Asymmetric Hierarchical Anchoring for Robust Audio–Visual Cross-Modal Generalization

Bixing Wu[1,*] **Yuhong Zhao**[1,*] **Zongli Ye**[1,3,*]
Jiachen Lian[2,†] **Xiangyu Yue**[3] **Gopala Anumanchipalli**[2]
[1]Zhejiang University, China [2]University of California, Berkeley, USA
[3]MMLab, Chinese University of Hong Kong, China
bixingwu@zju.edu.cn, jiachenlian@berkeley.edu
[*]Equal contribution. [†]Corresponding author.

Abstract

Audio–visual joint representation learning under Cross-Modal Generalization (CMG) aims to transfer knowledge from a labeled source modality to an unlabeled target modality through a unified discrete representation space. Existing symmetric frameworks often suffer from information allocation ambiguity, where the absence of structural inductive bias leads to semantic–specific leakage across modalities. We propose Asymmetric Hierarchical Anchoring (AHA), which enforces directional information allocation by designating a structured semantic anchor within a shared hierarchy. In our instantiation, we exploit the hierarchical discrete representations induced by audio Residual Vector Quantization (RVQ) to guide video feature distillation into a shared semantic space. To ensure representational purity, we replace fragile mutual information estimators with a GRL-based adversarial decoupler that explicitly suppresses semantic leakage in modality-specific branches, and introduce Local Sliding Alignment (LSA) to encourage fine-grained temporal alignment across modalities. Extensive experiments on AVE and AVVP benchmarks demonstrate that AHA consistently outperforms symmetric baselines in cross-modal transfer. Additional analyses on talking-face disentanglement experiment further validate that the learned representations exhibit improved semantic consistency and disentanglement, indicating the broader applicability of the proposed framework.

1 Introduction

Multimodal learning aims to mimic human perceptual integration, driving significant progress in tasks such as Visual Question Answering [1, 2, 3] and Audio–Visual Event Localization [4, 5, 6]. In this work, we focus on audio–visual joint representation learning, where early research predominantly employed implicit continuous representations (e.g., CLIP [7], ImageBind [8]) to align semantics via contrastive learning to reduce the distance between different modalities in high-dimensional semantic space, the field has increasingly shifted towards explicit discrete representations based on codebooks (e.g., [9, 10, 11, 12]) or prototypes [13, 14] to represent different modalities. As shown in Figure 1 (2), to mitigate the Modality Gap, a modality-specific branch was introduced, which enables discrete representations to capture cross-modal semantic information. The use of discrete space enables aggregation of similar input features in high-dimensional space, allowing complex feature representations to be achieved with a small number of latent codes. However, existing discretization methods often rely on the idealized assumption of perfect alignment, failing to address the semantic gaps and annotation cost disparities inherent in unconstrained videos. To overcome these limitations, Cross-Modal Generalization (CMG) task and the Unicode framework [15] were introduced. This

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

approach achieves fine-grained unified discrete representations on unlabeled paired data, enabling zero-shot transfer capabilities under single-modality training.

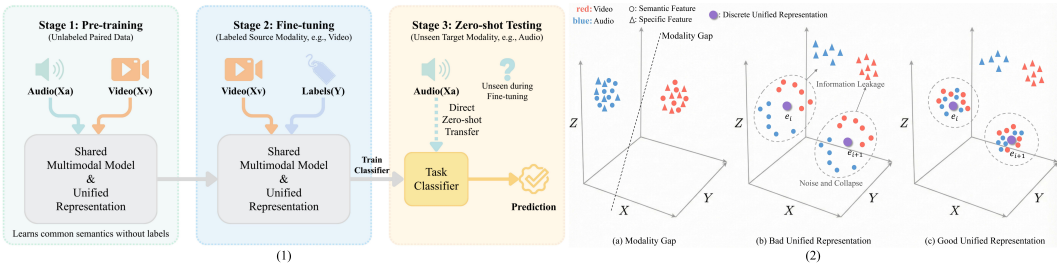

Figure 1: (1) **The Cross-Modal Generalization (CMG) Task Definition**; (2) **Illustration of the challenges and goals in audio-visual joint representation learning**. (a) *Modality Gap*: Audio (blue) and video (red) features are separated in the latent space due to modality-specific biases. This is why we need to separate the semantic and specific parts; (b) *Bad Unified Representation*: A naive unification leads to information leakage and semantic codebook collapse, where semantic and modality-specific factors are entangled around discrete codes; (c) *Good Unified Representation*: A well-structured unified space aligns cross-modal semantic features around shared discrete representations while isolating modality-specific variations.

Although the current explicit representations [9, 10, 11, 12] achieve certain effects on standard downstream tasks, our analysis and experiments show that this does not necessarily mean effective disentanglement of semantic and specific information. In experiments we designed, we find that due to the weak information retention capacity of the previous discrete semantic codebook, the shared space of the symmetric structure undergoes collapse, that is, semantic information tends to flow more towards the specific branches which have less constraints rather than towards the discrete unified representation desired by the CMG task. This exposes potential limitations of the symmetric structure in disentanglement, which we define as Information Allocation Ambiguity. These works [15, 16] often use Mutual Information (MI) estimators, such as the Contrastive Log-ratio Upper Bound (CLUB) [17], to minimize the correlation between the two streams. However, we find that MI minimization via variational upper-bound estimators (e.g., CLUB), which relies on learning a conditional model, becomes unreliable in unconstrained high-dimensional settings.

Our work attempts to answer a more structured question: in audio–visual joint representation, whether semantic information is explicitly concentrated in a designated shared subspace and does not leak into modality-specific branches. This distinction is particularly critical in generative or controllable modeling tasks. To address this, we propose an Asymmetric Hierarchical Anchoring (AHA) structure. Motivated by the fact that audio is commonly used as a conditioning signal in generative settings and often provides a compact and robust cue for high-level semantics under unconstrained visual noise [18, 19, 20, 21, 22], we anchor semantics on the audio modality, utilizing a single-branch Residual Vector Quantization (RVQ) [23] structure to construct an asymmetric semantic anchor point. RVQ has been widely adopted in generation and representation learning(especially in audio field) [24, 25, 26, 27] due to its ability to naturally decompose a signal into a coarse-to-fine hierarchy where the primary quantization layers capture the most salient semantic concepts. We designate this layer as a semantic anchor, forcing the video's semantic encoder to align with this predefined discrete target. This provides a structural prior that guides visual feature distillation, effectively resolving the allocation dilemma. On this basis, we introduce a more effective Gradient Reversal Layer (GRL)-based Adversarial Decoupler. By treating disentanglement as a min-max game, we suppress modality-specific information from the video semantic branch. Finally, we introduce Local Sliding Alignment (LSA) to refine cross-modal alignment at a granular temporal level for unconstrained sequences. This entire structure has been verified as effective through a series of our experiments. Our main contributions are summarized as follows:

- We introduce an Asymmetric architecture that leverages the hierarchical nature of audio RVQ as a structural anchor to provide a stable target for visual semantic distillation.

- We propose a robust GRL-based Adversarial Decoupler that outperforms traditional variational upper-bound MI estimators in separating semantics from modality-specific information.
- We propose AHA, a novel Cross-Modal Generalization framework that achieves effective semantic-specific disentanglement; the quality of the learned representations is further validated by a talking-face diagnostic experiment, suggesting broader applicability beyond classification tasks.

2 Related Works

Multimodal Unified Representations: Recent research on multimodal unified representation has primarily focused on bridging the semantic gap among heterogeneous data. Mainstream approaches typically align modalities by projecting them into a shared latent space [28, 29, 30] or employing modal-general encoders for unified cross-modal feature extraction [31, 32]. To further facilitate cross-modal synergy, cross-modal knowledge distillation is widely adopted for implicit information transfer [30, 33, 34], while representation bridging techniques enhance complementarity by connecting distinct continuous spaces [35, 36]. Furthermore, to enhance interpretability and structure, discretization methods based on codebooks or prototypes have gained increasing attention, mapping continuous features into compact discrete forms [11, 37, 38].

Mutual Information and GRL Disentanglement: In self-supervised learning, mutual information is commonly used for cross-modal alignment, such as in InfoNCE [39], Contrastive Predictive Coding (CPC) [39], and MINE [40], which typically implicitly maximize the mutual information lower bound through contrastive learning or neural estimators. In contrast, for disentanglement between information, variational upper-bound MI estimators like CLUB [17] have difficulty in fitting complex, high-dimensional conditional distributions in unconstrained scenarios. However, in the asymmetric anchor structure we introduced, adversarial learning via Gradient Reversal Layers (GRL) [41] benefits from the anchor, offering a more straightforward information stripping paradigm that provides superior robustness in identifying and eliminating biases without complex density estimation and has been applied in many applications [42, 43, 44, 45]. Consequently, we adapt this adversarial paradigm to purge modality-specific statistics, ensuring a purer semantic space for cross-modal transfer.

3 Method

3.1 Overview and Notations

As illustrated in Figure 2, our framework aligns cross-modal semantics through a dual-stream architecture comprising audio and video branches. Let x_a and x_v denote the input audio and video sequences, respectively. The architecture utilizes three primary encoders: **Audio Encoder (E_a)**: Extracts features from the audio input, denoted as the audio backbone output. **Video Semantic Encoder (E_{v_sem})**: Extracts the continuous semantic feature sequence $Z^V_{sem} \in \mathbb{R}^{T\times D}$, which captures high-level content shared across modalities. **Video Specific Encoder (E_{v_spec})**: Extracts the modality-exclusive feature sequence Z^V_{spec}, capturing video-specific attributes.

