Enhancing Multimodal Unified Representations for Cross Modal Generalization

Hai Huang[1†], Yan Xia[1†], Shengpeng Ji[1], Shulei Wang[1], Hanting Wang[1], Minghui Fang[1], Jieming Zhu[2*], Zhenhua Dong[2], Sashuai Zhou[1], Zhou Zhao[1‡]

[1]Zhejiang University, [2]Huawei Noah's Ark Lab
haihuangcode@outlook.com

Abstract

To enhance the interpretability of multimodal unified representations, many studies have focused on discrete unified representations. These efforts typically start with contrastive learning and gradually extend to the disentanglement of modal information, achieving solid multimodal discrete unified representations. However, existing research often overlooks two critical issues: **1)** The use of Euclidean distance for quantization in discrete representations often overlooks the important distinctions among different dimensions of features, resulting in redundant representations after quantization; **2)** Different modalities have unique characteristics, and a uniform alignment approach does not fully exploit these traits. To address these issues, we propose Training-free Optimization of Codebook (TOC) and Fine and Coarse cross-modal Information Disentangling (FCID). These methods refine the unified discrete representations from pretraining and perform fine- and coarse-grained information disentanglement tailored to the specific characteristics of each modality, achieving significant performance improvements over previous state-of-the-art models. The code is available at https://github.com/haihuangcode/CMG.

1 Introduction

Humans' capacity to integrate multimodal information, such as text, audio, and visual, has inspired research on extracting unified information from multimodal data (Harwath et al., 2018; Miech et al., 2019; Shvetsova et al., 2022; Monfort et al., 2021). Researchers aim to develop models that learn unified representations across modalities, using techniques like contrastive learning to map semantically similar multimodal data closer in the embedding space (Radford et al., 2021; Luo et al., 2022; Xu et al., 2021), achieving notable results in downstream tasks like zero-shot cross-modal retrieval. However, the unbounded nature of the continuous embedding space poses challenges in interpretability. To address this, recent works have explored constructing discrete embedding spaces with prototypes or codebooks, enhancing cross-modal learning and model interpretability (Liu et al., 2021a; Lu et al., 2022; Zhao et al., 2022; Xia et al., 2024).

While recent works has demonstrated incredible achievements in multimodal unified representation, there are limitations in terms of the efficiency of embedding space utilization and the granularity of alignment. **1)** According to previous work (Breiman, 2001; Wojtas and Chen, 2020), the significance of features varies across different dimensions, and selecting the appropriate dimensions can optimize the feature space, thereby speeding up inference and improving model performance. However, existing multimodal unified representation methods, whether through contrastive learning (Liu et al., 2021a), teacher-student distillation (Duan et al., 2022), or information disentanglement (Xia et al., 2024; Huang et al., 2025), overlook this issue due to the inherent constraints of the codebook and the quantization method based on Euclidean distance. **2)** In the unified discrete representation of multimodal data, some studies focus on coarse-grained semantic alignment (Duan et al., 2022), others on fine-grained alignment (Xia et al., 2024), and yet others consider both fine and coarse alignments simultaneously (Liu et al., 2021a). However, these approaches align text with audiovisual data in the same granularity, overlooking the inherent differences between modalities: audiovisual data have temporal fine-grained connections, whereas text represents holistic semantics.

To address the aforementioned issues, we propose two techniques: **TOC** and **FCID**.

1) Training-free Optimization of Codebook (TOC), inspired by training-free adapters (Zhang

[†]Equal Contribution
[*]Project Lead
[‡]Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 2353–2366

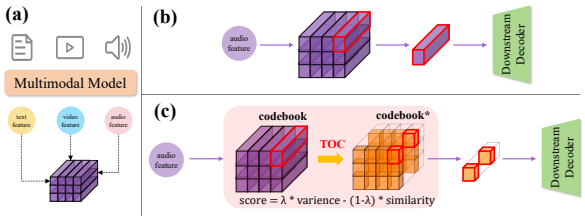

Figure 1: (a) Pretrained multimodal unified discrete representation. (b) vanilla downstream experiments using the quantized code from the unified representation. (c) After refinement with TOC, downstream experiments are conducted using only a subset of the dimensions.

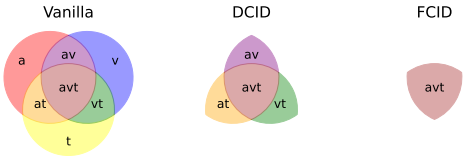

Figure 2: Multimodal Unified Representation of Vanilla method, DCID, FCID.

et al., 2022; Zhu et al., 2023) and drawing on the concept of feature importance (Breiman, 2001; Wojtas and Chen, 2020; Xue et al., 2022; Zhu et al., 2023), as illustrated in Figure 1, enables the use of a refined code for downstream tasks compared to the conventional approach, which directly uses the full quantized code during inference. The TOC calculation is independent of downstream tasks and does not require additional training. It only needs to be computed once, and the resulting dimensions can be reused for all subsequent inferences.

2) We adjust the order of multimodal alignment based on the inherent differences between text and audio-visual modalities and propose a **Fine and Coarse Cross-modal Information Disentangling (FCID)** architecture. As shown in Figure 2, vanilla methods for learning multimodal unified representations often retain modality-specific information. DCID (Xia et al., 2024) introduces disentanglement to separate modality-specific features; however, it does so by repeatedly aligning and decoupling pairs of modalities. For example, when aligning and disentangling audio and text, residual information such as *at* and *avt* is left behind. Similarly, when aligning audio and video, residuals like *av* and *avt* remain. These residuals, *at* and *av*, suggest that the disentanglement process is not fully effective. In contrast, FCID first aligns and disentangles audio and video while retaining fine-grained *av* and *avt* temporal information. It then performs coarse-grained disentangling with text, preserving the shared *avt* information, resulting in a more unified multimodal representation. The main contributions are summarized as follows:

- We propose **TOC**, a novel method for accurately identifying the importance of feature dimensions through efficient calculations, without the need for additional training. This versatile approach can be seamlessly applied to both multimodal unified and single-modal codebooks, offering a promising and adaptable solution.
- We introduce **FCID**, which disentangles information based on the distinct characteristics of text and audiovisual modalities, preserving both temporal and semantic information across all three modalities, and achieved a more unified multimodal representation.
- Our method outperforms the state-of-the-art across various cross-modal generalization tasks. Specifically, FCID, TOC, and their combination improve upon the SOTA by 2.16%, 1.06%, and 2.96%, respectively, across four downstream tasks. Additionally, we validated our approach on zero-shot cross-modal retrieval and cross-modal generation tasks.

