Towards Uniformity and Alignment for Multimodal Representation Learning

Anonymous Authors[1]

Abstract

Multimodal representation learning aims to construct a shared embedding space in which heterogeneous modalities are semantically aligned. Despite strong empirical results, InfoNCE-based objectives introduce inherent conflicts that yield *distribution gaps* across modalities. In this work, we identify two conflicts in the multimodal regime, both exacerbated as the number of modalities increases: (i) an *alignment–uniformity* conflict, whereby the repulsion of uniformity undermines pairwise alignment, and (ii) an *intra-alignment* conflict, where aligning multiple modalities induces competing alignment directions. To address these issues, we propose a principled decoupling of alignment and uniformity for multimodal representations, providing a conflict-free recipe for multimodal learning that simultaneously supports discriminative and generative use cases without task-specific modules. We then provide a theoretical guarantee that our method acts as an efficient proxy for a global Hölder divergence over multiple modality distributions, and thus reduces the distribution gap among modalities. Extensive experiments on retrieval and UnCLIP-style generation demonstrate consistent gains.

1. Introduction

Multimodal representation learning (Ruan et al., 2023; Girdhar et al., 2023) aims to construct a shared embedding space where semantically related signals from different modalities (e.g., image, text, audio, video, speech) are well aligned. A landmark example is CLIP (Radford et al., 2021), which employs an InfoNCE objective to align paired image–text representations by maximizing similarity for positive pairs while pushing negative pairs apart. This framework has since been extended beyond two modalities. For instance, ImageBind (Girdhar et al., 2023), VAST (Chen et al., 2023), and LanguageBind (Zhu et al., 2024) incorporate additional modalities (e.g., depth and audio) into a common space, while GRAM (Cicchetti et al., 2025) generalizes the cosine similarity to the Gramian volume among multiple modalities in InfoNCE and achieves promising performance.

Despite notable successes, InfoNCE-based methods exhibit inherent conflicts that induce distribution gaps (Liang et al., 2022; Shi et al., 2023; Yin et al., 2025). Consequently, UnCLIP-type generative models (e.g., DALL-E 2 (Ramesh et al., 2022) and Kandinsky (Razzhigaev et al., 2023)) add a diffusion module to transform CLIP embeddings. Prior work (Yin et al., 2025) shows that, in vision–language learning, this gap arises from a conflict between *uniformity* and *alignment* (Wang & Isola, 2020): the uniformity term spreads embeddings on the unit hypersphere, whereas the alignment term pulls positive (multimodal) pairs together. However, in the multimodal regime, it remains unclear how these conflicts evolve as the number of modalities M increases, which is crucial for balancing uniformity (discriminability for retrieval) and alignment (closing cross-modal distribution gaps for generation). This is challenging because each modality is jointly determined by multiple cross-modal interactions, making the conflicts hard not only to characterize but also to resolve without sacrificing either retrieval discriminability or generative alignment. To tackle the issue, we explicitly quantify these internal conflicts, which explains how distribution gaps worsen as M grows and directly guides the design of a conflict-free objective.

Contributions. In this work, we systematically analyze and address the inherent conflicts in multimodal InfoNCE that give rise to modality and distributional gaps. Our contributions are four-fold and are highlighted below.

First, we provide a theoretical analysis of the InfoNCE in multimodal settings ($M \geq 3$), which is non-trivial due to a more complex and heterogeneous representation geometry. We theoretically formalize two distinct conflicts: (1) an alignment-uniformity conflict (ζ_a), where uniformity forces oppose alignment, exacerbating distributional gaps across modalities (see Fig. 1 (a) and Proposition 2.2 in Sec. 2), and (2) an intra-alignment conflict (χ_a), driven by non-collinear positive embeddings across modalities, which widens the modality gap as the number M of modalities increases (see

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Fig. 1 (b) and Proposition 2.3 in Sec. 2). Together, these conflicts explain why multimodal InfoNCE struggles to scale: the same objective that enforces global uniformity undermines the alignment of positive pairs, especially as the modality count M grows.

Second, to resolve these issues, we propose *UniAlign*, which provides a principled decoupling of **uni**formity and **align**ment. Specifically, we enforce *intra-modality uniformity* within each modality's samples, ensuring uniform coverage on the unit hypersphere and preventing representation collapse. In parallel, we introduce an *anchor-based alignment* strategy that aligns embeddings of the same sample across modalities with respect to an anchor modality. This explicitly avoids competing alignment directions, thereby closing modality gaps without introducing competing forces. Therefore, UniAlign enhances both cross-modal discriminative and generative capability.

Third, beyond this geometric intuition, we provide a theoretical justification that our method minimizes the distribution gap. Specifically, we introduce a global Hölder divergence applicable to an arbitrary number of modality distributions. We then connect our decoupled losses to this divergence, showing that the intra-modality uniformity and anchor-based alignment terms act as efficient computational proxies for minimizing it, thereby providing formal theoretical justification.

Extensive experimental results demonstrate the consistent superiority of our UniAlign over InfoNCE-based baselines in representation quality, retrieval accuracy, and generation fidelity. Without additional task-specific modules, the learned embeddings support both discriminative (cross-modal retrieval) and generative (UnCLIP-style conditional generation) tasks, yielding around 2 R@1 gains and 10–40 lower FID, respectively. These results confirm that our decoupled principle not only resolves the modality and distributional gaps introduced by InfoNCE, but also provides a scalable recipe for robust and versatile multimodal learning.

2. Multimodal Conflict Analysis

In this section, we first revisit prior analyses of the InfoNCE objective for two modalities (vision and language). We then quantify two types of conflicts in multimodal learning and present a principled framework that characterizes how conflicts evolve as the number of modalities increases.