The core of our alignment strategy is a hierarchical Residual Vector Quantization (RVQ) module containing n layers which is divided into the Primary Codebook (shared semantic anchor) and Residual Specific Layers. After quantization, we obtain the discrete representations $\mathcal{A}_{unit} = \{A_1, \ldots, A_T\}$ for audio and $\mathcal{V}_{unit} = \{V_1, \ldots, V_T\}$ for video. Finally, the Audio Decoder (G_a) and Video Decoder (G_v) reconstruct the original signals from these disentangled representations.

3.2 Structure of Asymmetric Hierarchical Anchoring

To align cross-modal semantics, we propose Asymmetric Hierarchical Anchoring (AHA), which leverages RVQ's hierarchical structure to impose directional semantic constraints, enabling consistent alignment without symmetric decomposition.

Shared Semantic Alignment (Layers $1 \sim k$): Due to RVQ's tendency to aggregate semantic information in lower layers, the first k layers serve as a cross-modal Semantic Anchor. For the audio

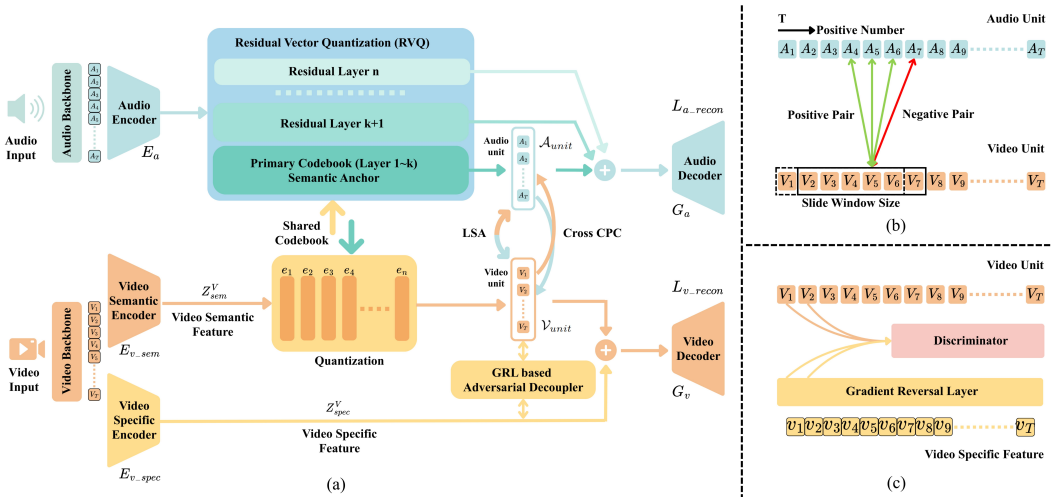

Figure 2: **Overview of the proposed AHA architecture**. (a) *The main pipeline of our entire model*. Through our experiments, we found that the effective semantic information in the audio RVQ is generally concentrated in the first two layers. Therefore, we make slight adjustments to k according to the information density required by the task. For the CMG task, to unify the comparison with previous state-of-the-art work, we set $k = 1$ by default. For the subsequent Talking Face Disentanglement Experiments Figure 4, we set $k = 2$; (b) *Framework of Local Sliding Alignment (LSA)*; (c) *Framework of Gradient Reversal Layer(GRL)-based Adversarial Decoupler*. We also discuss the choice of k and LSA hyperparameters in Appendix D.

branch, the input is processed by these shared layers as

$$\mathbf{q}_l = \text{Quantizer}(\mathbf{r}_{l-1}, \mathcal{C}_l \subset \mathcal{C}_{\text{shared}}); \quad l \leq k. \tag{1}$$

where $\mathcal{C}_l$ is the shared codebook at layer l, and $\mathbf{r}_{l-1}$ is the residual from the previous layer (or the input for $l = 1$).

For the video branch, the continuous semantic feature Z^V_{sem} is also projected onto this shared codebook space. This constraint forces the discretized video representation $\mathcal{V}_{unit}$ to capture only the high-level semantic content that is align with audio semantic and common to both modalities.

Asymmetric Refinement (Layers $k+1 \sim n$): Under the constraint of reconstruction loss, the audio branch extends beyond the shared layers to guide the modality-specific information flow to higher codebooks. The final inputs to the decoders are formulated as:

$$\mathcal{V}_{unit} = \sum_{l=1}^{k} \mathbf{q}_l^V; \mathcal{A}_{unit} = \underbrace{\sum_{l=1}^{k} \mathbf{q}_l^A}_{\text{Semantic Anchor}} + \underbrace{\sum_{l=k+1}^{n} \mathbf{q}_l^A}_{\text{Specific features}} \tag{2}$$

3.3 Adversarial Disentanglement via Contrastive Learning

In this section, we elaborate on the adversarial mechanism used to ensure orthogonality between the semantic and specific representations defined in subsection 3.1. Specifically, we aim to ensure that the specific feature Z^V_{spec} extracted by E_{v_spec} remains independent of the semantic content encapsulated in the Video Units $\mathcal{V}_{unit}$. To achieve this, we employ a Gradient Reversal Layer (GRL)-based Adversarial Decoupler shown in Figure 2(c).

Adversarial Contrastive Objective: Our objective is to decouple the specific features $v_t \in \mathcal{V}_{spec}$ and the semantic units $V_t \in \mathcal{V}_{unit}$. We formulate this as a min-max game involving E_{v_spec} and a conditional discriminator D_ϕ. The discriminator acts as a critic, attempting to distinguish the matched pair (v_t, V_t) from negative samples. Conversely, E_{v_spec} aims to deceive the discriminator via the GRL $\mathcal{R}_\lambda(\cdot)$. Let $s(\mathbf{x}, \mathbf{y}) = D_\phi(\mathcal{R}_\lambda(\mathbf{x}), \mathbf{y})$ denote the similarity score. The unified adversarial

objective is:

$$\min_{D_\phi} \max_{E_{v_spec}} \mathcal{L}_{grl} = -\mathbb{E}\left[\log \frac{\exp(s(v_t, V_t)/\tau)}{\sum_{V' \in \{\mathcal{V}_{unit}\} \cup \mathcal{N}} \exp(s(v_t, V')/\tau)}\right] \tag{3}$$

where τ is the temperature and $\mathcal{N}$ denotes the set of negative semantic units.

Velocity-Aware Anchor Sampling: We define the semantic velocity $\delta_t = \|V_{t+1} - V_t\|_2$. The probability of sampling a time step t as an anchor is proportional to its velocity: $P_{sample}(t) = \frac{\delta_t}{\sum_\tau \delta_\tau + \epsilon}$ This ensures the discriminator focuses on dynamic semantic changes rather than static frames. For implementation details, please refer to Appendix A.

3.4 Local Sliding Alignment

To handle the inherent asynchrony between visual and acoustic signals in unconstrained scenarios, we propose a Local Sliding Alignment (LSA) mechanism. As illustrated in Figure 2(b), instead of enforcing strict global matching, LSA encourages local synchronization between Audio Units $\mathcal{A}_{unit}$ and Video Units $\mathcal{V}_{unit}$. By explicitly modeling local alignment, LSA facilitates more robust cross-modal correspondence under real-world conditions.

We define a Local Search Scope $\Omega_t = \{j \mid |t - j| \leq R\}$ and a Positive Tolerance set $\mathcal{P}_t = \{j \mid |t - j| \leq n_{pos}\}$ (where $n_{pos} \leq R$). We construct a soft-target distribution $Y_{t,j}$ uniformly over $\mathcal{P}_t$. The predicted alignment probability $P_{t,j}$ is computed via a masked softmax over the audio and video units within Ω_t:

$$Y_{t,j} = \frac{1}{|\mathcal{P}_t|}\mathbb{I}(j \in \mathcal{P}_t), \quad P_{t,j} = \frac{\exp(A_t^\top V_j/\tau)}{\sum_{m \in \Omega_t} \exp(A_t^\top V_m/\tau)} \tag{4}$$

The alignment loss minimizes the bidirectional cross-entropy:

$$\mathcal{L}_{align} = -\frac{1}{2T}\sum_{t=1}^{T}\left(\sum_{j \in \Omega_t} Y_{t,j} \log P_{t,j}^{A \to V} + \sum_{j \in \Omega_t} Y_{t,j} \log P_{t,j}^{V \to A}\right). \tag{5}$$

3.5 Fine-Grained Alignment via Cross-CPC and MM-EMA

To bridge the modality gap at a fine-grained level, we employ Cross-Modal Contrastive Predictive Coding (Cross-CPC). We use modality-specific LSTMs to aggregate historical contexts from the quantized units, denoted as $\mathbf{h}_t^V$ and $\mathbf{h}_t^A$. Crucially, the context of one modality predicts the future semantic units of the counterpart (e.g., $\mathbf{h}_t^V$ predicts A_{t+step}). The symmetric InfoNCE objective is:

$$\mathcal{L}_{CPC} = \frac{1}{2}\left(\mathcal{L}_{V \to A} + \mathcal{L}_{A \to V}\right), \quad \mathcal{L}_{V \to A} \propto -\sum_{t,step} \log \frac{\exp(A_{t+step}^\top W_{step}^V \mathbf{h}_t^V)}{\sum_{A_j \in \mathcal{N}} \exp(A_j^\top W_{step}^V \mathbf{h}_t^V)}. \tag{6}$$

To encourage feature coupling in the shared latent space, we adopt Multi-Modal Exponential Moving Average (MM-EMA) for codebook updates. Unlike standard EMA, which updates each code vector based on a single modality, MM-EMA aggregates statistics from both audio and video features to ensure that each codebook vector $\mathbf{e}_i$ represents the centroid of the joint distribution.