2 Related Work

Multi-Modal Unified Representation: Recent work on multi-modal unified representations includes approaches that align modalities into a shared latent space (Petridis et al., 2018; Sarkar and Etemad, 2022; Andonian et al., 2022) and train modal-general encoders for cross-modal extraction (Chen et al., 2020b; Wang et al., 2022). Cross-modal distillation enables knowledge transfer between modalities (Sarkar and Etemad, 2022; Pedersoli et al., 2022), while bridging techniques connect representation spaces for improved unified representations (Wang et al., 2023). To enhance interpretability, many works use codebooks or prototypes (Duan et al., 2022; Lu et al., 2022; Liu et al., 2021a; Zhao et al., 2022; Xia et al., 2024; Huang et al., 2024; Fang et al., 2024). For instance, Duan et al. (2022) applies Optimal Transport to map features to prototypes, Xia et al. (2024) maps multimodal sequences to a common discrete semantic space. Our FCID framework addresses the inherent differences between text and audio-visual modalities through decoupling, enhancing multimodal unified representations.

Training Free Optimization: Recent works have explored various training-free methods to boost model performance. Tip-Adapter(Zhang et al., 2022) and APE(Zhu et al., 2023) enhance CLIP's few-shot classification, while a training-free method for diffusion models (Chen et al., 2024a) optimizes time steps and architecture for efficient image generation. The FuseDream (Liu et al., 2021b) combines CLIP and GANs for robust text-to-image generation. TEEN (Wang et al., 2024) offers a training-free solution for few-shot class-incremental learning. SCG-Diffusion (Wang et al., 2025) proposes a novel training-free method to improve alignment in Transformer-based Text-Guided Diffusion Models. Recent research in video generation (Chen et al., 2024b; Yang et al., 2024; Peng et al., 2024; Zhang et al., 2023) and multimodal large language models (Wu et al., 2024) also focuses on training-free techniques. We introduce TOC, the first training-free optimization method for discrete representation, broadening the scope of training-free approaches in the field.

3 Backgroud

Cross Modal Generalization (CMG) is a task introduced by Xia et al. (2024) that evaluates a model's ability to map diverse modalities into a unified discrete latent space. The model's ability for cross-modal zero-shot knowledge transfer is assessed in a setup where training is conducted on modality m_1 and testing is performed on modality m_2.

During training, the model learns a representation for inputs from one modality using the encoder Φ^{m_1} and the downstream decoder $\mathbf{D}$:

$$\mathbf{E}(\mathbf{D}(VQ(\Phi^{m_1}(\mathbf{x}_i^{m_1}))), \mathbf{y}_i^{m_1}), \tag{1}$$

where $\mathbf{x}_i^{m_1}$ is the input, $\mathbf{y}_i^{m_1}$ is the label, and $\mathbf{E}$ is the evaluation function. During testing, the model is evaluated on a different modality m_2, demonstrating its ability to generalize:

$$\mathbf{E}(\mathbf{D}(VQ(\Phi^{m_2}(\mathbf{x}_i^{m_2}))), \mathbf{y}_i^{m_2}). \tag{2}$$

Here, $m_1, m_2 \in a, b, c$ and $m_1 \neq m_2$. The parameters of both Φ^{m_1} and Φ^{m_2} are parameters frozen during training and testing, while only the parameters of $\mathbf{D}$ are updated during training.

4 Method

4.1 Training-free Optimization Codebook

Discrete unified representation spaces commonly employ a codebook structure, where modalities are updated based on the Euclidean distance between their features and the codebook codes. This dimension-equal-weighted update strategy does not consider the varying importance of feature dimensions, leading to redundancy in the final discrete space. According to previous work (Breiman, 2001; Wojtas and Chen, 2020), the importance of features varies across different dimensions. Eliminating redundant dimensions can help improve performance and accelerate computation. Therefore, we propose two metrics, Code Similarity and Code Variance, to refine features in the unified space.

Code Similarity: This metric aims to enhance the distinctiveness of codes by extracting feature dimensions that minimize code similarity. We represent the unified representation codebook of modalities as $\mathbf{e} \in \mathbb{R}^{H \times D}$, where H, D denote the size of the codebook and hidden dimension, respectively.

Assuming the existence of a classification dataset with C categories, acquiring its complete data enables the calculation of the average similarity, denoted as S. In an open-world setting, we may assume that the prior probabilities of all categories are equal, denoted as $\frac{1}{C}$. We adopt cosine similarity, $\delta(\cdot, \cdot)$, as the chosen metric:

$$S = \frac{1}{C^2} \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} \frac{1}{N^i N^j} \sum_{m=1}^{N^i} \sum_{n=1}^{N^j} \delta(\mathbf{x}^{i,m}, \mathbf{x}^{j,n}), \tag{3}$$

where $\mathbf{x}^{i,m}$ and $\mathbf{x}^{j,n}$ denote the input features for the m-th and n-th samples of categories i and j, respectively. N^i and N^j represent their respective total number of training samples.

Each code in the pretrained codebook, $\mathbf{e}^i \in \mathbb{R}^D, i \in [0, H)$, can be considered as a distinct semantic cluster center, representing a category. Therefore, we can simplify the average similarity calculation:

$$S = \frac{1}{H^2} \sum_{i=1}^{H} \sum_{\substack{j=1 \\ j \neq i}}^{H} \delta(\mathbf{e}^i, \mathbf{e}^j), \tag{4}$$

Our goal is to select Q dimensions out of D to enhance the distinctiveness of the codes. We introduce a binary flag $\mathbf{F} \in \{0, 1\}^D$, where $F_k = 1$ $(k = 1, ..., D)$ indicates that the k^{th} dimension $\mathbf{e}_k^i$ is selected, and $\mathbf{F}\mathbf{F}^\top = Q$. Our objective now

becomes finding the optimal $\mathbf{F}$ to minimize the Code Similarity:

$$\min_{\mathbf{F}} \quad S = \frac{1}{H^2} \sum_{i=1}^{H} \sum_{\substack{j=1 \\ j \neq i}}^{H} \delta(\mathbf{e}^i \odot \mathbf{F}, \mathbf{e}^j \odot \mathbf{F}), \quad (5)$$

where $\odot$ denotes element-wise multiplication.

We further suppose the Codebook has been L2-normalized, meaning that each code vector $\mathbf{e}^i \in \mathbb{R}^D$ has a unit length. Under this assumption, the cosine similarity between two code vectors $\mathbf{e}^i$ and $\mathbf{e}^j$ can be simplified as their dot product:

$$\delta(\mathbf{e}^i, \mathbf{e}^j) = \mathbf{e}^i \cdot \mathbf{e}^j, \quad (6)$$

where $\cdot$ denotes the dot product of two vectors. Then we can simplify the cosine similarity as

$$S = \sum_{k=d_1}^{d_Q} S_k = \sum_{k=d_1}^{d_Q} \left(\frac{1}{H^2} \sum_{i=1}^{H} \sum_{\substack{j=1 \\ j \neq i}}^{H} \mathbf{e}_k^i \cdot \mathbf{e}_k^j \right), \quad (7)$$

where $k = \{d_1, d_2, ..., d_Q\}$ denotes the indices of selected feature dimensions with $F_k = 1$, and $S_k = \frac{1}{L^2} \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \mathbf{e}_k^i \cdot \mathbf{e}_k^j$ represents the average inter-class similarity of the k^{th} dimension. Through straightforward derivation, we observe that solving the optimization problem is equivalent to selecting Q elements with the smallest average similarity.