2.1. Uniformity and Alignment Conflict of InfoNCE.

Let M be the number of modalities and B the batch size. For sample index $i \in \{1, \dots, B\}$ and modality $m \in \{1, \dots, M\}$, denote the ℓ_2-normalized embedding by $\mathbf{Z}^{(m)} = \{\mathbf{z}_i^{(m)} \in \mathbb{R}^d\}_{i=1}^B$. The generalized multi-modal InfoNCE objective (Oord et al., 2018) (over all pairs) is:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{\sum_{m\neq n} w_{mn}} \sum_{i=1}^{B} \sum_{m\neq n} w_{mn} \log p_{ii}^{(mn)} \quad \text{with } p_{ik}^{(mn)} = \frac{\exp(\mathbf{z}_i^{(m)\top} \mathbf{z}_k^{(n)}/\tau)}{\sum_{\ell=1}^{B} \exp(\mathbf{z}_i^{(m)\top} \mathbf{z}_\ell^{(n)}/\tau)}, \tag{1}$$

where $w_{mn} > 0$ denotes weight and $\tau > 0$ is the temperature. This loss has been extensively used in recent multimodal applications (Girdhar et al., 2023; Guzhov et al., 2022). Then for two modalities, InfoNCE can be decomposed into *alignment* and *uniformity* (Wang & Isola, 2020):

$$\begin{aligned} &\text{Alignment: } \mathbb{E}_{p_{\text{pair}}}\big[\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\|_2^2\big], \\ &\text{Uniformity: } \log \mathbb{E}_{p_{\text{data}}}\big[\exp(-\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\|_2^2/2\tau)\big], \end{aligned} \tag{2}$$

where p_{pair} is defined on cross-modal pairs, and p_{data} is the overall data distribution regardless of pairing. Uniformity spreads embeddings over the unit hypersphere, thereby avoiding collapse and promoting semantic coverage, while alignment pulls paired cross-modal representations together to enforce semantic consistency. In vision–language learning, Yin et al. (2025) demonstrates that uniformity *across* modalities ("inter-uniformity") conflicts with the alignment term, resulting in a systemic distributional gap (see evidence in Appendix F.2).

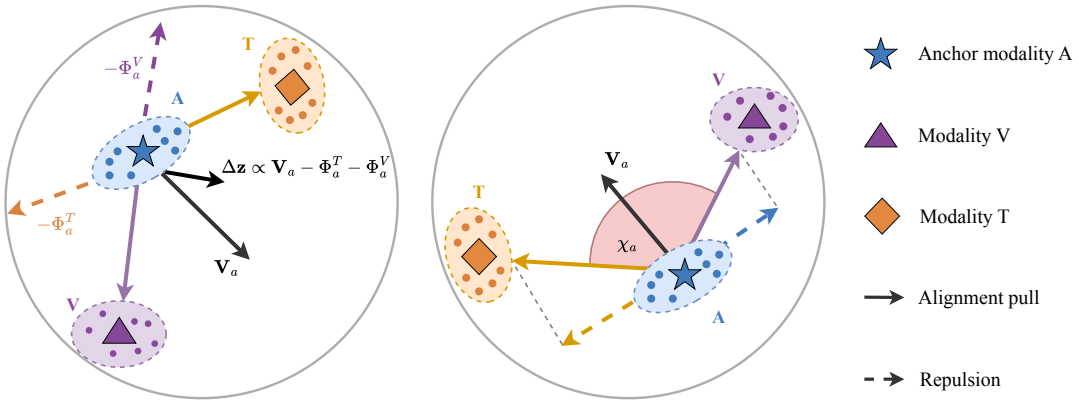
However, when extending to more modalities, the analysis is insufficient to present the relationship between conflict degree and the number of modalities, which is important to understand the learning mechanism of multimodal representation. Due to the more complex representation space geometry of multiple modalities, where each modality is influenced by multiple factors, quantifying the conflict in multimodal representation learning is challenging.

2.2. Systematic Multimodal Conflict Analysis

We first reveal two modes of conflict in multimodal learning with InfoNCE, and then prove that the two conflicts become severe when the number of modalities M increases by Proposition 2.2 and 2.3.

To quantify the conflict in multimodal learning, we first choose one modality $\mathbf{z}^{(a)}$ as the anchor, and analyze how it is influenced by other modalities from the gradient perspective. Differentiating Eq. (1) with respect to an anchor $\mathbf{z}_i^{(a)}$ exposes a "push–pull" structure. For a modality pair $(a \to n)$,

$$\nabla_{\mathbf{z}_i^{(a)}} \mathcal{L} = \underbrace{-\sum_{n\neq a} \frac{w_{an}}{\tau} \mathbf{z}_i^{(n)}}_{\mathbf{V}_a} + \underbrace{\sum_{n\neq a} \frac{w_{an}}{\tau} \sum_{k=1}^{B} p_{ik}^{(an)} \mathbf{z}_k^{(n)}}_{\mathbf{\Phi}_a}, \tag{3}$$



(a) Alignment–Uniformity Conflict *(b)* Intra-Alignment Conflict

Figure 1. **Two conflicts of multi-modal InfoNCE.** (a) *Alignment–uniformity:* positives are pulled together yet repelled by the uniformity force; (b) *Intra-alignment:* non-collinear positives induce angular tension. Both grow with M.

where the first term $\mathbf{V}_a$ aggregates only the matched positives across modalities and thus acts as an attraction (alignment) force, and the second term $\mathbf{\Phi}_a$ is a softmax-weighted mixture over the batch and captures the repulsion induced by negatives, i.e., the uniformity-related force.

Eq. (3) exposes two levels of conflicts. (i) *Inter-modality alignment–uniformity conflict*: when the uniformity push and alignment pull are directionally aligned, i.e., $\langle \mathbf{V}_a, \mathbf{\Phi}_a \rangle > 0$, the term $-\mathbf{V}_a + \mathbf{\Phi}_a$ cancels in the gradient, yielding a small update for $\mathbf{z}^{(a)}$ (see Fig. 1 (a)). This conflict is rooted in *inter-modality* uniformity, which induces repulsion across modalities and leads to a distribution gap. (ii) *Intra-alignment conflict*: the net alignment force weakens when the positive targets are not collinear (Fig. 1(b)). Since $\mathbf{V}_a = \sum_{n \neq a} \frac{w_{an}}{\tau} \mathbf{z}^{(n)}$ aggregates these positive pulls via a vector sum, collinear positives reinforce each other and yield a large $\|\mathbf{V}_a\|$, whereas non-collinear or opposing positives partially cancel, reducing the alignment signal.