Specifically, for code i, we maintain a moving average of cluster size N_i and embedding sum $\mathbf{m}_i$, updated as:

$$\begin{aligned} N_i^{(t)} &= \gamma N_i^{(t-1)} + (1-\gamma)(n_i^a + n_i^b), \\ \mathbf{m}_i^{(t)} &= \gamma \mathbf{m}_i^{(t-1)} + (1-\gamma)\left[\sum_{j=1}^{n_i^a} A_{i,j} + \sum_{j=1}^{n_i^b} V_{i,j}\right], \quad \mathbf{e}_i^{(t)} = \mathbf{m}_i^{(t)}/N_i^{(t)}, \end{aligned} \tag{7}$$

where n_i^a, n_i^b denote the number of features assigned to code i from each modality, and γ is the decay factor.

To further align encoder outputs with these shared centroids, we modify the commitment loss to include both modalities:

$$\mathcal{L}^{a}_{commit} = \beta \|A_i - \text{sg}[\mathbf{e}^{a}_i]\|_2^2 + \frac{\beta}{2}\|A_i - \text{sg}[\mathbf{e}^{b}_i]\|_2^2. \tag{8}$$

The commitment loss for video $\mathcal{L}^{v}_{commit}$ is defined in the same manner. This approach stabilizes quantization while pulling features from each modality closer to the shared semantic anchors.

3.6 Total Loss Function

The full objective function is a weighted combination of reconstruction, alignment, and disentanglement losses:

$$\mathcal{L}_{total} = L_{a_recon} + L_{v_recon} + \mathcal{L}_{VQ} + \mathcal{L}_{grl} + \mathcal{L}_{CPC} + \mathcal{L}_{align}, \tag{9}$$

where L_{a_recon} and L_{v_recon} correspond to the reconstruction losses shown in Figure 2(a), $\mathcal{L}_{VQ}$ sum of the commitment loss from MM-EMA and the commitment loss introduced by the additional audio RVQ layer. Each loss term targets a distinct component with naturally compatible gradient scales. We conducted sensitivity experiments and found performance remains robust across a range of weight combinations, with no single term dominating.

4 Experiment

4.1 Cross-Modality Downstream Tasks

4.1.1 Experimental Settings

Pre-training: We follow Unicode's CMG evaluation protocol, select the same audio backbone and video backbone, pre-train on unlabeled paired audio–video data, freeze the encoders, train the same linear head using labeled source-modality features and evaluate zero-shot on the target-modality (V→A / A→V). Unless otherwise specified, we use the same single-layer linear classifier and settings as Unicode. Pre-training is conducted on VGGSound-AVEL 40K [46, 47], a VGGSound-derived in-the-wild dataset with naturally noisy audio–visual correspondence. Recent re-annotation of VGGSound reports incomplete labels, overlapping classes, and modality misalignment, with 48.43% of original test samples containing incorrect labels or misaligned modalities and 40% containing background music, voice-over, or static images [48]. Thus, our CMG setting is not clean-pair pre-training but realistic noisy audio–visual learning. (Please see the Appendix B for more implementation details). We also discuss training efficiency in Appendix E.

Downstream Tasks: We evaluate four standard CMG tasks, each in both transfer directions: **(i) AVE** [4]: cross-modal event classification on AVE; **(ii) AVVP** [49]: cross-modal event localization on AVVP; **(iii) AVE → AVVP**: train on AVE classification and evaluate fine-grained localization on AVVP; **(iv) UCF(v) ↔ VGG(a)**: cross-dataset cross-modal classification between UCF-101 [50] (vision) and VGGSound (audio). (Details are shown in Appendix B) We report accuracy for classification tasks and F1 for localization tasks.

Baselines: We compare with MST [51], CODIS [9], TURN [52], CMCM [10], and Unicode [15]. We also report DCID [16] and FCID [16] as reference results. They use additional text supervision (AVT) during unified-representation pre-training.

4.1.2 Performance Analysis

Table 1 shows that AHA outperforms Unicode on all eight CMG transfer settings, with the largest gains on AVVP, a localization benchmark that is particularly sensitive to fine-grained temporal cross-modal alignment.

Improvements persist under dataset shift (AVE → AVVP and UCF(v) ↔ VGG(a)), suggesting that AHA generalizes beyond in-distribution evaluation and can transfer discriminative structure learned on AVE to fine-grained localization on AVVP.

Compared with AVT baselines (DCID/FCID) that use extra text supervision as a stronger semantic bridge across datasets, AHA remains competitive using only audio–video pairs and achieves the best results on AVE/AVVP/UCF(v) ↔ VGG(a), while AVT methods are stronger on AVE → AVVP.

Table 1: Cross-modal generalization on downstream CMG tasks. In the tables, V → A denotes training with video-modality labels and testing on audio, whereas A → V denotes the reverse. **Bold** indicates the best performance, and blue denotes the improvement over Unicode.

Method	AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)		Avg.
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
MST(AV)	19.5	23.1	22.7	24.5	29.5	36.4	45.7	43.1	30.56
CODIS(AV)	36.8	39.7	32.7	32.6	40.8	40.6	50.8	45.2	39.90
TURN(AV)	37.6	39.2	32.4	32.2	40.6	41.4	50.4	46.1	39.99
CMCM(AV)	46.3	45.8	36.1	35.2	47.1	48.2	51.2	48.3	44.78
Unicode(AV)	49.7	52.3	59.7	63.1	48.9	50.2	64.4	60.6	56.11
DCID(AVT)	54.5	55.0	40.9	41.6	56.5	53.6	68.1	61.7	53.99
FCID(AVT)	55.9	55.0	43.6	45.1	**57.4**	**58.5**	69.6	62.0	55.89
Ours(AV)	**57.1**	**59.4**	**73.4**	**70.8**	52.5	51.3	**70.3**	**63.1**	**62.24**
Improvements	+7.4	+7.1	+13.7	+7.7	+3.6	+1.1	+5.9	+2.5	+6.13

Table 2: Ablation studies on different loss.

Method	AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)		Avg.
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
Ours	**57.1**	**59.4**	**73.4**	**70.8**	**52.5**	**51.3**	**70.3**	63.1	**62.24**
w/o L_{align}	55.4	58.9	65.6	68.1	52.3	51.1	67.6	**63.3**	60.29
w/o L_{grl}	47.3	50.2	58.1	62.9	49.6	48.2	66.4	61.8	55.56
$L_{grl} \rightarrow L_{CLUB}$	54.2	56.4	64.9	65.5	52.4	50.8	65.1	62.7	59.00

Table 3: Ablation studies on different structures.

Method	AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)		Avg.
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
Ours (audio semantic anchor)	**57.1**	**59.4**	**73.4**	**70.8**	**52.5**	51.3	**70.3**	**63.1**	**62.24**
Symmetric audio-video structure	51.5	52.4	70.1	67.4	51.3	**51.6**	65.4	61.5	58.90
Video semantic anchor	54.7	54.9	69.0	66.1	49.7	48.1	66.5	62.3	58.91

Overall, these results suggest that our approach can more effectively extract shared semantics from pure audio–video pairs and support robust cross-modal transfer.

4.2 Ablation Study

Compared with traditional CMG models, AHA differs in three aspects: **(i)** an asymmetric architecture with an audio semantic anchor; **(ii)** GRL-based adversarial decoupler instead of MI-estimation-based decoupler; **(iii)** a local sliding alignment objective for fine-grained temporal alignment. We ablate each component while keeping the CMG protocol and downstream heads unchanged. Since the impact of Cross-CPC and MM-EMA has been discussed in previous work, we do not revisit it here.

Local Sliding Alignment (w/o L_{align}): As shown in Table 2, removing L_{align} consistently degrades performance, with the largest drop on AVVP. This matches our motivation that local sliding alignment provides a tolerant temporal correspondence constraint, sharpening segment-level audio–video matching under small offsets. Notably, the model remains clearly stronger than Unicode even without this term, suggesting that gains are not merely from adding another contrastive loss. Rather, L_{align} mainly refines fine-grained alignment on top of a transferable backbone.

Disentanglement (w/o L_{grl}): Removing GRL-based decoupler causes the largest degradation, especially on AVE and AVVP. Under CMG, the classifier is trained on the source modality only, without explicit disentanglement, modality-identifying cues can leak into the modality-specific branches during source training, creating a pronounced train–test mismatch when transferred to the target modality. These results indicate that L_{grl} is crucial for suppressing modality-specific leakage and enabling reliable zero-shot transfer.

Replacing GRL with CLUB ($L_{\text{grl}} \rightarrow L_{\text{CLUB}}$): Replacing GRL with CLUB partially recovers the drop but still underperforms the full model, with notable gaps on AVVP. This suggests that GRL-based adversarial decoupler provides a more effective and stable training signal than MI upper-bound minimization.

After validating the contributions of L_{infonce} and GRL disentanglement, we further ablate the architecture along two axes: **(i)** symmetric vs. asymmetric structure; **(ii)** the choice of semantic anchor Table 3.