Code Variance: Our goal is to reduce redundancy by removing feature dimensions with low variance across codewords, as these dimensions offer minimal discriminative value. The variance for the k^{th} feature dimension is formulated as:

$$V_k = \frac{1}{L} \sum_{i=1}^{L} (\mathbf{e}_k^i - \bar{\mathbf{e}}_k)^2, \quad (8)$$

where $\bar{\mathbf{e}}_k = \frac{1}{L} \sum_{i=1}^{L} \mathbf{e}_k^i$ represents the mean of the k^{th} dimension across all codewords. Similar to Code Similarity, we select the top Q dimensions with the highest variance to enhance discriminative power.

To combine the criteria of similarity and variance, a balance factor λ is introduced to compute the final metric for each feature dimension:

$$U_k = \lambda V_k - (1 - \lambda) S_k, \quad (9)$$

where $k = 1, \ldots, D$. The dimensions corresponding to the top-Q biggest values of U_k are chosen as the refined features.

4.2 Fine and Coarse cross-modal Information Disentangling

As shown in Figure 3, the Vanilla method uses contrastive learning (Liu et al., 2021a) or distillation (Duan et al., 2022) for cross-modal representation unification. DCID (Xia et al., 2024) introduces decoupling to separate modality-specific information. In FAC (Alayrac et al., 2020), audio and video are first aligned temporally and then with text for semantic alignment. We propose FCID, which combines FAC and DCID. It decouples and aligns audio and video, isolates modality-specific data, and then further aligns with text, discarding non-shared information for a unified multimodal representation. FCID is the first method to simultaneously tackle both modality differences and redundancy issues, inspired by FAC and DCID.

4.2.1 Fine cross-modal Information Disentangling

Given paired audio-video modalities, $(\mathbf{x}_i^a, \mathbf{x}_i^v)_{i=1}^N$, we utilize two fine modal-general encoders, Φ_f^a and Φ_f^v, to extract fine modal-general features $\mathbf{f}_i^a$ and $\mathbf{f}_i^v \in \mathbb{R}^{T \times D}$, and employ two fine modal-specific encoders, Ψ_f^a and Ψ_f^v, to obtain fine modal-specific features $\bar{\mathbf{f}}_i^a$ and $\bar{\mathbf{f}}_i^v \in \mathbb{R}^{T \times D}$ from the audio and video modalities, respectively. Here, N, T, and D represent the number of samples, the length of audio-video sequences, and the feature dimension, respectively. In subsequent equations, $m, n \in \{audio, video\}$:

$$\mathbf{f}_i^m = \Phi^m(\mathbf{x}_i^m), \ \bar{\mathbf{f}}_i^m = \Psi^m(\mathbf{x}_i^m). \quad (10)$$

Subsequently, we minimize the mutual information between the fine modal-specific features $\mathbf{f}_i^m$ and $\bar{\mathbf{f}}_i^m$. At the same time, maximize the mutual information between $\mathbf{f}_i^m$ and $\mathbf{f}_i^n$. The details of this approach are outlined below:

Mutual Information Minimization: CLUB (Cheng et al., 2020) could optimize the mutual information upper bound, demonstrating superior advantages in information disentanglement. Given two variables $\mathbf{x}$ and $\mathbf{y}$, the objective function of CLUB is defined as:

$$\begin{aligned} I_{vCLUB}(\mathbf{x}; \mathbf{y}) := \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\log q_\theta(\mathbf{y}|\mathbf{x})] \\ - \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{y})}[\log q_\theta(\mathbf{y}|\mathbf{x})]. \end{aligned} \quad (11)$$

We use CLUB to optimize the mutual information upper bound between fine modal-general features $\mathbf{f}_i^m$ and fine modal-specific features $\bar{\mathbf{f}}_i^m$,

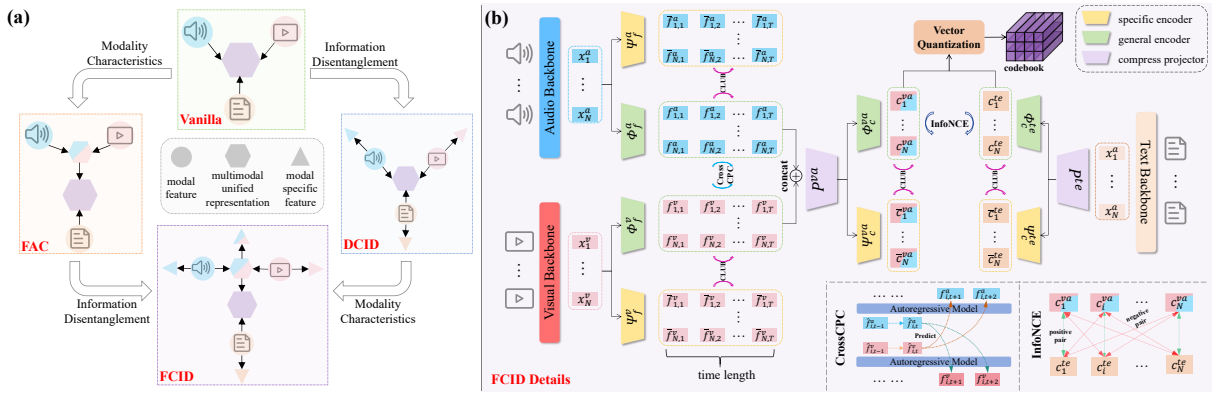

Figure 3: (a) A simple demonstration of four models: Vanilla method, FAC, DCID, and our proposed FCID. (b) Details of the FCID encoder. On the left, audio and video undergo fine-grained mutual information separation and alignment using modal-general encoders Φ_f^a, Φ_f^v and modal-specific encoders Ψ_f^a, Ψ_f^v. The CLUB module separates specific information $\bar{\mathbf{f}}_i^a, \bar{\mathbf{f}}_i^v$ from general information $\mathbf{f}_i^a, \mathbf{f}_i^v$, while CrossCPC aligns the general information across modalities. This is followed by compressing the features into unified audiovisual representations. On the right, coarse-grained mutual information separation and alignment are conducted with audiovisual data and text, resulting in a unified discrete representation across all three modalities. The decoder combines the quantized code with modality-specific features and computes the reconstruction loss, omitted in the figure.

where q_θ is the variational approximation of ground-truth posterior of $\mathbf{y}$ given $\mathbf{x}$ and can be parameterized by a network θ.

$$\begin{aligned}\hat{I}_{vCLUB_f} = &\frac{1}{N}\sum_{i=1}^{N}[\frac{1}{T}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{f}}_i^m|\mathbf{f}_i^m) \\ &-\frac{1}{N}\frac{1}{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{f}}_j^m|\mathbf{f}_i^m)].\end{aligned} \tag{12}$$

Mutual Information Maximization: Contrastive Predictive Coding (CPC) (Oord et al., 2018) aims to maximize the mutual information between sequence items by predicting future samples using autoregressive models and is widely adopted in self-supervised learning. Given fine general features $\mathbf{f}^a, \mathbf{f}^v \in \mathbb{R}^{T\times D}$, a prediction horizon of R steps, and a random time moment $t \in (0, \text{T-R}]$, two single-layer unidirectional LSTMs are used to summarize the information of all $\mathbf{f}_{\leq t}^a, \mathbf{f}_{\leq t}^v$, yielding three context representations as $\mathbf{o}_t^m = \text{LSTM}(\mathbf{f}_{\leq t}^m)$.

