{A}_k^-)}\Bigg]\end{aligned} \tag{16}$$

where $\mathcal{D}' = \exp(\mathcal{D}(\cdot,\cdot)/\tau)$, and B represents the batch size. For visual-textual contrast, we define $\langle \mathbf{F}^v, z\rangle$ as the $\mathcal{P}$ and $\langle \mathbf{F}^v, \ell_k^{\text{cf}}\rangle$ as the $\mathcal{N}$, without contrasting against other samples within the batch. Orthogonality determines the difference/similarity between counterfactual and factual samples. Thus, k-*th* hard samples (low counterfactual) require assigning higher weights $w_k = \frac{\sqrt{1-\alpha_k}}{\sum_{j=1}^{K}\sqrt{1-\alpha_j}}$ (for simplicity, disregard m and s). We formulate $\mathcal{L}_{\text{v}\leftrightarrow\text{l}}$ as,

$$\mathcal{L}_{\text{v}\leftrightarrow\text{l}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left[\frac{\mathcal{D}'(\mathbf{F}_i^v, z_i)}{\mathcal{D}'(\mathbf{F}_i^v, z_i) + \sum_{k\in\mathcal{N}(i)} w_k\cdot\mathcal{D}'(\mathbf{F}_i^v, \ell_k^{\text{cf}})}\right] \tag{17}$$

Similarly, we can formulate audio-text contrast $\mathcal{L}_{\text{a}\leftrightarrow\text{l}}$.

3.5. Loss Function

Our main loss consists of mask segmentation $\mathcal{L}_{\text{Seg}}$ and contrastive constraints $\mathcal{L}_{\text{contrast}}$. We formulate $\mathcal{L}_{\text{Seg}}$ by,

$$\mathcal{L}_{\text{Seg}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} \tag{18}$$

where sub-items are binary cross-entropy, dice [34], and focal loss [54]. Thus, the total loss can be formulated by,

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Seg}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}} + \lambda_{\text{CDCL}}\sum_{\text{p,q}\in\{\text{a,v,l}\}}\lambda_{\text{p}\leftrightarrow\text{q}}\mathcal{L}_{\text{p}\leftrightarrow\text{q}} \tag{19}$$

4. Experiments

Datasets. We conduct experiments on the AVS-Object [83] and AVS-Semantic (AVSS) [84] benchmarks. AVS-Object contains two subsets: single sound source segmentation (S4) and multiple sound source segmentation (M3), each with five sampled frames. The training/validation/testing sample capacities for S4 and M3 are 3452/740/740 and 296/64/64, respectively. Unlike AVS-Object, which only provides binary masks, AVSS further offers class labels and extends to ten sampled frames, comprising 12,356 (8498/1304/1554) videos across 70 categories.

Implementation Details. We leverage the PyTorch toolbox and execute our algorithm on NVIDIA A100 GPUs. All video frames are resized to 224×224. For the training phase, we employ the AdamW optimizer with the batch size of 8, the learning rate of 1e-4, and the total epochs of 60. During the testing phase, we do not apply any post-processing calibration, *e.g.,* test-time augmentation [56]. For fair comparison, we adopt ResNet-50 [17] and PVT-v2 [68] pretrained on ImageNet as the visual backbone networks, VGGish [21]

Table 1. Quantitative comparison of $\mathcal{J}$, $\mathcal{F}$ and $\mathcal{J}\&\mathcal{F}$ on the S4 [83], M3 [83], and AVSS [84] datasets. External models, *e.g.,* SAM [29]. ↑ represents the larger the better. Best performance in **bold**, second best underlined. ‡ represents data is unavailable