Conflict quantification (ζ_a, χ_a). To quantitatively analyze these two conflicts, we define the *alignment-uniformity conflict* $\zeta_a \in [-1, 1]$ to measure directional opposition between the alignment and uniformity forces, and introduce the intra-alignment conflict $\chi_a \in [0, 1]$ to quantify cancellation among non-collinear positive pulls within $\mathbf{V}_a$:

$$\begin{aligned} \zeta_a &\triangleq \cos\big(\mathbf{V}_a, \mathbf{\Phi}_a\big) = \mathbf{V}_a^\top \mathbf{\Phi}_a / \big(\|\mathbf{V}_a\|_2 \, \|\mathbf{\Phi}_a\|_2\big). \\ \chi_a &\triangleq 1 - \|\mathbf{V}_a\|_2 / \big(\sum\nolimits_{n \neq a} w_{an}/\tau\big), \end{aligned} \tag{4}$$

A high positive ζ_a (near 1) indicates severe conflict, which occurs when hard negatives lie in the same direction as positives. χ_a measures how much the positive "pull" vectors within $\mathbf{V}_a$ cancel due to non-collinearity. A value of $\chi_a = 0$ indicates perfect alignment (no conflict), whereas $\chi_a \to 1$ indicates severe conflict.

To study how the alignment–uniformity conflict scales with the number of modalities M, we next introduce a mild structural assumption on each per-modality uniformity component. Intuitively, it separates a systematic component that is coherent across modalities from residual variations that are largely modality-specific.

Assumption 2.1 (Systematic conflict per-modality)**.** Let $\hat{\mathbf{V}}_a = \mathbf{V}_a / \|\mathbf{V}_a\|$ denote the unit alignment direction for anchor a. For each modality $n \neq a$, the uniformity component admits the decomposition

$$\mathbf{\Phi}_a^{(n)} \;=\; c_n \, \hat{\mathbf{V}}_a \;+\; \varepsilon_n, \tag{5}$$

where $c_n \triangleq \langle \mathbf{\Phi}_a^{(n)}, \hat{\mathbf{V}}_a \rangle$ quantifies the magnitude of systematic conflict from modality n in this direction, and satisfies $c_n \geq c_0$ for positive constant c_0. The residuals ε_n are zero-mean, mutually independent, and have bounded covariance.

Assumption Justification. Most negatives in large-batch multimodal training are semantically unrelated and approximately isotropic, so their contributions largely average out as zero-mean noise. A small fraction of hard negatives, however, concentrates near the positive alignment direction and yields a consistent projection onto $\hat{\mathbf{V}}_a$. This motivates decomposing each term into a systematic component c_n (hard negatives) plus a residual ε_n (easy negatives and noise), and implies that as M increases, the systematic parts add coherently while residuals average out (see Appendix A).

Proposition 2.2 (Alignment–Uniformity Conflict)**.** *Let $\mathbf{\Phi}_a = \sum_{n \neq a} \mathbf{\Phi}_a^{(n)}$ be the total uniformity force on anchor a, and define $\zeta_a = \cos\big(\mathbf{V}_a, \mathbf{\Phi}_a\big)$. Under Assumption 2.1, the alignment–uniformity conflict converges to its maximum as the number of modalities M increases:*

$$\mathbb{E}[\zeta_a] \;=\; \mathbb{E}\big[\cos\big(\mathbf{V}_a, \mathbf{\Phi}_a\big)\big] \;\longrightarrow\; 1 \qquad \text{as } M \to \infty. \tag{6}$$

See its proof in Appendix A. This shows that even if the conflict from each modality (c_n) is small, their systematic accumulation inevitably causes the total, observable conflict (ζ_a) to become severe.

Proposition 2.3 (Intra-alignment Conflict)**.** *The expected intra-alignment conflict, $\mathbb{E}[\chi_a]$, is governed by M and the average pairwise alignment $\bar{\mu} = \mathbb{E}[\mathbf{z}_i^{(m)\top}\mathbf{z}_i^{(n)}] \in [0,1]$ for $m \neq n$ between modalities:*

$$\mathbb{E}[\chi_a] \;\geq\; 1 - \sqrt{(1+(M-2)\bar{\mu})/(M-1)}. \tag{7}$$

For imperfect alignment ($\bar{\mu} < 1$), the conflict increases with the number of modalities M and admits a non-zero asymptotic lower bound:

$$\liminf_{M\to\infty} \mathbb{E}[\chi_a] \;\geq\; 1 - \sqrt{\bar{\mu}}. \tag{8}$$

See its proof in Appendix B. Proposition 2.3 shows that the internal conflict of the alignment force gets severe with more modalities, resulting in ineffective alignment.

By combining Proposition 2.2 and 2.3, one can conclude that the standard multi-modal InfoNCE objective is fraught with a two-level conflict system: an intra-alignment conflict, and a classic alignment-uniformity conflict. Such conflicts result in a distinct distributional modality gap. This motivates the exploration of alternative frameworks that **decouple** these objectives, by optimizing uniformity separately and employing a more direct, conflict-free alignment mechanism.

3. Methodology

In Section 2, our analysis has identified two fundamental conflicts that impede multi-modal contrastive learning: the intra-alignment conflict (χ), and the alignment-uniformity conflict (ζ). To circumvent these issues, we propose a generic principle to decouple the learning objectives. Then, we show that our principle essentially minimizes the global distribution gap with a theoretical guarantee.

3.1. General Principle for Multimodal Learning

A general principle for multimodal learning

As the *alignment-uniformity* and intra-alignment conflicts are the root for the modality/distribution gap, avoiding these conflicts from the uniformity and alignment perspective is a good general principle:

1. **Intra-modality uniformity only:** encourage uniform spread *within each modality* $\{\mathbf{z}_i^{(m)}\}$ on $\mathbb{S}^{d-1}$, and avoid any *cross-modality* uniformity/repulsion.
2. **Conflict-free alignment:** Explicitly or implicitly maximize the consensus magnitude to avoid the non-collinearity problem between positive pairs.