For modality M, we first select a set Z_{neg} of N-1 random negative samples and one positive sample $\mathbf{f}_{t+r}^n$ from modality N, then use $\mathbf{o}_t^m$ to predict r-th future step $\mathbf{f}_{t+r}^n$ in modality N, and the loss for all modality can be optimized as:

$$L_{cpc}^{m2n} = -\frac{1}{R}\sum_{r=1}^{R}\log\left[\frac{\exp\left(\mathbf{f}_{t+r}^n W_r^m \mathbf{o}_t^m\right)}{\sum_{\mathbf{f}_j\in Z_{neg}}\exp\left(\mathbf{f}_j^n W_r^m \mathbf{o}_t^m\right)}\right]. \tag{13}$$

4.2.2 Coarse Cross-modal Information Disentangling

CCID initially sets up two projectors, P^{te} for compressing textual features and P^{va} for compressing the audiovisual modal-general features obtained from FCID. Subsequently, it configures two coarse modal-specific encoders, Ψ_c^{av} and Ψ_c^{te}, to extract coarse modal-specific features $\bar{\mathbf{c}}_i^{av}$ and $\bar{\mathbf{c}}_i^{te} \in \mathbb{R}^D$, and two coarse modal-general encoders, Φ_c^{av} and Φ_c^{te}, are employed to derive coarse modal-general features $\mathbf{c}_i^{av}$ and $\mathbf{c}_i^{te} \in \mathbb{R}^D$ from the audiovisual and textual modalities, respectively. In subsequent equations, $\mathcal{M}, \mathcal{N}, \in \{audiovisual, text\}$:

$$\begin{aligned}\mathbf{c}_i^{\mathcal{M}} &= \Phi_c^{\mathcal{M}}(P^{\mathcal{M}}(\mathbf{x}_i^{\mathcal{M}})), \\ \bar{\mathbf{c}}_i^{\mathcal{M}} &= \Psi_c^{\mathcal{M}}(P^{\mathcal{M}}(\mathbf{x}_i^{\mathcal{M}})),\end{aligned} \tag{14}$$

The subsequent process of information disentanglement and alignment is similar to that of Fine cross-modal Information Disentangling.

Mutual Information Minimization: We use CLUB to optimize the mutual information upper bound between coarse modal-general features $\mathbf{c}_i^{\mathcal{M}}$ and fine modal-specific features $\bar{\mathbf{c}}_i^{\mathcal{M}}$, similar to $\hat{I}_{vCLUB_f}$ in FCID:

$$\begin{aligned}\hat{I}_{vCLUB_c} = &\frac{1}{N}\sum_{i=1}^{N}[\frac{1}{T}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{c}}_i^{\mathcal{M}}|\mathbf{c}_i^{\mathcal{M}}) \\ &-\frac{1}{N}\frac{1}{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{c}}_j^{\mathcal{M}}|\mathbf{c}_i^{\mathcal{M}})].\end{aligned} \tag{15}$$

Mutual Information Maximization: Since the coarse information lacks a sequential structure, we transitioned the contrastive learning approach from CPC to InfoNCE, as described below:

$$L_{nce} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp(\text{sim}(c_i^{\mathcal{M}}, c_i^{\mathcal{N}})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(c_i^{\mathcal{M}}, c_j^{\mathcal{N}})/\tau)}\right]. \tag{16}$$

4.3 Final Loss

Then, we use the codebook to explicitly represent the unified multimodal representation, the latent codebook $\mathbf{e} \in R^{H\times D}$ is shared across modalities audio, video, and text, where T, H, D represent time, size of the discrete latent space, and hidden dimension, respectively. Apply vector quantized operation to map coarse model-general feature $\mathbf{f}_i^{av}, \mathbf{f}_i^{te}$ to discrete latent codes, $t \in [0, T)$:

$$\begin{aligned}&\hat{\mathbf{c}}_{i,t}^{\mathcal{M}} = VQ(\Phi_c^{\mathcal{M}}(\mathbf{x}_i^{\mathcal{M}})) = VQ(\mathbf{c}_{i,t}^{\mathcal{M}}) = e_l,\\ &\text{where } l = argmin_j||\Phi_c(x) - e_j||_2.\end{aligned} \tag{17}$$

Then, we combine $\hat{\mathbf{c}}_i^m$ with $\bar{\mathbf{c}}_i^m$ together to reconstruct original features:

$$\underbrace{\|\mathbf{x}_i^{\mathcal{M}} - D(\hat{\mathbf{c}}_i^{\mathcal{M}}; \bar{\mathbf{c}}_i^{\mathcal{M}})\|_2^2}_{\text{reconstruction loss}} + \underbrace{\beta\|\phi_k^{\mathcal{M}}(\mathbf{x}_i^{\mathcal{M}}) - \text{sg}[\mathbf{e}]\|_2^2}_{\text{commitment loss}}, \tag{18}$$

and employ Multimodal Exponential Moving Average (MMEMA) strategy to update codebook.

The overall objective of FCID is a combination of these loss functions across both layers:

$$L = L_{\text{recon}} + L_{\text{commit}} + L_{\text{contra}} + L_{\text{club}}, \tag{19}$$

where L_{recon} is the reconstruction loss that merges the modal-specific and modal-general results for each modality and compares them with the original input using MSE loss, L_{commit} is the commitment loss that computes the MSE loss between the modal-general results and their quantized codes, $L_{\text{contra}} = L_{\text{cpc}} + L_{\text{nce}}$ is the loss that enhances cross-modal alignment and inference by predicting future samples in one modality using information from another, and $L_{\text{club}} = \hat{I}_{vCLUB_f} + \hat{I}_{vCLUB_c}$ represents the mutual information loss concerning the modal-specific and modal-general results within each modality.

5 Experiment

5.1 Datasets and Tasks

5.1.1 Pretrain

Multimodal Unified Representation: The pretraining dataset uses VGGsound-AVEL40K (Chen et al., 2020a; Zhou et al., 2022) with text from Xia et al. (2024). **Single-modal Representation:** We trained a VQVAE (Van Den Oord et al., 2017) on the CelebA-HQ 30K (Karras et al., 2017) dataset and evaluated TOC's effect on selecting feature dimensions for reconstruction, assessing its transferability to other domains using the codebook.