Method	External Model	Backbone	AVS-Object (S4)			AVS-Object (M3)			AVSS		
			$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
AVSBench [83] ECCV22	✗	ResNet-50	78.80	72.79	84.80	52.84	47.88	57.80	‡	‡	‡
ECMVAE [47] ICCV23	✗	ResNet-50	81.42	76.33	86.50	54.70	48.69	60.70	‡	‡	‡
CATR [33] ACM MM23	✗	ResNet-50	80.70	74.80	86.60	59.05	52.80	65.30	‡	‡	‡
AQFormer [25] IJCAI23	✗	ResNet-50	81.70	77.00	86.40	61.30	55.70	66.90	‡	‡	‡
AVSegFormer [13] AAAI24	✗	ResNet-50	80.67	76.54	84.80	56.17	49.53	62.80	27.12	24.93	29.30
AVSC [36] ACM MM23	✗	ResNet-50	81.13	77.02	85.24	55.55	49.58	61.51	‡	‡	‡
BAVS [37] TMM24	✓	ResNet-50	81.63	77.96	85.29	56.30	50.23	62.37	27.16	24.68	29.63
QDFormer [35] CVPR24	✗	ResNet-50	81.80	77.60	86.00	61.55	59.60	63.50	‡	‡	‡
UFE [41] CVPR24	✗	ResNet-50	83.23	78.96	87.50	60.19	55.88	64.50	‡	‡	‡
CAVP [6] CVPR24	✗	ResNet-50	83.84	78.78	88.89	61.48	55.82	67.14	32.83	30.37	35.29
SEIM [32] ACM MM24	✗	ResNet-50	81.40	76.60	86.20	60.05	54.50	65.60	34.55	31.90	37.20
COMBO [76] CVPR24	✓	ResNet-50	85.90	81.70	90.10	60.55	54.50	66.60	35.30	33.30	37.30
CPM [8] ECCV24	✗	ResNet-50	**85.92**	81.37	**90.47**	65.40	59.80	71.00	37.05	34.53	39.57
Ours	✓	ResNet-50	85.90	**82.78**	89.01	**66.93**	**61.77**	**72.08**	**39.00**	**36.22**	**41.78**
AVSBench [83] ECCV22	✗	PVT-v2	83.30	78.70	87.90	59.25	54.00	64.50	32.50	29.80	35.20
DiffusionAVS [46] arXiv23	✗	PVT-v2	85.79	81.38	90.20	64.54	58.18	70.90	‡	‡	‡
AVSBG [16] AAAI24	✗	PVT-v2	86.06	81.71	90.40	60.95	55.10	66.80	‡	‡	‡
CATR [33] ACM MM23	✗	PVT-v2	85.50	81.40	89.60	64.50	59.00	70.00	35.65	32.80	38.50
AVSegFormer [13] AAAI24	✗	PVT-v2	85.98	82.06	89.90	63.83	58.36	69.30	39.33	36.66	42.00
UFE [41] CVPR24	✗	PVT-v2	86.78	83.15	90.40	66.43	61.95	70.90	‡	‡	‡
SEIM [32] ACM MM24	✗	PVT-v2	87.35	83.50	91.20	65.80	60.30	71.30	44.10	41.30	46.90
COMBO [76] CVPR24	✓	PVT-v2	88.30	84.70	91.90	65.20	59.20	71.20	44.10	42.10	46.10
Ours	✓	PVT-v2	**90.07**	**86.63**	**93.51**	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**

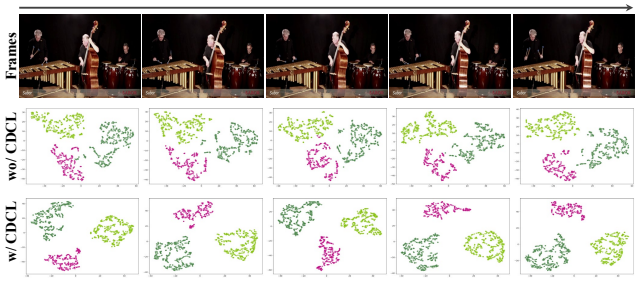

Figure 3. t-SNE visualization. Drum (•), Marimba (•), Cello (•).

pretrained on Youtube-8M as the audio backbone. To generate implicit orthogonal text, we employ Latent Diffusion [53] to reduce computational complexity. Following [76], we leverage Multi-Scale Deformable Attention Transformer as the pixel decoder. The loss balancing hyper-parameter settings and more experiments are included in the supplementary material [1].

Evaluation Metrics. We utilize the Jaccard index $\mathcal{J}$, F-score $\mathcal{F}$, and mean value $\mathcal{J}\&\mathcal{F}$ as metrics for evaluation. Higher values indicate better model performance.