1 avoids the cross-modality uniformity conflict but still pushes the embeddings uniformly spreading in a unit hypersphere. 2 avoids the non-collinear positive pulls in the consensus vector.

Following this generic principle, we present our design for the uniformity and alignment terms in Euclidean space. See a summarization in Table 4 of Appendix D.

Uniformity. Let $\mathbf{Z}^{(m)} = \{\mathbf{z}_i^{(m)} \in \mathbb{R}^d\}_{i=1}^B$ denote a batch of unit-normalized embeddings from modality m. To promote uniformity of multimodal representations and mitigate inter-modality conflict, we adopt an *intra-modality* uniformity term to prevent collapse:

$$U\big(\mathbf{Z}^{(m)}\big) = \frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{1}{B-1}\sum\nolimits_{j\neq i} \kappa\big(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(m)}\big)\right), \quad \text{with} \quad \kappa(\mathbf{z}_i,\mathbf{z}_j) = \exp\big(-\|\mathbf{z}_i-\mathbf{z}_j\|_2^2/2\tau^2\big), \tag{9}$$

where $\tau > 0$ is the temperature and κ is a Gaussian kernel. The sample-wise gradient satisfies $\nabla_{\mathbf{z}_i^{(m)}} U = -\frac{1}{\tau^2}\sum_{j\neq i} p_{ij}\,(\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)})$, $p_{ij} = \frac{\exp\left(-\|\mathbf{z}_i^{(m)}-\mathbf{z}_j^{(m)}\|_2^2/(2\tau^2)\right)}{\sum_{\ell\neq i}\exp\left(-\|\mathbf{z}_i^{(m)}-\mathbf{z}_\ell^{(m)}\|_2^2/(2\tau^2)\right)}$, so the softmax weights p_{ij} decay exponentially with distance, effectively suppressing far-away contributions. Consequently, the uniformity term concentrates gradients on nearby hard negatives (semantically similar samples) while leaving distant points largely unaffected, improving local uniformity and preventing collapse. Note that, as the $U(\mathbf{Z}^{(m)})$ is defined for each modality separately, which is different from the Φ_a term in Eq. (1), hence, the conflict ζ_a is avoided.

Conflict-free alignment. To address the intra-alignment conflict (χ) inherent in standard multimodal ($M \geq 3$) contrastive objectives, we propose an anchor-based alignment. This design enforces a single positive-pull direction per sample, thereby avoiding non-collinearity among positive pairs.

Specifically, we choose one modality $\mathbf{Z}^{(a)}$ as the anchor, which serves as a reference direction. As in other modalities, we apply intra-modality uniformity to the anchor embeddings, while aligning all remaining modalities to this anchor. As a result, each sample is aligned along a single direction, preventing competing (non-collinear) positive pulls.

Hence, the most straightforward alignment loss on the Euclidean space can be defined by:

$$L_{\text{align}} = \frac{1}{B(M-1)}\sum\nolimits_{i=1}^{B}\sum\nolimits_{n\neq a}\big\|\mathbf{z}_i^{(a)} - \mathbf{z}_i^{(n)}\big\|_2^2. \tag{10}$$

Overall objective. The final objective with hyperparameter $\lambda_{\text{align}} > 0$ is defined by:

$$\mathcal{L} = \sum\nolimits_{m=1}^{M} U\big(Z^{(m)}\big) + \lambda_{\text{align}}\, L_{\text{align}}. \tag{11}$$

Hypersphere space. Our principle is not restricted to Euclidean space and can be extended to manifolds. Since both

InfoNCE and our uniformity term encourage representations to spread on the unit hypersphere, a natural alternative is to measure similarity via the geodesic distance on $\mathbb{S}^{d-1}$:

$$d_{\mathbb{S}}(\mathbf{z}^i, \mathbf{z}^j) = \arccos\big(\langle \mathbf{z}^i, \mathbf{z}^j \rangle\big), \qquad \|\mathbf{z}^i\|_2 = \|\mathbf{z}^j\|_2 = 1,$$
$$k_{\mathbb{S}}(\mathbf{z}^i, \mathbf{z}^j; \tau) = \exp\Big(-d_{\mathbb{S}}(\mathbf{z}^i, \mathbf{z}^j)^2 / 2\,\tau^2\Big). \tag{12}$$

This yields a drop-in hyperspherical variant within the same framework, illustrating the flexibility of our design. We validate this choice in Table 3 and summarize Euclidean vs. manifold instantiations in Table 4 (Appendix D).

3.2. Theoretical Analysis from Divergence Perspective

In the previous section, we introduced our objective based on the proposed principle. A natural question is whether this objective is theoretically guaranteed to reduce the distribution gap across modalities. Here, we show that optimizing intra-modality uniformity and cross-modality alignment minimizes a global distribution divergence, thereby mitigating the cross-modal (distribution) gap.

Classical divergences (Jenssen et al., 2006; Shlens, 2014) are typically defined between two distributions, but our setting involves M modalities. We therefore introduce a new *global Hölder divergence* to jointly measure the discrepancy among all modality distributions. Let $\{p_m(\mathbf{z})\}_{m=1}^M$ denote the densities of the M modalities. By Hölder's inequality, they satisfy

$$\left| \int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z} \right|^M \le \prod_{m=1}^M \int |p_m(\mathbf{z})|^M d\mathbf{z}, \tag{13}$$

and takes equality if and only if $p_1 = \cdots = p_M$. This inequality motivates the definition of the global Hölder divergence as the log of the ratio between the two sides of Eq. (13):

$$D_{\text{Hölder}} = -\log \frac{\int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z}}{\Big(\prod_{m=1}^M \int |p_m(\mathbf{z})|^M d\mathbf{z}\Big)^{\frac{1}{M}}}$$
$$= \underbrace{\frac{1}{M}\sum_{m=1}^M \log \int |p_m(\mathbf{z})|^M d\mathbf{z}}_{\text{Uniformity Term}} \underbrace{- \log \int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z}}_{\text{Alignment Term}}. \tag{14}$$

We empirically estimate this global divergence in a non-parametric way via a kernel density estimator (KDE) with a Gaussian kernel. Under the KDE plug-in estimator derived in Appendix C, the term $\frac{1}{M}\sum_{m=1}^M \log \int |p_m(\mathbf{z})|^M d\mathbf{z}$ can be approximated by averaging Gaussian similarities between samples within the same modality, which motivates our intra-modality uniformity objective $U(Z^{(m)})$ in Eq. (9) (up to constants). Likewise, the second term $-\log \int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z}$ measures how much the modality distributions overlap in the shared embedding space. It can be approximated by averaging Gaussian similarities between embeddings from different modalities. Maximizing this overlap admits an efficient instance-matching surrogate in the low-temperature limit ($\tau \to 0$), motivating L_{align} in Eq. (10). Therefore, optimizing intra-modality uniformity together with cross-modality alignment provides a tractable way to reduce the global Hölder divergence in Eq. (14).