Symmetric vs. Asymmetric: The symmetric dual audio–video structure (Details are shown in Appendix B) is consistently weaker than our asymmetric design, with clear drops on AVE and UCF(v) ↔ VGG(a). This is consistent with our analysis of allocation ambiguity, without a directed anchor, semantic information can drift into modality-specific branches, weakening the shared code used for cross-modal transfer.

Choice of Anchor: Within the asymmetric framework, the audio anchor yields the strongest overall transfer and improves most on AVVP, where fine-grained temporal alignment is critical. Under dataset shift (AVE → AVVP), the video anchor degrades noticeably, while the audio anchor remains substantially higher, the symmetric structure is comparable on A → V but is weaker overall due to its larger drops on AVVP. These results support audio anchoring as a directional constraint that reduces semantic allocation ambiguity and yields a more transferable modality-invariant code in CMG tasks.

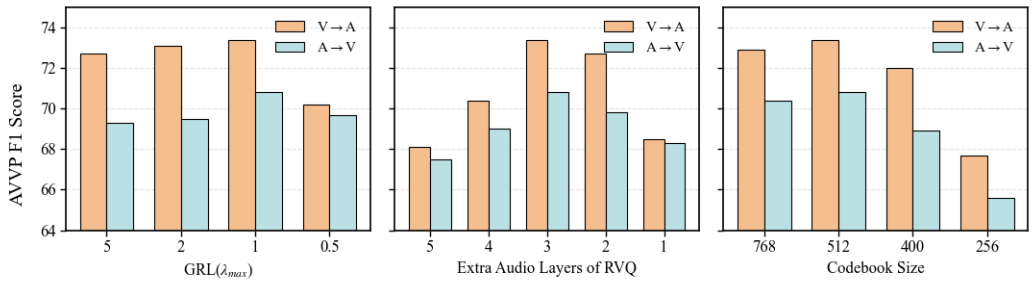

Figure 3: The impact of GRL's adversarial intensity (larger λ_{max} indicates stronger adversarial effect), RVQ's number of extra audio layers and Codebook size on experimental accuracy. Additional ablations on the anchor-layer choice k and LSA hyperparameters are provided in Appendix D.

4.3 Talking Face Disentanglement Experiment



Figure 4: (a) **Talking Face Disentanglement Experiment Results**: We selected two videos from sources not in the training set to demonstrate the effect. To better showcase the experimental results, we selected frames where Specific Video and Semantic Video have significant differences (for example: whether eyes are closed and mouth is open) for comparison. We set the identity information as the corresponding identity vector from the Semantic Video. For comparison, we also tested (i) removing GRL from our asymmetric structure (w/o L_{grl}). (ii) replacing GRL with CLUB in the asymmetric structure($L_{\text{grl}} \rightarrow L_{\text{CLUB}}$). (iii) adding GRL to a symmetric structure identical to Unicode (Symmetric). (Please see the Appendix C for more experimental results); (b) **PCA & UMAP Results**: For the same selected video features, we extracted specific features and semantic features, and performed Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP).

To address the limitation that previous CMG experiments could only implicitly observe semantic information and disentanglement through data, which lacks intuitiveness, we propose a novel Talking Face Disentanglement Experiment that explicitly tests the model's disentanglement capability.

Experimental Settings: Based on our original pipeline Figure 2(a), we use the motion encoder from LIA [53, 54] as the video backbone to extract identity information from the input target video that represents its identity, and extracts video features that contain highly entangled motion information of various facial features. In addition, we use Wav2Vec2.0 [55] as the audio backbone to extract audio features. To better demonstrate the effect of modality transfer, consistent with the CMG experiment approach, we add very sparse phoneme supervision on the time step at the audio shared RVQ side of all models. We selected a filtered TalkVid [56] subset (approximately 30.6 hours) and trained for approximately ten hours on four A6000 GPUs (Please see the Appendix C for more implementation details).

Expected Results: According to our hypothesis about the Talking Face Disentanglement Experiment, the video information shared with audio should be related to mouth movements, while information representing other facial actions should flow to modality-specific branches. Therefore, the ideal effect of our Talking Face Disentanglement Experiment is to achieve transfer of mouth movements. To elaborate, we can input a semantic video and extract the motion vector representing audio-related mouth movements after passing through the shared RVQ in its pipeline. For another specific video, we can extract the motion vector representing non-mouth movements (such as eye blinking) after passing through the specific encoder in its pipeline. These two can be concatenated and output, then passed through LIA and concatenated with the target identity information, and rendered through LIA's renderer. Finally, a fused video is generated that combines the mouth movements from the semantic video and audio-unrelated movements from the specific video. This also validates the model's decoupling ability, namely whether mixed features common to both videos will appear.

Result Analysis: From Figure 4(a), we can see that our current AHA structure can effectively decouple information, enabling alignment of audio-irrelevant actions (such as eye movements) in specific video and alignment of mouth movements in semantic video, with minimal information leakage. When removing the GRL loss (w/o L_{grl}), the disentanglement effect becomes significantly worse, as evidenced by many cases where mouth movements are closed at the same frame in semantic video, while the frame without GRL has the mouth open, and vice versa. There are also cases where eye movements do not correspond to the corresponding frames in specific video. After replacing GRL with CLUB ($L_{\text{grl}} \rightarrow L_{\text{CLUB}}$) and retraining, the improvement in overall disentanglement effect is not obvious. This demonstrates that variational MI estimators are weaker than GRL in handling information disentanglement in such complex information environments. For the symmetric structure, despite strong decoupling with GRL, the lack of audio as an anchor point for semantic information causes most information to leak into the modality-specific branches, resulting in poor disentanglement performance.

4.3.1 Further Analysis

Table 4: Quantitative comparison in Visual-to-Visual Lip Synchronization (V2V-LS) and Mouth Landmark RMSE.

	Ours	w/o L_{grl}	$L_{\text{grl}} \rightarrow L_{\text{CLUB}}$	Symmetric
V2V-LS ↓	**5.98**	6.77	6.82	6.40
Mouth RMSE ↓	**5.51**	16.59	18.26	14.61

Table 5: Quantitative comparison in Peak Signal-to-Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS).

	Ours	w/o L_{grl}	$L_{\text{grl}} \rightarrow L_{\text{CLUB}}$	Symmetric
PSNR ↑	**29.42**	26.46	27.49	26.76
LPIPS ↓	**0.0468**	0.0703	0.0575	0.0730

PCA reflects global variance directions. As shown in Figure 4, AHA yields clearly different variance patterns with limited overlap between semantic and specific features, suggesting weaker cross-branch leakage. Removing GRL or replacing it with CLUB results in heavier overlap, while the symmetric structure is dominated by the higher-variance branch, consistent with feature collapse rather than structural disentanglement. UMAP further confirms this: under AHA, semantic and specific features show more separated neighborhoods, whereas other variants exhibit highly mixed local structures and similar manifold backbones.

We further evaluate lip synchronization via V2V-LS (frame-wise lip semantic embedding differences [57]) and Mouth RMSE (geometric misalignment of 2D mouth landmarks), and reconstruction quality via PSNR and LPIPS. As shown in Table 4 and Table 5, AHA achieves the best performance across all four metrics, indicating improved semantic coherence, reconstruction fidelity, and perceptual quality. Detailed metric definitions are provided in Appendix C.

5 Conclusion

This paper investigates audio-visual joint representation learning under Cross-Modal Generalization (CMG), noting that current symmetric discrete unified representation frameworks are prone to information allocation ambiguity. We propose Asymmetric Hierarchical Anchoring (AHA), which leverages the hierarchical structure of audio RVQ to construct shared semantic anchors for directed video-to-audio semantic distillation. AHA further introduces a GRL-based adversarial decoupler to suppress semantic information leakage and Local Sliding Alignment (LSA) to improve fine-grained

temporal alignment. Extensive experiments on AVE, AVVP, cross-dataset transfer, and a Talking Face Disentanglement Experiment show that AHA improves cross-modal generalization, disentanglement quality, and reconstruction fidelity.

Limitations: While AHA demonstrates strong performance under the current CMG setting, this work mainly focuses on audio-visual representation learning with a fixed semantic anchor and standard downstream evaluation protocols. Future work will explore adaptive anchor selection across tasks and modality pairs, extend AHA to broader multimodal settings, and conduct more systematic evaluations under open-domain noise, weak cross-modal correspondence, and long temporal contexts.

References

[1] Antol, S., A. Agrawal, J. Lu, et al. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433. 2015.

[2] Lei, J., L. Yu, M. Bansal, et al. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379. 2018.

[3] Li, G., Y. Wei, Y. Tian, et al. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118. 2022.

[4] Tian, Y., J. Shi, B. Li, et al. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[5] Xu, H., R. Zeng, Q. Wu, et al. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3893–3901. 2020.

[6] Geng, T., T. Wang, J. Duan, et al. Uniav: Unified audio-visual perception for multi-task video event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[7] Radford, A., J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[8] Girdhar, R., A. El-Nouby, Z. Liu, et al. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190. 2023.

[9] Duan, J., L. Chen, S. Tran, et al. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15651–15660. 2022.

[10] Liu, A., S. Jin, C.-I. Lai, et al. Cross-modal discrete representation learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3013–3035. 2022.

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

[12] Chen, Z., C. Wang, X. Chen, et al. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. *arXiv preprint arXiv:2503.06764*, 2025.