5.1.2 Downstream

The unified representation pre-trained models will be evaluated on several downstream tasks using different datasets. **Cross-modal event classification on AVE dataset:** (Tian et al., 2018) training on one modality and evaluating on another. **Cross-modal event localization on AVVP dataset:** (Tian et al., 2020) localizing events in one modality and transferring to the other. **Cross-dataset localization/classification:** training on classification in AVE and evaluating localization in AVVP, transferring across datasets. Cross-modal classification between UCF-101 (Soomro et al., 2012) visual clips and VGGSound (Chen et al., 2020a) audio clips. The decoder in all of the above experiments consists of a single linear layer. **Cross-modal Zero-shot Retrieval:** We adopt a process similar to the test set (Yu et al., 2018), which consists of 500 pairs from MSCOCO (Chen and Dolan, 2011), assessing zero-shot retrieval capability for visual-text alignment. Clotho (Drossos et al., 2020) assesses zero-shot retrieval capability for audio-text alignment. Flickr Sound (Senocak et al., 2018) assesses zero-shot retrieval capability for audio-visual alignment. **Cross-modal Generation:** We use a 2-layer MLP and the IP-Adapter (Ye et al., 2023) as downstream decoders. The MLP is trainable during training but frozen during testing, while the IP-Adapter remains frozen throughout. By leveraging IP-Adapter's image-to-image capability, the MLP bridges multimodal unified representation and image generation, enabling audio-to-image and text-to-image generation during testing. The model was fine-tuned on 4,500 FlickrSound (Senocak et al., 2018) image-audio pairs over 80,000 steps with a batch size of 8, and evaluated on 500 additional pairs. Please refer to Appendix A for details on the compared works, evaluation metrics, and hyperparameter settings.

Figure 4: Example results of reconstructions using random and TOC masking.

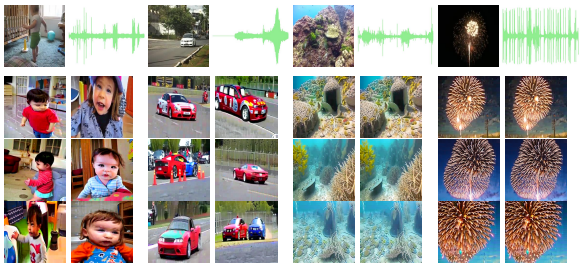

Figure 5: Example results of cross-modal image generation experiments conducted by FCID+TOC.

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	
CODIS	36.8	39.7	32.7	32.6	40.8	40.6	50.8	45.2	39.90
TURN	37.6	39.2	32.4	32.2	40.6	41.4	50.4	46.1	39.99
CMCM	46.3	45.8	36.1	35.2	47.1	48.2	51.2	48.3	44.78
SimMMDG	49.5	51.7	39.3	39.7	52.9	52.7	64.5	58.8	51.14
DCID	54.1	**55.0**	40.4	40.8	53.0	52.4	67.1	60.6	52.93
FCID	**55.2**	54.9	**42.4**	**44.5**	**55.3**	**57.4**	**69.4**	**61.6**	**55.09**
CODIS+TOC	37.2	41.3	33.1	33.9	41.9	42.4	51.2	47.3	41.04(+1.14)
TURN+TOC	38.3	40.5	33.2	32.9	41.5	43.3	51.5	46.8	41.00(+1.01)
CMCM+TOC	46.9	47.2	37.9	36.2	49.8	50.1	52.3	49.1	46.19(+1.41)
DCID+TOC	54.5	**55.0**	40.9	41.6	56.5	53.6	68.1	61.7	53.99(+1.06)
FCID+TOC	**55.9**	**55.0**	**43.6**	**45.1**	**57.4**	**58.5**	**69.6**	**62.0**	**55.89**(+0.80)

Table 1: Comparison with SOTA Methods on four audio-visual tasks. (SimMMDG represents recent great work in multimodal domain generalization, is incompatible with TOC as it does not use discrete representations.)

5.2 Performance Analysis

In the all tables, **bold** numbers indicate the best results, while green values in parentheses show the performance improvement attributed to the TOC.

TOC: As shown in Table 1 and Table 2, TOC optimizes methods with discrete representation spaces, facilitating at least a 0.80% improvement in average results for cross-modal generalization tasks, and a minimum average increase of 0.41% for cross-modal zero-shot retrieval tasks. Additionally, as illustrated in Table 4 and Figure 5, TOC excels in cross-modal generation tasks, improving image-to-image ($I2I$), audio-to-image ($A2I$), and text-to-image ($T2I$) generation performance.

We also explored extending TOC to unimodal discrete representation space. Using a VQVAE model trained on the CelebA-HQ 30K dataset (Karras et al., 2017), we tested reconstructions with a subset of the codeword dimensions. Table 3 shows that R100-avg represents the average outcome of 100 random codeword dimension selections for reconstruction, where TOC masks the least important dimensions. The MSE for TOC reconstructions with 25% to 87.5% of dimensions was significantly lower than the average MSE from random selections. 'Count' represents the number of times the MSE from 100 random selections is greater than the MSE from TOC. Figure 4 displays reconstruction samples, for all columns except the 'origin' column, with the left half showing reconstructions with randomly masked dimensions and the right half showing TOC-reconstructed images, demonstrating the superior effectiveness of TOC-selected dimensions. For more results, see Appendix B.5.

We also conducted experiments on the impact of TOC on the cosine similarity of the codebook. Please refer to Appendix B.1 for details.

FCID: Reviewing Tables 1 and 2 clearly shows that FCID and FCID+TOC consistently outperform all other methods across a variety of tasks. Compared to the previous SOTA, FCID achieves an average improvement of 2.16% in four cross-modal generalization tasks and an average improvement of 0.75% in three cross-modal zero-shot retrieval tasks. As shown in Table 4, these approaches also demonstrate a clear advantage in cross-modal generation tasks. All results suggest that our methods can more effectively process and understand cross-modal information.

As shown in Figure 5, the top row displays four image-audio pairs, and the three rows below show images generated from these samples. FCID, fine-tuned only with images, achieves A $\rightarrow$ I results that closely resemble I $\rightarrow$ I outcomes. Notably, in the last two examples, the generated images are identical, indicating that these image-audio pairs map to the same code in the codebook, demonstrating strong modal alignment. For additional T $\rightarrow$ I examples and results, see Appendix B.6.

We demonstrate multimodal quantization activations in the discrete representation spaces of DCID (Xia et al., 2024) and FCID using tri-modal audio-video-text data from VALOR32K (Chen et al., 2023). Figures 6 and 7 show the code activations in the DCID and FCID codebooks. Red points indicate single modality activations > 95%, green points show activation across all three modalities $\geq$ 5%, and blue points fall between these categories. More green dots indicate closer alignment of discrete representations across modalities in the

Method	V↔T (R@10)	A↔T (R@10)	V↔A (R@10)	Avg.
CMCM	7.20	14.87	15.60	7.11
DCID	8.30	16.70	17.20	8.14
FCID	**9.60**	**18.19**	**17.50**	**8.89**
CMCM+TOC	7.70	15.33	16.10	7.52(+0.41)
DCID+TOC	8.80	17.08	17.80	8.56(+0.42)
FCID+TOC	**10.40**	**19.04**	**18.40**	**9.42**(+0.53)

Table 2: Results on three cross-modal zero-shot retrieval tasks, averaged across two directions, with Avg. as the average of R@1, R@5, and R@10 (details in Table 10).