3.3. Tuple-Level Extensions

Our design principle is *instantiation-agnostic*: it targets uniformity across samples and alignment across modalities, while remaining compatible with a broad class of loss constructions. Beyond anchor-wise (pairwise) formulations, a complementary direction is to treat each multimodal tuple as a *single structured sample* and regularize its *within-tuple geometry*. This tuple-level view strengthens both uniformity and alignment from a different angle.

We propose a tuple-level complement that (i) enforces *uniform dispersion* of tuples in the representation space, and (ii) encourages *cross-modal collinearity* within each tuple.

Tuple-level uniformity $U(\mathbf{C})$. Given a multimodal tuple $\{\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(M)}\}$, we represent it by a weighted and normalized centroid $\mathbf{c}_i$, and denote $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^B$:

$$\mathbf{c}_i = \sum_{m=1}^M w_m\, \mathbf{z}_i^{(m)} \Big/ \Big\| \sum_{m=1}^M w_m\, \mathbf{z}_i^{(m)} \Big\|_2, \quad \text{with} \quad w_m \ge 0, \quad \sum_{m=1}^M w_m = 1. \tag{15}$$

We then apply the same uniformity objective to $\mathbf{C}$, i.e., $U(\mathbf{C})$ (default $w_m = 1/M$), which encourages tuples (rather than individual modalities) to spread out and improves global separability.

Tuple-level alignment L_{vol}. To capture agreement among all modalities within each tuple, we penalize the *Gram-determinant volume* spanned by $\{\mathbf{z}_i^{(m)}\}_{m=1}^M$, which equals zero when the modality embeddings are collinear. Let $\mathbf{G}_i \in \mathbb{R}^{M\times M}$ be the Gram matrix (Cicchetti et al., 2025) with $[\mathbf{G}_i]_{mn} = \langle \mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)} \rangle$. The induced volume is proportional to $\sqrt{\det(\mathbf{G}_i)}$, and we define

$$L_{\text{vol}} = \frac{1}{B}\sum_{i=1}^B \sqrt{\det(\mathbf{G}_i)}. \tag{16}$$

Minimizing L_{vol} explicitly promotes cross-modal collinearity, complementing the anchor-based alignment.

4. Related Work

CLIP (Radford et al., 2021) pioneered aligning two modalities (vision and language) using the InfoNCE objective (Oord et al., 2018). It has enabled substantial progress

Table 1. **Zero-shot multimodal text-to-video (T2V) and video-to-text (V2T) retrieval results (Recall@1).** Our method, UniAlign, consistently outperforms baselines in most tasks. * denotes the results without tuple-level losses ($U(\mathbf{C})$ and L_{vol}).

Method	Modality	MSR-VTT		DiDeMo		ActivityNet		Average	
		T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
UMT (Liu et al., 2022)	T–V	33.3	–	34.0	–	31.9	–	33.1	–
OmniVL (Wang et al., 2022a)	T–V	34.6	–	33.3	–	–	–	34.0	–
UMT-L (Li et al., 2023)	T–V	40.7	37.1	48.6	49.9	41.9	39.4	43.7	42.1
TVTSv2 (Zeng et al., 2023)	T–V	38.2	–	34.6	–	–	–	36.4	–
ViCLIP (Wang et al., 2023a)	T–V	42.4	41.3	18.4	27.9	15.1	24.0	25.3	31.1
VideoCoCa (Yan et al., 2022)	T–V	34.3	64.7	–	–	34.5	33.0	34.4	48.9
Norton (Lin et al., 2024)	T–V	10.7	–	–	–	–	–	10.7	–
ImageBind (Girdhar et al., 2023)	T–V	36.8	–	–	–	–	–	36.8	–
InternVideo-L (Wang et al., 2022b)	T–V	40.7	39.6	31.5	33.5	30.7	31.4	34.3	34.8
HiTeA (Ye et al., 2023)	T–V	34.4	–	43.2	–	–	–	38.8	–
mPLUG-2 (Xu et al., 2023)	T–V	47.1	–	45.7	–	–	–	46.4	–
VideoPrism-b (Zhao et al., 2024)	T–V	51.4	50.2	–	–	49.6	47.9	50.5	49.1
LanguageBind (Zhu et al., 2024)	T–V	44.8	40.9	39.9	39.8	41.0	39.1	41.9	39.9
VAST (Chen et al., 2023)	T–VA	49.3	43.7	49.5	48.2	51.4	46.8	50.1	46.2
GRAM (Cicchetti et al., 2025)	T–VA	54.2	50.5	54.2	**52.2**	59.0	50.4	55.8	51.0
UniAlign* (Ours)	T–VA	57.7	53.2	55.2	51.9	59.2	**52.5**	57.4	52.5
UniAlign (Ours)	T–VA	**58.7**	**54.6**	**58.2**	51.6	**59.4**	51.7	**58.8**	**52.6**

in image–text retrieval (Jang et al., 2024; Koukounas et al., 2024; Huang et al., 2024) and text-to-image (T2I) generation (Ramesh et al., 2022; Rombach et al., 2022). CLIP-style contrastive objectives have since been applied to additional modality pairs, including audio–text (Elizalde et al., 2023; Wu et al., 2023) and point cloud–text (Zhang et al., 2022). Beyond pairs, recent work such as CMRC (Wang et al., 2023b), CLIP4VLA (Ruan et al., 2023), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2024) extends CLIP by introducing more modalities (e.g., video, audio, depth, IMU) into a unified space using pairwise InfoNCE objectives. In parallel, VAST (Chen et al., 2023), mPLUG-2 (Xu et al., 2023), and InternVideo2 (Wang et al., 2024) advance the state of the art through large-scale training and architectural refinements. Complementing these trends, GRAM (Cicchetti et al., 2025) introduces the cross-modality Gram matrix to replace pairwise cosine similarity in InfoNCE with a volume score given by the modality Gram matrix to better handle multimodal alignment.