4.1. Main Results

Quantitative Comparison. In Table 1, based on the visual backbone, *i.e.,* ResNet-50 or PVT-v2, we categorize existing AVS methods into CNN-based and Transformer-based (partially overlapping). Our approach outperforms the second-best model, COMBO, by 1.77%, 4.69%, and 4.06% in terms of $\mathcal{J}\&\mathcal{F}$ on three datasets. However, the superiority is not prominent on the S4 dataset. We analyze that the representation of visual scenes from single-source audio is relatively simple and pure, requiring less discriminative ability and representation space range. Conversely, in complex visual scenes with multiple-source audio, the lack of semantics may lead to catastrophic misalignment between audio and visual content. We focus on contrasting and analyzing two methods of incorporating text. BAVS [37] introduces object category information, while our implicit text provides more comprehensive semantic understanding of video (including sequence and background) without prior knowledge. Compared to TeSO [72] [2], which explicitly uses scene descriptions and relies on offline advanced LLMs (*e.g.,* LLaMA2 [63]) for sounding object perception, each step may result in incomplete and suboptimal outcomes. We instead directly match and fuse in the feature space, reducing manual intervention and noise introduction through end-to-end learning. Moreover, our approach outperforms DiffusionAVS [46], which constructs latent diffusion model conditioned on audio and introduces paired contrastive learning, by an average of 4.82%. We may attribute

[1] Project: https://winter-flow.github.io/project/ICL

[2] Using Swin-Base [43] as the visual backbone network.

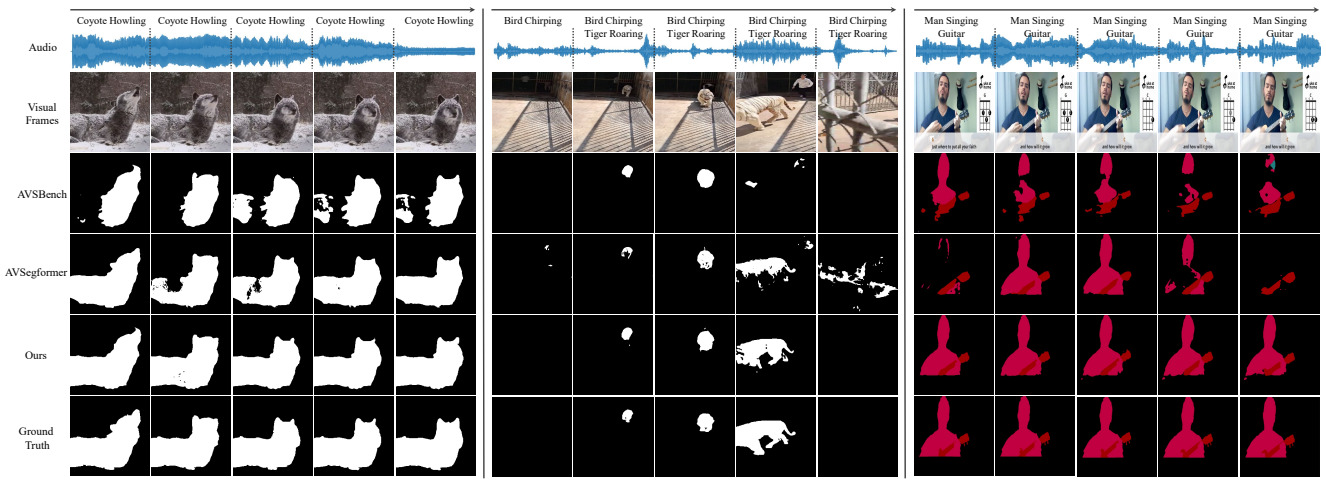

Figure 4. Qualitative comparison. From left to right, the samples are sourced from the S4, M3, and AVSS datasets. **Left:** The white snow on the coyote creates camouflage making it difficult to segment completely. **Middle:** Rapid movement and continuous noise interference (bird chirping); **Right:** Multiple audio and visual categories affected by subtitle background stripe.