Mask (%)	R100-avg↓	TOC↓	Count↑
87.5	0.0621	0.0231	100
75.0	0.0477	0.0159	100
62.5	0.0335	0.0109	100
50.0	0.0229	0.0086	100
37.5	0.0141	0.0062	96
25.0	0.0075	0.0039	90

Table 3: Reconstruction errors under random vs. TOC masking.

Method	I2I↓	A2I↓	T2I↓
CMCM	129.56	130.93	148.93
DCID	121.44	123.28	141.16
FCID	**116.06**	**117.26**	**135.52**
CMCM+TOC	124.25	125.37	144.93
DCID+TOC	118.30	119.96	135.93
FCID+TOC	**113.95**	**115.14**	**130.98**

Table 4: Cross-modal image generation results (lower is better).

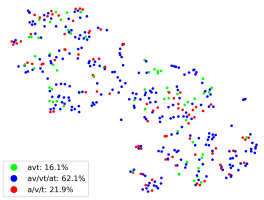

Figure 6: DCID's codebook activate

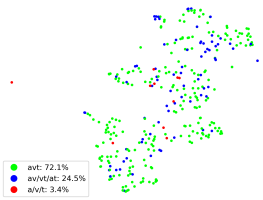

Figure 7: FCID's codebook activate

codebook, while more red dots reflect greater divergence. FCID outperforms DCID in aligning the three modalities, with fewer codes activated by a single modality, demonstrating its improved ability to learn unified tri-modal representations.

For details on the computational efficiency and loss variation of FCID pre-training, please refer to Appendix B.2 and B.3, respectively.

5.3 Ablation Study

Loss: We conducted ablation experiments on all losses in Equation 19, excluding the contrastive loss, as it is fundamental to multimodal alignment in this work. Without contrastive learning, alignment cannot be achieved, rendering its ablation unnecessary. As shown in Table 5, we observe that L_{commit} has a minor impact on results, as its primary role is to align unified discrete representations with quantized features. In contrast, L_{recon} is crucial for ensuring semantic completeness after disentanglement, minimizing information incompleteness during this process. Finally, L_{club} directly evaluates the success of information disentangling, and its absence significantly degrades model performance.

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	**55.9**	**55.0**	**43.6**	**45.1**	**57.4**	**58.5**	**69.6**	62.0	**55.89**
w/o L_{club}	51.3	51.6	39.5	40.7	50.6	51.1	63.3	57.6	50.71
w/o L_{recon}	53.3	53.6	42.5	43.3	56.1	57.5	66.8	**62.3**	54.43
w/o L_{commit}	54.6	54.8	42.9	44.7	57.2	56.8	67.9	61.4	55.04

Table 5: Ablation studies on the impact of different loss.

TOC: As presented in Table 6, the two components of TOC, when applied individually, led to improvements in FCID performance by 0.54% and 0.10% across eight metrics, respectively. The combined effect of both components resulted in a total improvement of 0.80%. Refining the discrete representation space using either Code Similarity or Code Variance alone effectively reduces feature redundancy and enhances model performance in downstream tasks. However, Code Similarity demonstrates a greater impact than Code Variance. When both components are combined, the best performance is achieved, highlighting the effectiveness of our proposed TOC design.

S	R	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	
-	-	55.2	54.9	42.4	44.5	55.3	57.4	69.4	61.6	55.09
✓	-	55.8	54.5	**43.6**	**45.7**	56.8	58.3	69.2	61.1	55.63
-	✓	55.6	**55.0**	43.4	44.8	56.2	54.8	**69.8**	61.9	55.19
✓	✓	**55.9**	**55.0**	**43.6**	45.1	**57.4**	**58.5**	69.6	**62.0**	**55.89**

Table 6: Ablation studies on the impact of TOC (S and R represent Code Similarity and Code Variance, respectively)

FCID: We focus our ablation studies on the key disentanglement components, $\hat{I}vCLUB_f$ and $\hat{I}vCLUB_c$, which involve A_{CLUB}, V_{CLUB} and AV_{CLUB}, TE_{CLUB}, respectively. Table 7 shows that A_{CLUB} and V_{CLUB} significantly impact audiovisual-related downstream tasks. TE_{CLUB} also influences results, as improperly disentangled textual information can affect performance by containing irrelevant or missing AV data. Similarly, using only AV_{CLUB} still retains some modality-specific information, but the disentanglement and alignment with text help separate these features.

A_{CLUB}	V_{CLUB}	AV_{CLUB}	TE_{CLUB}	Avg.
-	-	-	-	50.71
✓	-	-	-	52.79
-	✓	-	-	53.40
-	-	✓	-	51.78
-	-	-	✓	51.59
✓	✓	-	-	54.34
-	-	✓	✓	52.46
✓	✓	✓	✓	**55.09**

Table 7: Ablation studies on the impact of FCID (detailed results for each task are provided in Table 11)

6 Conclusion

Inspired by feature importance and training-free optimization, we propose TOC, the first training-free optimization method for discrete representation space, enhancing both multimodal and single-modal representations. We also introduce FCID, a framework that integrates disentanglement with modality-specific characteristics to achieve fine-grained audio-video temporal alignment and coarse-grained text semantic alignment. This disentanglement separates modality-specific information, yielding a unified multimodal representation. Extensive experiments on cross-modal classification, localization, retrieval, and generation tasks validate the effectiveness of our approach.

Limitations

TOC: We assumed equal prior probabilities for different classes in the open-world setting, a reasonable assumption in many cases. However, in real-world applications, class distributions can be imbalanced, with some classes having an excessive number of instances, while others may have too few. Consequently, for broader practical applicability, TOC requires further optimization and deeper exploration to account for such variabilities.
FCID: is specifically designed for tri-modal representations (audio, video, and text). However, in scenarios involving only two modalities, only the Fine or Coarse components of FCID's disentanglement and alignment processes are applicable.

Acknowledgments

We thank MindSpore (http://mindspore.cn) for the partial support of this work, which is a new deep learning computing framework.

References

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in neural information processing systems*, 33:25–37.

Alex Andonian, Shixing Chen, and Raffay Hamid. 2022. 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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020a. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Minghao Chen, Iro Laina, and Andrea Vedaldi. 2024a. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353.

Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. 2024b. Delta-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. 2024. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2022. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660.

Minghui Fang, Shengpeng Ji, Jialong Zuo, Hai Huang, Yan Xia, Jieming Zhu, Xize Cheng, Xiaoda Yang, Wenrui Liu, Gang Wang, et al. 2024. Ace: A generative cross-modal retrieval framework with coarse-to-fine semantic modeling. *arXiv preprint arXiv:2406.17507*.

David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hai Huang, Shulei Wang, and Yan Xia. 2025. Semantic residual for multimodal unified discrete representation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Jieming Zhu, Zhenhua Dong, and Zhou Zhao. 2024. Unlocking the potential of multimodal unified discrete representation through training-free codebook optimization and hierarchical alignment. *arXiv preprint arXiv:2403.05168*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021a. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*.

Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021b. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, George Tzanetakis, and Kwang Moo Yi. 2022. Estimating visual information from audio through manifold learning. *arXiv preprint arXiv:2208.02337*.