Despite the success of these methods, embeddings from different modalities still exhibit distinct *distribution gaps* (Fig. 2), largely attributable to the InfoNCE objective. Prior studies (Zhou et al., 2023; Liang et al., 2022; Shi et al., 2023) have reported this phenomenon in vision–language learning: Liang et al. (2022) observe that the InfoNCE objective can encourage modality gaps, while Yin et al. (2025) provide a theoretical account showing that uniformity and alignment (Wang & Isola, 2020) conflict, inducing persistent distributional discrepancies. However, these analyses are restricted to the bimodal case; a principled understanding of the conflict mechanisms in the *multimodal* regime remains lacking, partly due to the geometric complexity of shared representation spaces. In this work, we systematically analyze these conflicts for general multimodal learning and, based on this analysis, propose a generic principle for multimodal representation learning.

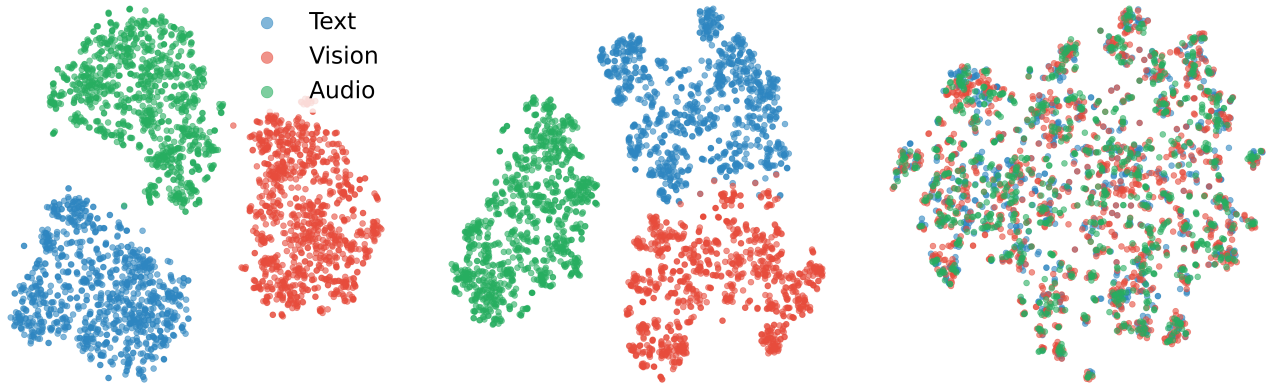
5. Experiments

We evaluate UniAlign from two aspects: (i) embedding separability via cross-modal retrieval (Sec. 5.1), and (ii) the distributional modality gap via cross-modal generation with fixed image decoders (Sec. 5.2). Additional implementation details and results are provided in Appendix F.

5.1. Cross-modal Retrieval

Experimental setting. Following GRAM (Cicchetti et al., 2025), we train on VAST150K (Chen et al., 2023) and evaluate *zero-shot* video retrieval on three standard benchmarks: MSR-VTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), and ActivityNet (Caba Heilbron et al., 2015). For fair comparison, we use the same modality encoders as VAST/GRAM: BERT-B for text, BEATs (Chen et al., 2022) for audio, and EVA-CLIP ViT-G (Sun et al., 2023) for video. We report zero-shot Recall@1 (R@1) for both text-to-video (T2V) and video-to-text (V2T) retrieval. For joint retrieval, where text queries the most similar video-audio tuple (T-VA) and vice versa, we follow GRAM and rank samples by the *volume* score (the determinant of the Gramian matrix).

Experimental results. Table 1 shows that UniAlign consistently outperforms strong baselines on zero-shot T2V/V2T retrieval. Notably, UniAlign* already yields solid gains even without tuple-level losses ($U(\mathbf{C})$ and L_{vol}). Starting



(a) ImageBind (paired InfoNCE). *(b)* GRAM (volume InfoNCE). *(c)* Ours

Figure 2. **T-SNE visualizations of vision, text, and audio features.** InfoNCE-type of objective results in clear distribution gaps (2a and 2b). Our method mitigates the distribution gap (2c).

Table 2. **Ablation on $U(C)$ and L_{vol}.** Both tuple-level uniformity $U(C)$ and tuple-level alignment L_{vol} yield consistent gains.

$U(C)$	L_{vol}	T2V	V2T	Avg.
✗	✗	36.5	36.8	36.6
✗	✓	38.3	36.9	37.6
✓	✗	37.4	39.8	38.6
✓	✓	40.2	43.4	41.8

from the same VAST pretrained weights, UniAlign improves VAST by 8.7 R@1 on T2V and 6.4 R@1 on V2T, demonstrating the effectiveness of our principle-instantiated objective. Moreover, since VAST uses pairwise InfoNCE, it is affected by both conflicts. GRAM introduces a Gramian volume score inside InfoNCE, which encourages collinearity (the volume is minimized when vectors become collinear) and thus alleviates the intra-alignment conflict, but the alignment-uniformity conflict remains. By explicitly addressing both conflicts, UniAlign further improves over GRAM by 3.0 R@1 on T2V and 1.6 on V2T.

Ablation study. To understand the effects of the tuple-level uniformity and alignment, $U(C)$ and L_{vol}, we ablate both components on MSR-VTT (training and testing). To isolate the contribution of our objectives, we use plain cosine-similarity retrieval, excluding common post-hoc refinements such as similarity matrix or image-text matching re-ranking. As shown in Table 2, both the tuple-level uniformity $U(C)$ and alignment L_{vol} improve embedding separability and consistently boost retrieval performance. More ablations are provided in Appendix F.3.