[13] Snell, J., K. Swersky, R. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[14] Chen, D., Z. Wu, F. Liu, et al. Protoclip: Prototypical contrastive language image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[15] Xia, Y., H. Huang, J. Zhu, et al. Achieving cross modal generalization with multimodal unified representation. In *Advances in Neural Information Processing Systems*, vol. 36, pages 63529–63541. 2023.

[16] Huang, H., Y. Xia, S. Ji, et al. Enhancing multimodal unified representations for cross modal generalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2353–2366. 2025.

[17] Cheng, P., W. Hao, S. Dai, et al. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[18] Aytar, Y., C. Vondrick, A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[19] Narasimhan, M., S. Ginosar, A. Owens, et al. Strumming to the beat: Audio-conditioned contrastive video textures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3761–3770. 2022.

[20] Biner, B. C., F. M. Sofian, U. B. Karakaş, et al. Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878*, 2024.

[21] Cheng, H., L. Lin, C. Liu, et al. Dawn: Dynamic frame avatar with non-autoregressive diffusion framework for talking head video generation. *arXiv preprint arXiv:2410.13726*, 2024.

[22] Gao, X., L. Hu, S. Hu, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

[23] Lee, D., C. Kim, S. Kim, et al. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532. 2022.

[24] Défossez, A., J. Copet, G. Synnaeve, et al. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[25] Kim, J., T. Moon, K. Lee, et al. Efficient generative modeling with residual vector quantization-based tokens. *arXiv preprint arXiv:2412.10208*, 2024.

[26] Zhang, R., X. Hao, Y. Han, et al. Mbcodec: Thorough disentangle for high-fidelity audio compression. *arXiv preprint arXiv:2509.17006*, 2025.

[27] Kim, H., J. Lee, J. Morton, et al. Improving test-time performance of rvq-based neural codecs. *arXiv preprint arXiv:2509.19186*, 2025.

[28] Petridis, S., T. Stafylakis, P. Ma, et al. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018.

[29] Andonian, A., S. Chen, R. Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441. 2022.

[30] Sarkar, P., A. Etemad. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pages 14875–14885. 2024.

[31] Chen, Y.-C., L. Li, L. Yu, et al. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[32] Wang, T., W. Jiang, Z. Lu, et al. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022.

[33] Huo, F., W. Xu, J. Guo, et al. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015. 2024.

[34] Jeong, J., S. Lee, D. Park, et al. Multi-modal knowledge distillation-based human trajectory forecasting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24222–24233. 2025.

[35] Wang, Z., Y. Zhao, H. Huang, et al. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023.

[36] Grassucci, E., G. Cicchetti, D. Comminiello. Closing the modality gap enables novel multimodal learning applications. In *Second Workshop on Representational Alignment at ICLR 2025*. 2025.

[37] Jin, Y., K. Xu, L. Chen, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

[38] Yu, Q., M. Weber, X. Deng, et al. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024.

[39] Oord, A. v. d., Y. Li, O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[40] Belghazi, M. I., A. Baratin, S. Rajeshwar, et al. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[41] Ganin, Y., V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[42] Zhang, B. H., B. Lemoine, M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. 2018.

[43] Yu, Y., K. H. R. Chan, C. You, et al. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in neural information processing systems*, 33:9422–9434, 2020.

[44] Zhang, Y., L. Yan, Z. Qin, et al. Towards disentangling relevance and bias in unbiased learning to rank. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5618–5627. 2023.

[45] Ju, Z., Y. Wang, K. Shen, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

[46] Zhou, J., L. Zheng, Y. Zhong, et al. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8436–8444. 2021.

[47] Zhou, J., D. Guo, M. Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2023.

[48] Zverev, D., T. Wiedemer, A. Prabhu, et al. Vggsounder: Audio-visual evaluations for foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1027–1037. 2025.

[49] Tian, Y., D. Li, C. Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision – ECCV 2020*, pages 436–454. 2020.

[50] Soomro, K., A. R. Zamir, M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[51] You, H., L. Zhou, B. Xiao, et al. Learning visual representation from modality-shared contrastive language-image pre-training. In *Computer Vision – ECCV 2022*, pages 69–87. 2022.

[52] Zhao, Y., C. Zhang, H. Huang, et al. Towards effective multi-modal interchanges in zero-resource sounding object localization. In *Advances in Neural Information Processing Systems*, vol. 35, pages 38089–38102. 2022.

[53] Wang, Y., D. Yang, F. Bremond, et al. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[54] Ki, T., D. Min, G. Chae. Float: Generative motion latent flow matching for audio-driven talking portrait. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14699–14710. 2025.

[55] Baevski, A., Y. Zhou, A. Mohamed, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[56] Chen, S., H. Huang, Y. Liu, et al. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. *arXiv preprint arXiv:2508.13618*, 2025.

[57] Shi, B., W.-N. Hsu, K. Lakhotia, et al. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

[58] Chung, J. S., A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[59] Rehman, A., J. Cai, J.-J. Zhang, et al. Bfa: Real-time multilingual text-to-speech forced alignment. *arXiv preprint arXiv:2509.23147*, 2025.

[60] Hempel, T., A. A. Abdelrahman, A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022.

[61] Horé, A., D. Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. 2010.

[62] Zhang, R., P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595. 2018.

[63] Prajwal, K., R. Mukhopadhyay, V. P. Namboodiri, et al. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492. 2020.

[64] Devlin, J., M.-W. Chang, K. Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186. 2019.

[65] Mirsky, Y., W. Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

A Implementation of Adversarial Disentanglement(Section 3.3)

Algorithm 1 Adversarial Disentanglement with Velocity-Aware Sampling

Require: Video Encoders E_{v_sem}, E_{v_spec}; Discriminator D_ϕ; Batch $\mathcal{B}$.
Hyperparameters: Max steps N_{max}; Temp τ; Smoothing ϵ; GRL Limit λ_{max}; Learning rate α;
Model parameters θ.
Helper Functions: Score $s(\mathbf{x}, \mathbf{y}) = D_\phi(\mathcal{R}_\lambda(\mathbf{x}), \mathbf{y})$.
Ensure: Disentangled Specific Features $\mathbf{z}_{spec}$.
Initialize step $p \leftarrow 0$
while $p < N_{max}$ **do**
1. Feature Extraction
$V \leftarrow$ Quantizer$(E_{v_sem}(\mathcal{B}))$; {Video Units}
$v \leftarrow E_{v_spec}(\mathcal{B})$ {Video Specific Features}
2. Velocity-Aware Sampling
for each sequence in batch **do**
$\delta_t \leftarrow \|V_{t+1} - V_t\|_2$
$P(t) \leftarrow \delta_t / (\sum \delta_\tau + \epsilon)$;
Sample $\mathcal{K} \sim$ Multinomial$(P(t))$
$\hat{V}, \hat{v} \leftarrow$ Gather$(\{V, v\}, \mathcal{K})$
end for
3. Adversarial Objective
$\lambda \leftarrow \lambda_{max} \cdot \left(\frac{2}{1+\exp(-10 \cdot p/N_{max})} - 1 \right)$
Sample negatives $\mathcal{N}$ from batch
$\mathcal{L}_{grl} \leftarrow -\log \frac{\exp(s(\hat{v}, \hat{V})/\tau)}{\sum_{V' \in \{\hat{V}\} \cup \mathcal{N}} \exp(s(\hat{v}, V')/\tau)}$
4. Optimization
{Update Discriminator (Minimize Loss)}
$\theta_D \leftarrow \theta_D - \alpha \nabla_{\theta_D} \mathcal{L}_{grl}$
{Update Specific Encoder (Maximize Loss via GRL)}
$\theta_{E_{v_spec}} \leftarrow \theta_{E_{v_spec}} - \alpha \lambda \nabla_{\theta_{E_{v_spec}}} \mathcal{L}_{grl}$
{gradient is negated by $\mathcal{R}_\lambda$ during backprop, achieving effective gradient ascent on E_{v_spec}}
$p \leftarrow p + 1$
end while

B Details of CMG Downstream Tasks

Pre-train Details:

Following the standard CMG evaluation protocol, we adopt the same backbone architectures as Unicode [15]. Specifically, for each 1-second visual segment we uniformly sample 16 RGB frames and extract deep convolutional feature maps from VGG-19 after the final pooling stage; we then apply global average pooling across the 16 frames to obtain $7 \times 7 \times 512$-D visual feature maps. For the audio stream, we employ a VGG-like network pre-trained on AudioSet to extract 128-D features for each 1-second audio segment. For both modalities, we further project semantic representations into a shared 512-D embedding space.

To capture modality-specific information, the video branch uses a convolutional encoder to produce spatial feature maps of shape $3 \times 3 \times 2048$; we apply spatial average pooling followed by a linear projection to obtain video-specific features of shape $T \times 512$. The audio branch uses a linear encoder to produce audio features of shape $T \times 256$. The video-specific features are used for reconstruction and for adversarial disentanglement via the GRL-based decoupler (operating against video semantic units). The audio and video semantic units are aligned via Local Sliding Alignment. In contrast to prior unified-representation approaches (e.g., DCID [16], FCID [16]) that rely on additional text supervision, our Asymmetric Hierarchical Anchoring (AHA) framework is pre-trained solely on paired audio–video data in this downstream tasks.