Bo Peng, Xinyuan Chen, Yaohui Wang, Chaochao Lu, and Yu Qiao. 2024. Conditionvideo: Training-free condition-guided video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4459–4467.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE.

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

Pritam Sarkar and Ali Etemad. 2022. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. *arXiv preprint arXiv:2211.13929*.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263.

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

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Few-shot class-incremental learning via training-free prototype calibration. *Advances in Neural Information Processing Systems*, 36.

Shulei Wang, Wang Lin, Hai Huang, Hanting Wang, Sihang Cai, WenKang Han, Tao Jin, Jingyuan Chen, Jiacheng Sun, Jieming Zhu, et al. 2025. Towards transformer-based aligned generation with self-coherence guidance. *arXiv preprint arXiv:2503.17675*.

Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. 2022. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR.

Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. 2023. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114.

Maksymilian Wojtas and Ke Chen. 2020. Feature importance ranking for deep learning. *Advances in Neural Information Processing Systems*, 33:5105–5114.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. 2024. Controlmllm: Training-free visual prompt learning for multimodal large language models. *arXiv preprint arXiv:2407.21534*.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. 2024. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36.

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

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.

Shaoshu Yang, Yong Zhang, Xiaodong Cun, Ying Shan, and Ran He. 2024. Zerosmooth: Training-free diffuser adaptation for high frame rate video generation. *arXiv preprint arXiv:2406.00908*.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 471–487.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Yang Zhao, Chen Zhang, Haifeng Huang, Haoyuan Li, and Zhou Zhao. 2022. Towards effective multi-modal interchanges in zero-resource sounding object localization. *Advances in Neural Information Processing Systems*, 35:38089–38102.

Jinxing Zhou, Dan Guo, and Meng Wang. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. 2021. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. 2023. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. *arXiv preprint arXiv:2304.01195*.

A Implementation Details

Compared Works: The models we compare include the most outstanding recent developments in multimodal unified discrete representations and models that excel in multimodal domain generalization: CODIS (Duan et al., 2022), TURN (Zhao et al., 2022), CMCM (Liu et al., 2021a), SimMMDG (Dong et al., 2024), and DCID (Xia et al., 2024). These methods are implemented on our tasks, and their performance is evaluated on multi downstream tasks.

Evaluation Metrics: For the AVE (Tian et al., 2018), VGGSound-AVEL (Zhou et al., 2022, 2021), and UCF101 (Soomro et al., 2012) datasets, precision is used as the metric. The F1-score is utilized for assessing the AVVP (Tian et al., 2020) and AVE→AVVP generalization task, and recall is utilized for zero-shot retrieval (Chen and Dolan, 2011; Drossos et al., 2020). Mean Square Error (MSE) is employed to evaluate the reconstruction quality of

TOC on the CelebA-HQ 30K dataset (Karras et al., 2017). Additionally, Fréchet Inception Distance (FID) (Heusel et al., 2017) is used to assess the model's capability in cross-modal generalization. **Hyperparameter Settings:** The β of L_{commit} is set to 0.25, and in the TOC formulation, the parameter λ is set to 0.3, and in L_{nce}, the parameter τ is set to 1.0. All results presented in table 1, 2, 4, 6, 7 were obtained with a codebook size set to 400 and an embedding dimension set to 256. The table 3 involves VQVAE with a codebook size of 128 and an embedding dimension of 128. The ablation study on codebook size is discussed in Table 9.

B Experimental Supplement

B.1 The Impact of TOC on Cosine Similarity of Codebook

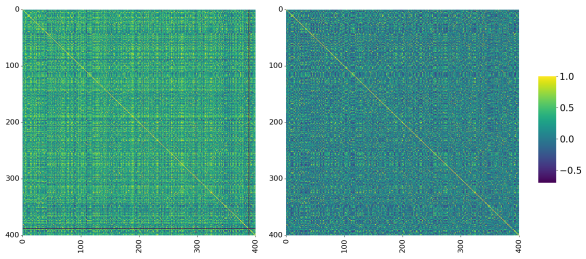

Figure 8: Left: Cosine similarity in the original codebook. Right: Cosine similarity after TOC.

We conducted an evaluation of TOC on the open-source pre-trained model of DCID (Xia et al., 2024). As depicted in Figure 8, the distinctiveness of the codes with the features obtained after TOC computation is notably enhanced.

B.2 Computational efficiency

Comparing computational efficiency and resource requirements provides a more comprehensive view of the trade-off between performance improvement and computational cost. As shown in Table 8, we present a comparison of the computational efficiency of CMCM (Liu et al., 2021a), DCID (Xia et al., 2024), and our method:

Model	GPU Memory Usage	Time per Epoch	Total Epochs
DCID	100%	100%	5
CMCM	81%	76%	8
FCID	96%	99%	5

Table 8: Model Comparison: Relative GPU Memory Usage and Time per Epoch (Compared to DCID), and Total Training Epochs

Under the same pretraining conditions with a batch size of 80 and using an RTX 3090 GPU, the results are shown in Table 8. In CMCM, contrastive learning is applied to both fine-grained and high-level representation for each modality. DCID performs alignment and disentanglement of all modalities at the fine-grained level (Alignment of pairwise combinations of the three modalities, along with their respective disentanglement). Our proposed FCID first applies disentangling and contrastive learning to fine-grained features for audio and video, then applies coarse-grained disentangling and contrastive learning to the compressed audio-visual and text features. Specifically, there are 2 alignment operations and 4 disentangling operations (fine-grained alignment of audio and video features, coarse-grained alignment of audiovisual and text features, and disentangling of these 4 feature types).

For GPU Memory Usage and Time per Epoch, CMCM has the lowest consumption in both cases. FCID uses less GPU memory than DCID, while Time per Epoch is nearly the same, mainly due to the projection compressing the temporal dimension before performing coarse-grained alignment and disentanglement, reducing computational data. DCID and FCID benefit from faster convergence due to disentangling and only require 5 epochs to achieve optimal performance, while CMCM needs warm-start strategy and takes 8 epochs to converge, resulting in the longest total training time.

During downstream inference, no disentangling or alignment is required. Only the corresponding encoders and quantization are used. Compared to DCID, our FCID adds only simple projection layers and coarse-grained general and specific encoders constructed with MLP for audiovisual data, making the added inference time almost negligible. The text modality, on the other hand, requires less time than DCID due to coarse-grained compression. The inference speed difference between CMCM, DCID, and FCID is minimal.

Another module we propose, TOC, only requires a single computation of less than 10 seconds after obtaining the discrete representations, and it only requires CPU computation. By reducing the feature dimensions, it can also accelerate training and inference speeds in downstream tasks.

B.3 Loss Changes During Pre-training

As shown in Equation 19, our model involves a total of four losses. Among them, L_{commit} is a

loss commonly used in VQ-related models. Its purpose is to ensure that the model's output maintains continuity in the latent space when mapped to the discrete space. This is not the main contribution of this paper. Here, we focus on the changes in the other three losses. As illustrated in the Figure 9, the overall trend shows a clear downward slope. Among them, L_{club} converges first, and when it reaches zero, it indicates successful decoupling of the modalities. The decrease in L_{contra} reflects an increase in the alignment of multimodal information, while the decrease in L_{recon} signifies improved preservation of complete semantics after the information is decoupled and re-merged.