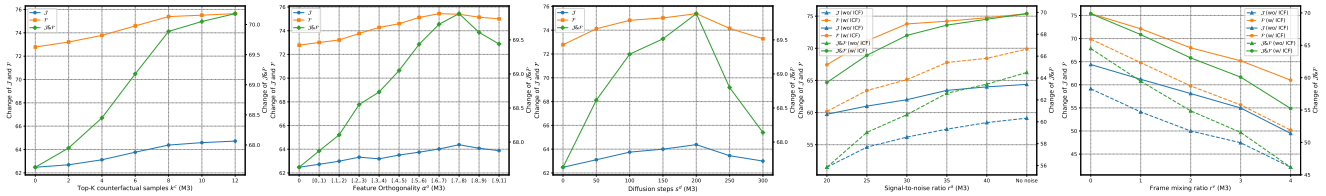

Figure 5. Quantitative ablation of Top-K k^c, orthogonality α^o, number of diffusion steps s^d, signal-to-noise ratio r^a, and frame mixing r^v.

the performance gap to: 1) Inherent uncertainties (instabilities) in audio, *i.e.,* semantic biases caused by multiple factors (*e.g.,* noise disturbances); 2) Insufficient decoupling, lacking accurate reference and sample partition.

Qualitative Comparison. In Figure 4, we present visualization comparison results under various settings. With increasing audio and visual complexity, other methods produce some false negatives or false positives. Our approach remains effective in establishing robust cross-modality correlations and capturing temporal changes, accurately segmenting sounding objects and determining categories.

4.2. Analysis and Discussion

We analyze the effect of each component and framework (Table 2, Table 3 and Figure 6) and the settings of critical hyper-parameters (Figure 5) on the M3 and AVSS datasets.

Crucial Components. The SC cannot work independently of the MIT. We integrate components through one-step and two-step processes. We find that: 1) Directly involving MIT-generated implicit text as an additional cue through cross-attention in modality interaction leads to insignificant performance gains; 2) When text information is not provided, the CDCL degenerates into visual-audio distribution contrast. As preceding components are gradually introduced, the gains increase. We analyze: 1) The implicit text is not purified and aligned, making carried uncertainties and heterogeneous features may disrupt the internal structure of

Table 2. Quantitative ablation of components and strategies.

Component/Strategy			M3			AVSS		
			$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
MIT	SC	CDCL	64.51	59.13	69.88	41.86	38.48	45.23
✓	✗	✗	66.64	61.44	71.83	44.10	40.83	47.37
✗	✗	✓	65.84	60.77	70.90	43.72	39.99	47.45
✓	✓	✗	68.03	62.94	73.11	45.66	42.55	48.76
✓	✗	✓	67.63	62.48	72.78	46.04	42.96	49.11
✓	✓	✓	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**
Frame	Video	Segment	64.51	59.13	69.88	41.86	38.48	45.23
✓	✗	✗	65.07	59.89	70.25	42.34	38.89	45.79
✓	✓	✗	65.68	60.43	70.92	43.11	39.77	46.45
✓	✓	✓	**66.64**	**61.44**	**71.83**	**44.10**	**40.83**	**47.37**
Feature-level *	Inter-sample	Intra-sample	67.63	62.48	72.78	46.04	42.96	49.11
✓	✗	✗	67.21	61.96	72.45	45.84	42.43	49.24
✗	✓	✗	68.87	63.52	74.22	47.33	44.47	50.19
✗	✗	✓	68.82	63.83	73.81	47.07	44.11	50.02
✗	✓	✓	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**
Discrete	Continuous	$\mathcal{L}_{\text{ortho}}$	67.63	62.48	72.78	46.04	42.96	49.11
✓	✗	✗	68.51	63.13	73.88	46.83	43.85	49.81
✗	✓	✗	69.18	63.86	74.49	47.31	44.28	50.33
✗	✓	✓	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**
Pair Swap	Audio Replacement	Text Revision	67.63	62.48	72.78	46.04	42.96	49.11
✓	✗	✗	68.16	62.89	73.42	46.56	43.71	49.40
✗	✓	✗	68.35	63.05	73.64	46.61	43.45	49.77
✗	✗	✓	68.44	63.31	73.56	46.91	43.95	49.86
$\mathcal{L}_{\text{v}\leftrightarrow\text{t}}$	$\mathcal{L}_{\text{a}\leftrightarrow\text{t}}$	$\mathcal{L}_{\text{a}\leftrightarrow\text{v}}$	68.03	62.94	73.11	45.66	42.55	48.76
✓	✗	✗	68.69	63.42	73.96	46.73	43.80	49.66
✓	✓	✗	69.32	63.85	74.78	47.24	44.29	50.19
✓	✓	✓	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**
Prototype	Feature	Distribution	68.03	62.94	73.11	45.66	42.55	48.76
✓	✗	✗	68.32	63.12	73.52	46.22	43.26	49.17
✗	✓	✗	68.59	63.19	73.99	46.90	43.79	50.01
✗	✗	✓	**69.89**	**64.38**	**75.39**	**48.16**	**45.03**	**51.28**