For Cross-CPC, we set the prediction horizon to one step and apply it between the two modality semantic sequences during unified-representation pre-training. For GRL-based disentanglement,

we first update the GRL discriminators once with frozen encoders, and then update the full model (encoder, CPC, decoder) with GRL reversal in the main pass. We use a learning rate of 1×10^{-4}, set $\gamma = 0.99$ for MM-EMA, and adopt a batch size of 96.

Details of all downstream tasks:

Table 6: Details of all downstream tasks.

Task	Pretrained Modality	Downstream Dataset	Generalization Direction
Cross-Modal Event Classification	Audio-Visual	AVE	A→V V→A
Cross-Modal Event Localization	Audio-Visual	AVVP	A→V V→A
Cross Modal & Dataset Event Localization	Audio-Visual	AVE & AVVP	A(AVE)→V(AVVP) V(AVE)→A(AVVP) UCF(v)↔VGG(a)

Cross-Modal Event Classification (AVE) [4]: The AVE dataset contains 28 event categories and each audio/video clip is 10 seconds long. Given discrete representations extracted from one modality, we attach a two-layer MLP classifier to map the sequence into 28-way class logits, followed by a softmax and a standard cross-entropy loss. To test video-to-audio generalization (V→A), we train the classifier using video-derived discrete sequences and then directly replace the input with audio-derived sequences at test time, keeping both the frozen encoder and the learned classifier unchanged; audio-to-video (A→V) is evaluated analogously.

Cross-Modal Event Localization (AVVP) [49]: The AVVP dataset includes 25 event categories, and a single clip may contain multiple events. We again use the frozen encoder to obtain a length-T discrete sequence, and employ a two-layer MLP localization head to predict 25-dimensional event scores (per segment). We apply a sigmoid activation and optimize with a multi-label classification objective (binary cross-entropy) against the ground-truth labels. All remaining training settings follow those used for cross-modal event classification.

Cross-Modal & Cross-Dataset Event Localization (AVE → AVVP; UCF(v)↔VGG(a)): To further assess transferability across datasets, we consider the 12 overlapping event categories shared by AVE and AVVP: dog, car, helicopter, violin fiddle, frying food, motorcycle, acoustic guitar, banjo, baby cry, chainsaw, cat, accordion. We train the localization head using single-modality inputs from the AVE training set and directly evaluate on the opposite modality in the AVVP validation set, without any additional adaptation. Following prior practice, we report the F1 score as the primary metric, while keeping the optimization protocol identical to the AVVP localization setting. We also tested cross-modal classification tasks, performing classification between the visual modality on a subset of UCF-101 [50] (16 classes) and the audio modality on a subset of VGGSound-AVEL [46, 47] (16 classes).

Brief Pipeline of symmetric structure: Below is a schematic diagram of the symmetric structure in CMG Downstream Tasks and in the Talking Face Disentanglement Experiment. Except for slight differences in the peripheral structure (i.e., audio or video backbone, encoder, decoder), the main structure is shown in Figure 5.

Table 7: Extension to AVT setting.

Method	AVE		AVVP		AVE→AVVP		UCF↔VGG		Avg.
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
DCID (AVT)	54.5	55.0	40.9	41.6	56.5	53.6	68.1	61.7	53.99
FCID (AVT)	55.9	55.0	43.6	45.1	57.4	58.5	69.6	62.0	55.89
Ours (AVT)	**59.1**	**61.7**	**73.9**	**72.1**	**59.5**	**60.3**	**73.5**	**66.8**	**65.86**

Extension to Text–Vision–Audio (AVT) Setting: Following DCID/FCID [16], we incorporate BERT-based text features while keeping audio as the semantic anchor. As shown in Table 7, AHA (AVT) achieves the best results across all metrics, confirming extensibility beyond the pure audio–visual setting.

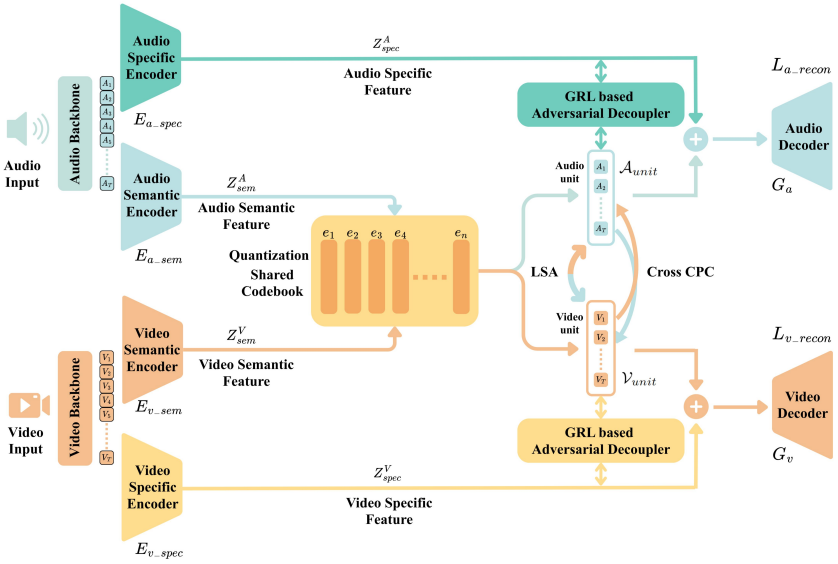

Figure 5: Brief Pipeline of symmetric structure

C Details of Talking Face Disentanglement Experiment

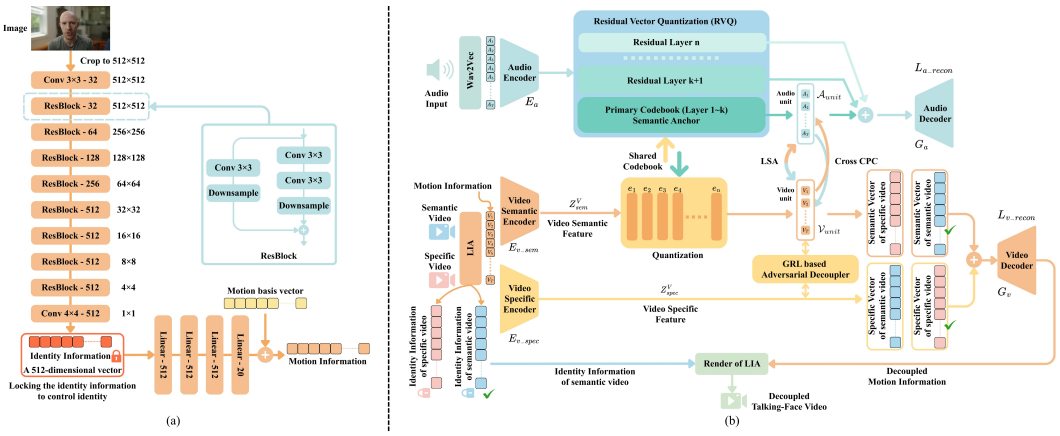

Figure 6: (a) LIA's Brief Pipeline for Extracting Identity Information and Motion Information; (b) Brief Pipeline for Talking Face Disentanglement Experiment

Based on our initial AHA Pipeline, we replaced the Video backbone with LIA [53, 54] and the Audio Backbone with Wav2Vec2.0 [55]. The pipeline for the Talking-Face Disentanglement Experiment is shown in Figure 6(b). For the Video Backbone, LIA's encoder extracts Identity Information from the input target image that represents identity of this image, and extracts Motion Information containing highly entangled facial features from it. We use this Motion Information as the video feature, with the detailed pipeline shown in Figure 6(a). Additionally, for the Audio Backbone, we used Wav2Vec to extract audio features. After processing with LIA and Wav2Vec, the final video input dimension and audio input dimension are both 512-D.

Specifically, for our decoupling process, we select a video not present in the training data as the Semantic Video. After passing through the Encoder of LIA, we obtain the Identity Information of Semantic Video and Motion Information of Semantic Video. Additionally, we select another video also not present in the training data as the Specific Video. After similarly passing through the Encoder of LIA, we obtain the Identity Information of Specific Video and Motion Information of Specific Video. We freeze the obtained Identity Information and input the Motion Information into our AHA, obtaining the Semantic Vector of Semantic Video, Semantic Vector of Specific Video,

Specific Vector of Semantic Video, and Specific Vector of Specific Video respectively. Finally, we select the Semantic Vector of Semantic Video and Specific Vector of Specific Video, concatenate them, and send them to the Decoder to obtain the Decoupled Motion Information. Together with the frozen Identity Information of Semantic Video, we input them into the Renderer of LIA to finally obtain our Decoupled Talking-Face Video.

Therefore, the expected effect is to generate a fused Decoupled Talking-Face Video that combines the mouth movements obtained from the Semantic Video with the audio-independent motions (such as blinking actions) from the Specific Video.

For our training details, we selected a subset of the TalkVid [56] dataset (approximately 38.4 hours). After our data processing, we ultimately used approximately 30.6 hours of data and trained for approximately 10 hours across 4 A6000 GPUs, covering 250 epochs.