5.2. Cross-modal Generation

To evaluate the distributional modality gap, we use a simple proxy: if multiple modalities are well aligned to a shared embedding distribution, embeddings from non-image modalities (e.g., audio or text) should be seamlessly decodable by an image generator trained on image embeddings. Under this view, cross-modal generation quality provides a direct indicator of cross-modal alignment. Accordingly, we adopt UnCLIP-style generators (Ramesh et al., 2022), which use a separate image generator trained on image embeddings.

Dataset. We use the VGGSound dataset (Chen et al., 2020) to evaluate the generation performance. VGGSound is an audio-visual correspondent dataset, allowing us to build a semantically aligned vision-audio-text triplet. It has around 200K video clips, annotated with 309 sound classes. The dataset does not provide the video caption. Hence, we use the captioner provided by VAST (Chen et al., 2023) to generate video captions. We select 1024 videos for testing and utilize the remaining ones for training.

Experimental setting. We map all modalities into a shared, image-anchored embedding space and evaluate two encoder-decoder configurations compatible with UnCLIP-style generators. (i) *ViT-H*: CLIP ViT-H/14 image-text encoders paired with the ImageBind audio encoder, compatible with the Stable UnCLIP decoder (Ramesh et al., 2022). (ii) *ViT-bigG*: CLIP ViT-bigG-14 vision-text encoders combined with the BEATs audio encoder (Chen et al., 2022), compatible with the Kandinsky decoder (Razzhigaev et al., 2023). For comparison, we re-train GRAM and use the released ImageBind weights pretrained on large-scale data. We evaluate text-to-image (T2I), audio-to-image (A2I), and modality interpolation generation using Fr'echet Inception Distance (FID) (Heusel et al., 2017).

T-SNE visualization. We visualize the joint embedding space using 2D t-SNE (Fig. 2) to illustrate modality gaps under different training objectives. We extract text, vision, and audio embeddings from VGGSound and compute t-SNE. As shown in Fig. 2, training with an InfoNCE-type objective (both paired and volume InfoNCE) yields clearly separated, modality-specific clusters (i.e., modality distribution gap), whereas our method produces substantially tighter cross-modal alignment, significantly reducing the distribution gap.

Table 3. **Cross-modal generation with different decoders.** We report FID (↓). Kandinsky and Stable UnCLIP, evaluated in self-reconstruction by feeding image embeddings to the decoder (marked $*$), serve as upper-bound references. Our method consistently outperforms both baselines.

Decoder	Method	T2I ↓	A2I ↓	(T+A)→I ↓	Avg. ↓
Kandinsky	Kandinsky	-	-	-	32.99*
	GRAM	62.11	106.97	92.63	87.23
	Ours (Geodesic)	**45.35**	**50.75**	**48.19**	**48.09**
	Ours (Euclidean)	**42.72**	**50.51**	**46.56**	**46.60**
Stable UnCLIP	Stable UnCLIP	-	-	-	34.61*
	ImageBind	50.17	53.59	46.81	50.19
	GRAM	45.53	55.40	47.15	49.36
	Ours (Geodesic)	**39.88**	**40.16**	**40.80**	**40.23**
	Ours (Euclidean)	**39.63**	**39.95**	**41.03**	**40.20**

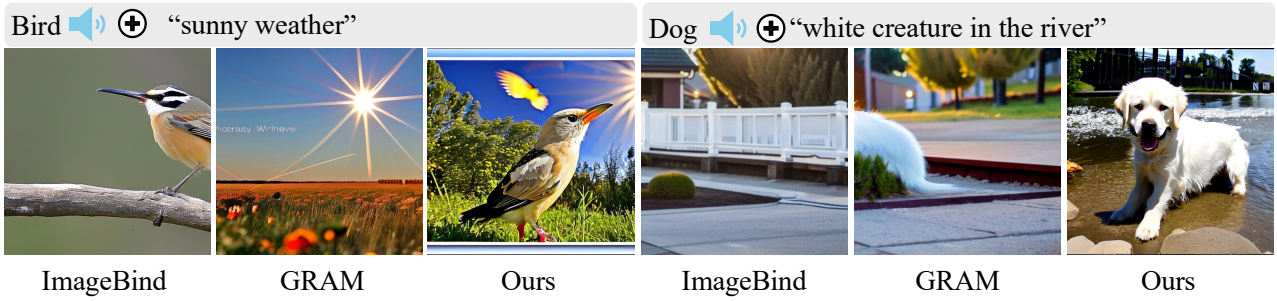


Figure 3. **Modality-interpolation generation results (T+A) → I.** When interpolating between text and audio representations, our method has a better ability to fuse the semantic information across modalities, leading to better generation.

Results. We compare against GRAM and ImageBind using both *Kandinsky* and *Stable UnCLIP* decoders. When conditioned on image embeddings, these decoders achieve FID 32.99 and 34.61, respectively, serving as dataset-specific upper bounds (decoder self-reconstruction). UniAlign yields substantial improvements in cross-modal generation for both text-to-image (T2I) and audio-to-image (A2I) over InfoNCE-trained baselines. As shown in Table 3, UniAlign significantly outperforms GRAM with *Kandinsky*, and improves over ImageBind and GRAM by about 10 FID with *Stable UnCLIP*. These gains are consistent across architectures and decoders, indicating the robustness of our objective and suggesting a reduced modality gap via tighter distributional alignment. We also evaluate a geodesic-kernel variant, which performs on par with the Euclidean version, further supporting the generality of our principle. Additional qualitative results are provided in Appendix F.

Modality interpolation. If multiple modalities are aligned to a shared distribution, a straightforward application is *embedding interpolation*, which blends information from different modalities directly in the shared space for image synthesis (as opposed to conditioning a generator via cross-attention from a single modality). Prior work has primarily demonstrated this for vision–language with DALL·E 2 (Ramesh et al., 2022). Here, we go beyond vision-language interpolation for generation. We interpolate modality embeddings (e.g., $(T+A)/2$ and generate images with Kandinsky and Stable UnCLIP decoders. Our method outperforms baselines both quantitatively (Table. 3 with lower FID) and qualitatively (Fig. 3), indicating an improved ability to fuse complementary semantics across modalities. We attribute these gains to the reduced cross-modal distribution gap and the resulted smoothness of the shared embedding manifold.