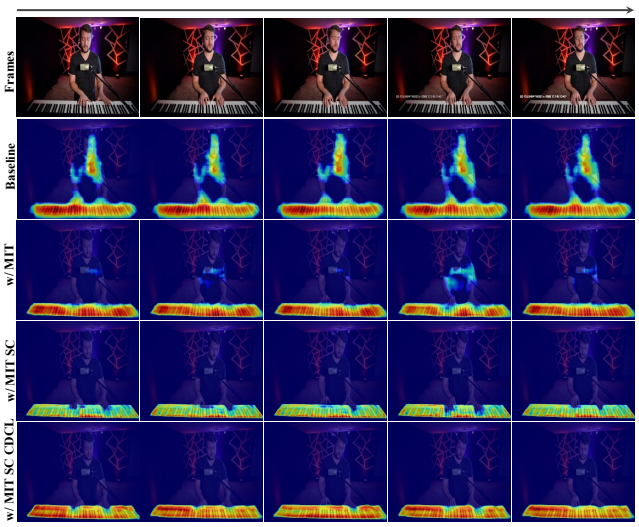

Figure 6. Qualitative ablation of proposed components. From top to bottom, the focus region transitions from salient object (person) to sounding object (piano) and is further enhanced by the CDCL.

Table 3. Effect of integrating the ICF into other approaches.

Method	M3			AVSS		
	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
AVSBench [83] ECCV22	59.25	54.00	64.50	32.50	29.80	35.20
w/ ICF	**62.30**	**57.31**	**67.28**	**37.02**	**34.15**	**39.88**
AVSegFormer [13] AAAI24	63.83	58.36	69.30	39.33	36.66	42.00
w/ ICF	**67.31**	**62.25**	**72.37**	**43.44**	**40.12**	**46.75**

other modalities, thereby reducing the discriminative representation of the shared space. Instead, with the assistance of the CDCL, textual features are implicitly injected, avoiding dedicated modules and suboptimal alignments. 2) With the addition of factual and counterfactual text, the expanded embedding space and accurate references promote modality cohesion and disentanglement.

Multi-granularity Semantics. Segment-level semantic understanding achieve the maximum gain, based on integrating video-level and frame-level. This may be attributed to segments balancing short-term and continuous contexts. Long sequence modeling provides global optimization directions, while intra-frame modeling offers scene spatial correlations, achieving complementarity.

Counterfactual Dimensions, Space and Constraint. The synergistic counterfactual effect within and between samples is greater than single one. We employ VQVAE [64] to replace Diffusion to generate orthogonal representations in the discrete space, yet the performance is not promising. We attribute to the the potential decrease in semantic coherence and sample diversity when forcibly partitioning features into several groups. Conversely, continuous space is more conducive to ensuring the reasonableness and varying degrees. $\mathcal{L}_{\text{ortho}}$ further controls deviations in the denoising.

Counterfactual Strategies. Inspired by [27], we randomly swap half of the video and audio samples with other samples in the datasets to recombine and construct counterfactual pairs, *i.e, Pair Swap*. Similar to [58], we obtain sounding categories and utilize Text-To-Audio model [24] to generate audio with different spectra for replacement, *i.e., Audio Replacement*. We further leverage Qwen2-VL [67] to generate descriptions of videos and replace nouns to generate counterfactual explicit text, *i.e., Text Revision*. The gain of *Text Revision* is the largest, while *Pair Swap* is the smallest. We analyze that: 1) Exchanging samples from the dataset alters the sequential structure but does not fundamentally change the representation space; 2) Clues generated by external models expand the representation space, while text enriches semantic variations. The above strategies require significant computational and storage expenses, and are difficult to dynamically adjust based on input. Our method employ online end-to-end training to adaptively equip counterfactual at different positions.