Data Processing Details:

1. Face Recognition: Due to transitions or narration in YouTube videos, TalkVid may contain video segments with audio but no actual human faces. Therefore, we used Face Recognition to detect whether the video contains human faces and removed videos without faces.
2. Silence Detection: Since some videos contain prolonged silence (exceeding 1/3 of the video duration), which is useless for training purposes, we also performed silence detection on the videos.
3. Audio-Visual Synchronization: Since most data in TalkVid comes from YouTube, some videos are similar to film reviews or commentary videos that contain facial action videos, but the audio and corresponding video cannot be properly synchronized. Based on this problem, we made certain adjustments based on the Python code from [58], and filtered out videos that do not meet our requirements for audio-visual synchronization.
4. Phoneme Extraction: To better demonstrate the modality transfer effect, consistent with the CMG experiment approach, we add a phoneme supervision on the audio shared RVQ that is very sparse in the time dimension (statistically around 20% of frames participate), i.e., adding a simple classification head to perform phoneme classification on labeled frames. We use BFA [59] to extract phonemes.
5. Large Angle Face Detection: Since some videos contain significant angle changes in face orientation, mouth information may be lost due to large-angle movements, which affects our experiments. Therefore, we used [60] to detect large angles in videos and removed videos where face rotation exceeds 60 degrees from the dataset.

More Talking Face Disentanglement Experiment Examples:

More experimental results can be found in Figure 8 and Figure 9. From the extensive experimental results, we can see the robustness of our model in terms of decoupling stability and reconstruction fidelity.

Detailed diagrams of PCA and UMAP:

To more clearly see the individual distribution of specific features and semantic features respectively, we present the PCA and UMAP details of specific feature only and semantic feature only in Figure 7.

Evaluation Metrics:

1. **PSNR** [61]: Peak Signal-to-Noise Ratio (PSNR) measures the reconstruction quality of an image by comparing it with the reference image, higher values indicate better fidelity.
2. **LPIPS** [62]: Learned Perceptual Image Patch Similarity (LPIPS) is used to measure the perceptual similarity between reconstructed image and real image based on the pre-trained AlexNet features.
3. **V2V-LS (Ours)**: We propose a video-to-video lip-sync metric inspired by Wav2Lip [63], using visual embeddings from AV-HuBERT [57]. The lip region embeddings of the two videos are compared via frame-wise differences to compute a score, quantifying how well the source video drives the target lip movements.

4. **Mouth RMSE**: Root-mean-square error (RMSE) computed over the 2D mouth landmarks of two videos, capturing frame-wise geometric misalignment of the lips; lower values indicate better lip synchronization.

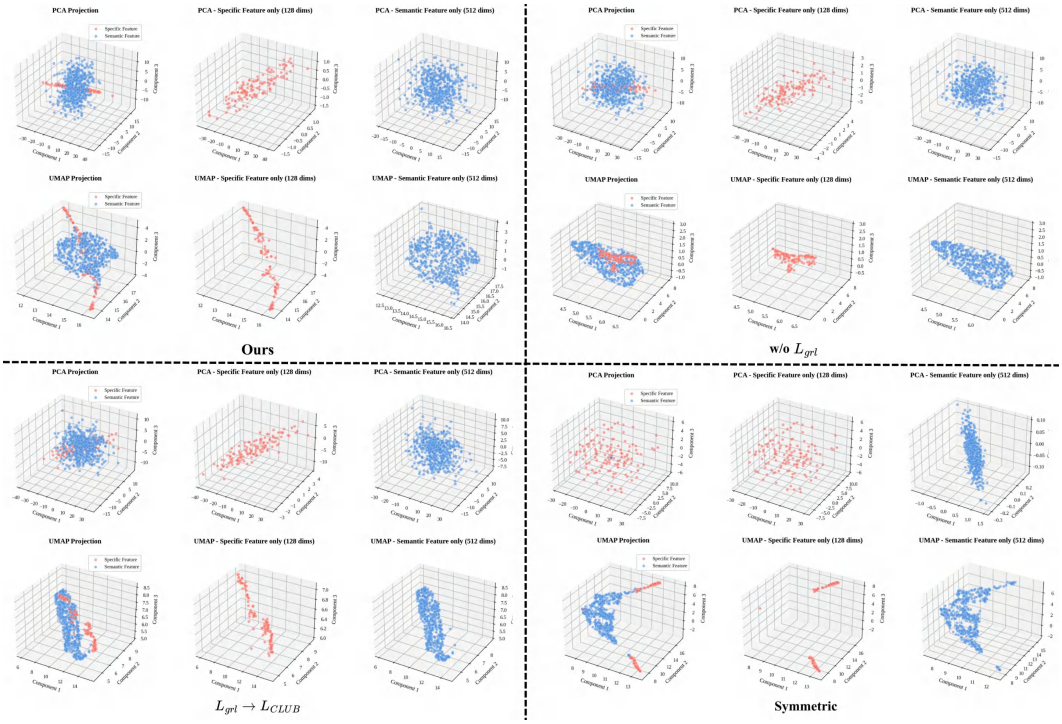

Figure 7: Detailed diagrams of PCA and UMAP

D More Ablation Studies

Sensitivity to Anchor Layer k. Table 8 sweeps $k \in \{1, 2, 3, 4\}$. Performance is robust across all values; k=1 and k=2 yield the best results, consistent with the hypothesis that primary semantic content concentrates in lower RVQ layers. We use k=1 for CMG (fair comparison with single-codebook baselines) and k=2 for Talking Face (higher semantic density required).

Table 8: Sensitivity to shared RVQ anchor layers k.

k	AVE V→A	AVE A→V	AVVP V→A	AVVP A→V	V2V-LS ↓	PSNR ↑
1	57.1	**59.4**	73.4	70.8	6.43	28.01
2	**57.6**	58.9	**73.5**	**71.4**	**5.98**	**29.42**
3	56.7	58.2	72.9	71.3	6.14	28.72
4	56.4	58.3	71.7	70.2	6.77	27.65

Sensitivity to LSA Hyperparameters. Table 9–Table 10 sweep R and n_{pos}. Performance plateaus for $R \geq 5$; only R=1 (overly narrow scope) hurts. n_{pos}=1 is optimal: larger tolerance blurs the alignment target. Both parameters follow simple task-granularity guidelines and require no per-dataset tuning.

E Training Details

Table 11 shows a portion of the hyperparameters used during our training process and their corresponding values. Figure 10 shows the GRL loss during our training process. As can be seen, after a

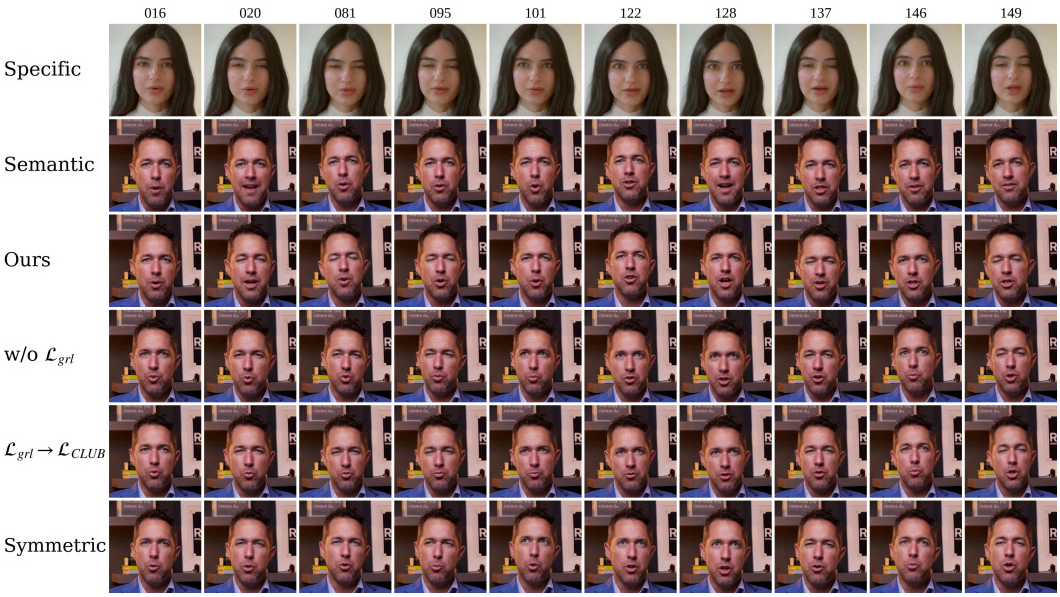

Figure 8: More Talking Face Disentanglement Experiment Examples (1)

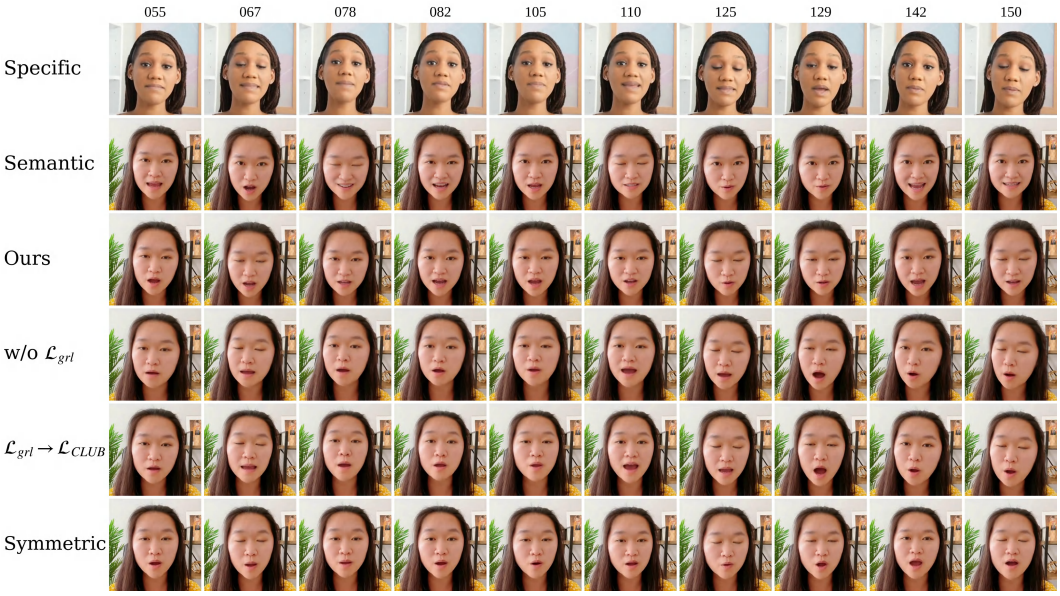

Figure 9: More Talking Face Disentanglement Experiment Examples (2)

Table 9: LSA search scope R sweep (fixed n_{pos}=1).