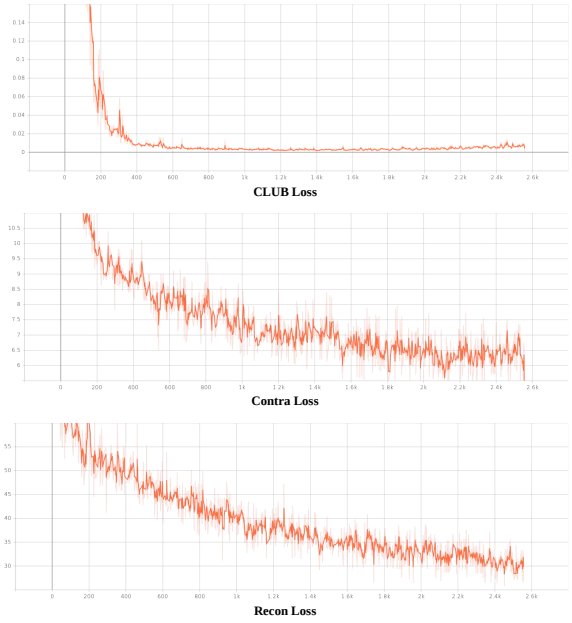

Figure 9: Changes in Losses During Pre-training

Codebook Size	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	
256	52.9	52.3	38.8	43.2	53.7	53.9	**70.8**	56.4	52.75
300	52.8	54.1	42.1	44.1	54.1	**58.5**	69.6	60.4	54.46
400	**55.2**	**54.9**	**42.4**	**44.5**	**55.3**	57.4	69.4	**61.6**	**55.09**
512	54.4	52.4	40.0	42.6	54.1	56.9	70.3	59.3	53.75
800	52.2	54.6	41.6	43.9	53.1	56.7	69.6	59.7	53.93
1024	52.8	54.5	40.4	41.6	**55.3**	55.9	65.8	58.6	53.11

Table 9: Ablation Studies on the Impact of Codebook Size

B.4 Codebook size

Table 9 presents the performance of the FCID model across various codebook sizes. It is observed that the model achieves the best average results when the codebook size is set to 400. Conversely, using either a excessively large or small codebook size may lead to insufficient semantic learning or inadequate semantic expression, resulting in decreased model performance.

B.5 Reconstruction

As shown in Figure 10, for all columns except the 'origin' column, the images on the left represent reconstructions with random masks, while the images on the right illustrate reconstructions using the dimensions with the highest TOC retention scores. It is evident that TOC significantly outperforms random masking in reconstructions with mask ratios ranging from 25.0% to 87.5%, with the performance gap becoming increasingly pronounced as the mask ratio increases.

Figure 10: More results of reconstructions using random and TOC masking.

B.6 Generation

As shown in Figure 11, thanks to multimodal unified representations, the results of cross-modal image generation from audio and text closely resemble actual images. As evident in samples 2 and 6, despite the audio not mentioning specific details

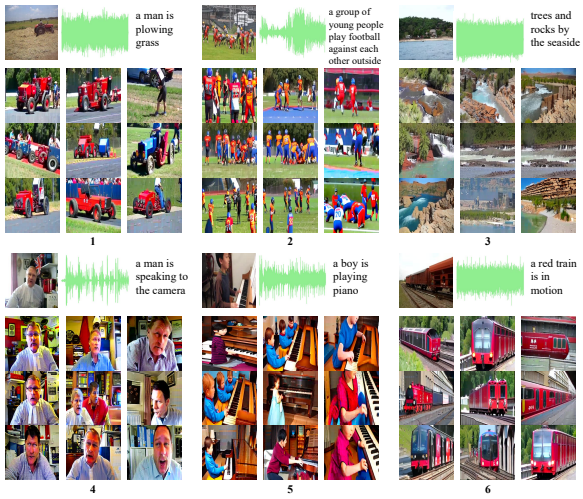

Figure 11: More results of cross-modal generation.

Method	MSCOCO(V↔T)			Clotho(A↔T)			FlickrSound(V↔A)			Avg.
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
CMCM (Liu et al., 2021a)	0.50	4.20	7.20	1.62	8.04	14.87	2.20	9.80	15.60	7.11
DCID (Xia et al., 2024)	0.80	**5.00**	8.30	2.06	9.00	16.70	**3.10**	11.10	17.20	8.14
FCID	**1.30**	4.90	**9.60**	**2.87**	**10.73**	**18.19**	**3.10**	**11.80**	**17.50**	**8.89**
CMCM+TOC	0.70	4.50	7.70	1.93	8.43	15.33	2.40	10.60	16.10	7.52(+0.41)
DCID+TOC	1.10	**5.30**	8.80	2.59	9.00	17.08	3.60	11.80	17.80	8.56(+0.42)
FCID+TOC	**1.50**	5.10	**10.40**	**3.16**	**11.15**	**19.04**	**3.80**	**12.20**	**18.40**	**9.42**(+0.53)

Table 10: Details of comparison with SOTA methods on three cross-modal zero-shot retrieval tasks, all results are calculated as the mean across two directions.

A_{CLUB}	V_{CLUB}	AV_{CLUB}	TE_{CLUB}	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	
-	-	-	-	51.3	51.6	39.5	40.7	50.6	51.1	63.3	57.6	50.71
✓	-	-	-	52.4	53.5	40.9	42.4	53.1	54.2	66.0	59.8	52.79
-	✓	-	-	53.1	53.4	41.7	43.2	53.9	54.7	67.1	60.1	53.40
-	-	✓	-	52.2	51.9	40.2	41.7	52.4	52.5	64.2	59.1	51.78
-	-	-	✓	51.7	51.5	40.6	41.8	52.5	52.9	63.5	58.2	51.59
✓	✓	-	-	54.2	54.0	41.4	43.9	**55.9**	56.1	67.9	61.3	54.34
-	-	✓	✓	52.9	52.6	40.8	42.1	52.5	53.9	65.7	59.2	52.46
✓	✓	✓	✓	**55.2**	**54.9**	**42.4**	**44.5**	55.3	**57.4**	**69.4**	**61.6**	**55.09**

Table 11: Details of ablation studies on the impact of FCID

such as the color of clothing and trains, these elements are still accurately generated, which can be attributed to the discrete unified representation serving as a central semantic hub for multiple modalities. In contrast, the results from Text-to-Image (T→I) are noticeably inferior to those from Image-to-Image (I→I) and Audio-to-Image (A→I). This difference is exemplified in the first image generated from sample 1's text, where the action of a car mowing grass is mistakenly transformed into a man mowing grass. This discrepancy arises because the semantic connections between images and audio are stronger than those generated through model-based text, which merely mentioned 'man' and 'plowing grass' without specifying the tool used for plowing.