Modality Contrastive Loss. The tri-modal contrastive loss exhibits complementarity. We rely on masked average pooling or mean to generate prototypes to depict holistic information. Performance averages 1.76% lower than distribution-based approach. We argue that distribution mining captures the general statistical probabilities of representations, avoiding anomalies and unstable interferences. However, treating each element indiscriminately with only the mean dilutes the correct representations. In Figure 3, we illustrate the transition from loose crossover to clustered separation before and after using the CDCL.

Critical Hyper-parameters. To balance performance and cost, we set $k^c = 8$, $\alpha^o = [0.7, 0.8)$, $s^d = 200$ (the larger s^d, the closer the features approach pure noise). As r^a decreases and r^v (1/4 of M3) increases, the performance gap becomes more significant, with about a 10% gain of $\mathcal{J}\&\mathcal{F}$.

Limitation. We achieve matrix heterogeneity by Gram-Schmidt with the complexity of $\mathcal{O}(n^2)$. Besides, numerous crucial hyper-parameters require iterative experiments to determine the optimal combination. In the future, we aim to reduce the complexity (*e.g.,* SVD [30]) and explore automated parameter selection methods (*e.g.,* MoE [26]).

5. Conclusion

Based on causal inference and semantic guidance, we aim to eliminate modality biases and model complex spatiotemporal patterns. Unlike explicitly generating textual descriptions and modifying attributes, we construct counterfactual on the implicitly generated text from the foundation model leveraging *'if not...then...'*. The strategy prevents the model from relying on statistical regularities to form spurious associations. By orthogonalizing the latent space and introducing constraint factors to ensure controllable counterfactual generation, we further establish decoupling and coherence among modalities in the Gaussian space. Our ICF is easily integrated into other frameworks.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091, the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

References

[1] Donna Rose Addis, Alana T Wong, and Daniel L Schacter. Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7):1363–1377, 2007. 1

[2] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiangkang Deng, and Xiatian Zhu. Unsupervised audio-visual segmentation with modality alignment. *arXiv preprint arXiv:2403.14203*, 2024. 3

[3] Åke Björck. Numerics of gram-schmidt orthogonalization. *Linear Algebra and Its Applications*, 197:297–316, 1994. 4

[4] Randy L Buckner and Daniel C Carroll. Self-projection and the brain. *Trends in cognitive sciences*, 11(2):49–57, 2007. 1

[5] Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Le Lu, Jieping Ye, and Nenghai Yu. Bootstrapping audio-visual segmentation by strengthening audio cues. *arXiv preprint arXiv:2402.02327*, 2024. 1, 4

[6] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 2, 3, 5, 6

[7] Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. Cpm: Class-conditional prompting machine for audio-visual segmentation. In *European Conference on Computer Vision*, pages 438–456. Springer, 2024. 2, 3

[8] Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. Cpm: Class-conditional prompting machine for audio-visual segmentation. In *European Conference on Computer Vision*, pages 438–456. Springer, 2025. 6

[9] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11676–11685, 2022. 3

[10] Zhen Chen, Zongming Zhang, Wenwu Guo, Xingjian Luo, Long Bai, Jinlin Wu, Hongliang Ren, and Hongbin Liu. Asiseg: Audio-driven surgical instrument segmentation with surgeon intention understanding. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13773–13779. IEEE, 2024. 3

[11] Giorgio Coricelli, Hugo D Critchley, Mateus Joffily, John P O'Doherty, Angela Sirigu, and Raymond J Dolan. Regret and its avoidance: a neuroimaging study of choice behavior. *Nature neuroscience*, 8(9):1255–1262, 2005. 1

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[13] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 1, 3, 4, 6, 8

[14] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7533–7541, 2024. 3

[15] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721, 2021. 1

[16] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2067–2075, 2024. 1, 3, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[19] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13332–13341, 2024. 1