R	AVE V→A	AVE A→V	AVVP V→A	AVVP A→V
1	52.6	54.7	66.7	67.9
3	55.9	58.3	72.1	69.2
5 (ours)	57.1	**59.4**	73.4	**70.8**
7	**57.2**	59.2	**73.6**	70.6

Table 10: LSA positive tolerance n_{pos} sweep (fixed R=5).

n_{pos}	AVE V→A	AVE A→V	AVVP V→A	AVVP A→V
1 (ours)	**57.1**	**59.4**	**73.4**	**70.8**
2	56.1	57.8	70.5	68.4
3	53.1	54.9	65.2	66.7

certain number of steps, the GRL loss stabilizes and fluctuates around 4.28. This value is determined by $ln(batchsize \times GRL\ negative\ sample\ ratio) = ln(96 \times 0.75) = 4.28$ in our experiment. Fluctuations around this value indicate that the GRL adversarial process is proceeding normally during our model training.



Figure 10: Changes in GRL Loss During Training

Training Efficiency: All models were trained for 250 epochs on 4 NVIDIA A6000 GPUs with batch size 96. Our proposed AHA completed training in approximately **9.7** hours, while symmetric structure required **13.3** hours, demonstrating a **27%** improvement in training efficiency.

F Existing Assets and Licenses

We use existing datasets, pretrained components, and reference implementations only for non-commercial academic research and evaluation. We cite the original papers and official sources where applicable, and follow the corresponding licenses or terms of use. The main assets used in this work are summarized below.

- **AVE** [4]: Used for cross-modal event classification and AVE-to-AVVP transfer evaluation. The official release provides dataset access and citation information, but we did not find an explicit public dataset license. We therefore use it only for academic evaluation and cite the original paper.
- **AVVP / LLP** [49]: Used for cross-modal event localization. The official project repository is released under the GNU General Public License v3.0. We use the released annotations/features only for academic evaluation and cite the original paper.
- **VGGSound and VGGSound-AVEL** [46, 47]: Used for audio–visual pre-training and cross-dataset evaluation. VGGSound is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0), while the copyright of the original videos remains with their respective owners.

Table 11: Some hyperparameters and their values.

Hyperparameter	Value
Codebook size	512
Embedding dimension	512
CPC hidden dimension	256
CPC context dimension	256
CPC LSTM layers	2
CPC prediction steps	1
GRL velocity sample ratio	0.4
GRL negative sample ratio	0.75
Main optimizer	AdamW
GRL optimizer	Adam (learning rate = 1e-4)
Learning rate scheduler	CosineAnnealingWarmupLR
Warmup epochs	5
Minimum learning rate	1e-6
GRL temperature ratio τ	0.1
GRL λ_{max}	1
Audio extra layers	3
Positive tolerance set (CMG Tasks)	1
Positive tolerance set (Talking Face Experiment)	3
LSA window size (CMG Tasks)	5
LSA window size (Talking Face Experiment)	31
Shared RVQ Layers (CMG Tasks)	1
Shared RVQ Layers (Talking Face Experiment)	2

- **UCF-101** [50]: Used for cross-dataset visual-to-audio evaluation. The official website provides the dataset for research use, but we did not find an explicit public license in the official release. We use it only for academic evaluation and cite the original dataset paper.
- **TalkVid** [56]: Used in the Talking Face Disentanglement Experiment. The official repository states that the dataset is released under CC BY-NC 4.0 for non-commercial research use, and that the accompanying code is released under the Apache License 2.0.
- **LIA** [53]: Used as the motion encoder/renderer in the diagnostic talking-face disentanglement evaluation. The official repository is released under the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0). We use it only for non-commercial research.
- **Wav2Vec2.0 / fairseq** [55]: Used as the audio backbone in the talking-face experiment. The fairseq repository is MIT-licensed and states that the license applies to pretrained models as well.
- **VGG-19 and AudioSet-pretrained audio features**: For CMG experiments, we follow the standard protocol and use VGG-19 visual features and AudioSet-pretrained audio features. TorchVision is released under the BSD 3-Clause License, and AudioSet annotations are released under CC BY 4.0, with the AudioSet ontology under CC BY-SA 4.0.
- **BERT-based text features** [64]: Reference AVT baselines use BERT-based text features. The official Google BERT repository states that its code and pretrained models are released under the Apache License 2.0.

For datasets derived from online videos, we use only the released annotations, features, or clips for academic evaluation and respect the terms of the original dataset providers and source platforms.

G Broader Impacts

This work studies audio–visual joint representation learning under Cross-Modal Generalization (CMG). Its potential positive impact lies in improving the robustness of audio–visual representation learning under modality shift and reducing reliance on extensive labeled data, which may benefit downstream audio–visual understanding, localization, and transfer-learning applications.

This work also includes a Talking Face Disentanglement Experiment as a diagnostic evaluation of semantic–specific disentanglement. In this experiment, we use an existing renderer to visualize fused

facial motions by combining mouth-related motion from one sequence with other facial motions from another sequence. Although this setting is used only for representation analysis rather than for releasing a deployable face-generation system, related techniques could potentially be misused for synthetic media manipulation, impersonation, or non-consensual deepfakes. The societal risks of deepfakes have been discussed in prior work [65].

We therefore recommend responsible use of this line of research, including using data with appropriate consent and governance, clearly disclosing or labeling synthetic outputs, and conducting safety evaluations before any deployment in controllable audio–visual generation systems. We strongly condemn malicious use and advocate responsible and ethical research practices.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the abstract, introduction, and conclusion, including AHA, the audio-RVQ semantic anchor, the GRL-based adversarial decoupler, and Local Sliding Alignment. These claims are supported by experiments on AVE, AVVP, cross-dataset transfer, ablation studies, and the Talking Face Disentanglement Experiment.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a Limitations paragraph in the conclusion. It discusses that the current study focuses on audio-visual representation learning with a fixed semantic anchor and standard downstream protocols, and points to future evaluations under broader multimodal, open-domain, noisy, weakly aligned, and long-context settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or lemmas requiring proofs. The mathematical formulations are used to define the model architecture, loss functions, and training objectives.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the model architecture, loss functions, training pipeline, downstream evaluation protocol, datasets, baselines, ablation settings, and implementation details in the main paper and appendices. The paper provides a reasonable avenue for reproducing the main experimental setup through detailed methodological and experimental descriptions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper uses existing datasets and pretrained components that are cited in the paper. The paper provides detailed architecture, training, and evaluation descriptions to support reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main training and testing settings, including datasets, backbones, transfer directions, evaluation metrics, downstream heads, learning rate, batch size, MM-EMA decay, training epochs, and hyperparameter sensitivity studies. Additional details are provided in the appendices for CMG downstream tasks, the Talking Face Disentanglement Experiment, ablations, and training efficiency.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports extensive quantitative comparisons and ablation studies, but does not currently include error bars, confidence intervals, multi-seed variance, or formal statistical significance tests. The reported results should therefore be interpreted as single-setting empirical comparisons.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the main compute setup, including training for 250 epochs on 4 NVIDIA A6000 GPUs with batch size 96, as well as the approximate training time for AHA and the symmetric baseline. The Talking Face Disentanglement Experiment also reports training on 4 A6000 GPUs for approximately ten hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and are not aware of any aspect of the research that violates it. The research uses existing datasets and standard experimental protocols.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impacts in the Broader Impacts section, including potential benefits for robust audio–visual representation learning and reduced reliance on labeled data, as well as risks related to synthetic media manipulation in talking-face applications. It also recommends responsible use practices such as consent-based data use, disclosure of synthetic outputs, and safety evaluation before deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new high-risk dataset, pretrained generative model, face-generation system, or deployable model checkpoint. Therefore, additional safeguards for responsible release are not applicable in the current submission.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the original creators of the datasets, pretrained models, and reference implementations used in the experiments. The Existing Assets and Licenses section lists the main assets, their citations, official sources, and available licenses or terms of use, including cases where an explicit public dataset license could not be found.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, codebase, model checkpoint, or other reusable asset. The contribution is a new method and empirical analysis built on existing datasets and models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or newly conducted studies with human participants. All experiments are conducted using existing datasets and pretrained components.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new crowdsourcing or human-subject experiments and does not collect new data from study participants. Therefore, IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, experiments, and scientific contributions of this work do not involve LLMs as an important, original, or non-standard component. Any use of LLMs, if applicable, is limited to writing, editing, or formatting assistance and does not affect the methodology, experimental design, or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.