[20] Hauke R Heekeren, Sean Marrett, Peter A Bandettini, and Leslie G Ungerleider. A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010): 859–862, 2004. 3

[21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 5

[22] Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54 (2-3):217–223, 1996. 5

[23] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 2

[24] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio gen-

eration with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 8

[25] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 3, 6

[26] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 8

[27] Xun Jiang, Zhuoyuan Wei, Shenshen Li, Xing Xu, Jingkuan Song, and Heng Tao Shen. Counterfactually augmented event matching for de-biased temporal sentence grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6472–6481, 2024. 8

[28] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26467–26476, 2024. 2

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6

[30] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980. 8

[31] Chengen Lai, Shengli Song, Sitong Yan, and Guangneng Hu. Improving vision and language concepts understanding with multimodal counterfactual samples. In *European Conference on Computer Vision*, pages 174–191. Springer, 2024. 2

[32] Jiaxu Li, Songsong Yu, Yifan Wang, Lijun Wang, and Huchuan Lu. Selm: Selective mechanism based audio-visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3926–3935, 2024. 3, 6

[33] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1485–1494, 2023. 3, 6

[34] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 5

[35] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3402–3413, 2024. 3, 6

[36] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 1, 3, 6

[37] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 2024. 2, 3, 6

[38] Chen Liu, Peike Patrick Li, Qingtao Yu, Hongwei Sheng, Dadong Wang, Lincheng Li, and Xin Yu. Benchmarking audio visual segmentation for long-untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22712–22722, 2024. 3

[39] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in neural information processing systems*, 34:9204–9215, 2021. 4

[40] Jinxiang Liu, Yu Wang, Chen Ju, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. *arXiv preprint arXiv:2305.11019*, 2023. 3

[41] Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26328–26339, 2024. 3, 6

[42] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 3

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[44] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: a progressive training strategy for audio-visual semantic segmentation. In *European Conference on Computer Vision*, pages 311–327. Springer, 2024. 3

[45] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26742–26751, 2024. 2

[46] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 3, 6

[47] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 3, 6

[48] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36:17208–17221, 2023. 3

[49] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023. 2

[50] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 6

[54] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 5

[55] Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70 (1):117–129, 1985. 5

[56] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021. 5

[57] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 5

[58] Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. Looking similar sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26907–26918, 2024. 8

[59] Artem Sokolov, Swapnil Bhosale, and Xiatian Zhu. 3d audio-visual segmentation. *arXiv preprint arXiv:2411.02236*, 2024. 3

[60] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7259–7268, 2024. 3

[61] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 2

[62] Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. Language-guided audio-visual source separation via trimodal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10584, 2023. 3

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[64] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 8

[65] Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15471–15481, 2022. 3

[66] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13139–13150, 2024. 3

[67] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[68] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 5

[69] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 1

[70] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. *arXiv preprint arXiv:2309.07929*, 2023. 3

[71] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5669–5677, 2024. 3

[72] Yaoting Wang, Peiwen Sun, Yuanchao Li, Honggang Zhang, and Di Hu. Can textual semantics mitigate sounding object segmentation preference? In *European Conference on Computer Vision*, pages 340–356. Springer, 2024. 2, 6

[73] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[74] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[75] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6449–6457, 2024. 1

[76] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does

matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27134–27143, 2024. 3, 6

[77] Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13438–13448, 2024. 2

[78] Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14001–14010, 2024. 1

[79] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19027–19036, 2023. 3

[80] Mingfeng Zha, Yunqiang Pei, Guoqing Wang, Tianyu Li, Yang Yang, Wenbin Qian, and Heng Tao Shen. Weakly-supervised mirror detection via scribble annotations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6953–6961, 2024. 3

[81] Mingfeng Zha, Guoqing Wang, Yunqiang Pei, Tianyu Li, Xiongxin Tang, Chongyi Li, Yang Yang, and Heng Tao Shen. Heterogeneous experts and hierarchical perception for underwater salient object detection. *IEEE Transactions on Image Processing*, 2025. 3

[82] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020. 3

[83] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 2, 5, 6, 8

[84] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 5, 6

[85] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *European Conference on Computer Vision*, pages 452–469. Springer, 2024. 1