6. Conclusion

We introduced a conflict-aware principle for multimodal representation learning that decouples *uniformity* from *alignment*, overcoming key limitations of InfoNCE when modality number $M \geq 3$. By promoting intra-modality uniformity and anchoring positive alignment, our method directly reduces cross-modal distribution gaps. A divergence-based analysis further shows that these objectives serve as tractable estimators for minimizing a global discrepancy, providing theoretical guarantees. Empirically, the learned embeddings achieve strong performance in video retrieval and cross-modal generation with UnCLIP decoders, while t-SNE visualizations confirm improved modality integration. Overall, our approach offers a conflict-free and theoretically grounded framework for unifying discriminative and generative multimodal tasks without task-specific modules.

Limitation. Our study mainly focuses on fine-tuning rather than large-scale pretraining. Extending the proposed principle to large-scale pretraining would require substantial compute resources and is left for future work.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., and Liu, J. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

Cicchetti, G., Grassucci, E., Sigillo, L., and Comminiello, D. Gramian multimodal representation learning and alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ftGnpZrW7P.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Guzhov, A., Raue, F., Hees, J., and Dengel, A. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., Dai, Q., Dai, X., Chen, D., Luo, C., et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.

Jang, Y. K., Kang, J., Lee, Y. J., and Kim, D. Mate: Meet at the embedding–connecting images with long texts. *arXiv preprint arXiv:2407.09541*, 2024.

Jenssen, R., Principe, J. C., Erdogmus, D., and Eltoft, T. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

Koukounas, A., Mastrapas, G., Günther, M., Wang, B., Martens, S., Mohr, I., Sturua, S., Akram, M. K., Martínez, J. F., Ognawala, S., et al. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*, 2024.

Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., and Qiao, Y. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19948–19960, 2023.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Lin, Y., Zhang, J., Huang, Z., Liu, J., Wen, Z., and Peng, X. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024.

Liu, Y., Li, S., Wu, Y., Chen, C.-W., Shan, Y., and Qie, X. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural

language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., and Dimitrov, D. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruan, L., Hu, A., Song, Y., Zhang, L., Zheng, S., and Jin, Q. Accommodating audio modality in clip for multimodal processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9641–9649, 2023.

Shi, P., Welle, M. C., Björkman, M., and Kragic, D. Towards understanding the modality gap in clip. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Shlens, J. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*, 2014.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G., and Yuan, L. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022a.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022b.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023a.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

Wang, Z., Zhao, Y., Huang, H., Liu, J., Yin, A., Tang, L., Li, L., Wang, Y., Zhang, Z., and Zhao, Z. Connecting multimodal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023b.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pp. 38728–38748. PMLR, 2023.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Yan, S., Zhu, T., Wang, Z., Cao, Y., Zhang, M., Ghosh, S., Wu, Y., and Yu, J. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J., and Huang, F. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15405–15416, 2023.

Yin, W., Xiao, Z., Zhou, P., Yu, S., Shen, J., Sonke, J.-J., and Gavves, E. Distributional vision-language alignment by cauchy-schwarz divergence. *arXiv preprint arXiv:2502.17028*, 2025.

Zeng, Z., Ge, Y., Tong, Z., Liu, X., Xia, S.-T., and Shan, Y. Tvtsv2: Learning out-of-the-box spatiotemporal visual representations at scale. *arXiv preprint arXiv:2305.14173*, 2023.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

Zhao, L., Gundavarapu, N. B., Yuan, L., Zhou, H., Yan, S., Sun, J. J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.

Zhou, C., Zhong, F., and Öztireli, C. Clip-pae: projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9, 2023.

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C. W., Li, Z., Liu, W., and Yuan, L. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

A. Proof of Proposition 1

Proposition 1 (Alignment–uniformity Conflict). Let $\mathbf{\Phi}_a = \sum_{n \neq a} \mathbf{\Phi}_a^{(n)}$ be the total uniformity force on anchor a. Assume each per-modality component admits the decomposition:

$$\mathbf{\Phi}_a^{(n)} \;=\; c_n \, \hat{\mathbf{V}}_a \;+\; \boldsymbol{\varepsilon}_n. \tag{17}$$

Here, $\hat{\mathbf{V}}_a \triangleq \mathbf{V}_a / \|\mathbf{V}_a\|$ is the direction of the total alignment force. The scalar $c_n \triangleq \mathbf{\Phi}_a^{(n)} \cdot \hat{\mathbf{V}}_a$ quantifies the magnitude of systematic conflict from modality n in this direction, and satisfies $c_n \geq c_0$ for some positive constant c_0. The vector $\boldsymbol{\varepsilon}_n$ is a random perturbation unique to modality n, assumed to be zero-mean, mutually independent, and with bounded covariance.

Under these assumptions, the overall alignment-uniformity conflict ζ_a converges to its maximum value as the number of modalities M increases:

$$\mathbb{E}[\zeta_a] = \mathbb{E}\big[\cos\big(\mathbf{V}_a, \mathbf{\Phi}_a\big)\big] \;\to\; 1 \quad \text{as } M \to \infty. \tag{18}$$

Proof. The proof proceeds in three stages. First, we provide a formal justification for the decomposition of the per-modality uniformity force. Second, we derive a precise expression for the conflict metric ζ_a based on this decomposition. Finally, we analyze the asymptotic behavior of this expression as the number of modalities $M \to \infty$.

Justification of the Decomposition The decomposition of $\mathbf{\Phi}_a^{(n)}$ is a formalization of the geometric principle of orthogonal projection. For any vector $\mathbf{\Phi}_a^{(n)}$ and a given direction defined by the unit vector $\hat{\mathbf{V}}_a$, we can uniquely decompose $\mathbf{\Phi}_a^{(n)}$ into a component parallel to $\hat{\mathbf{V}}_a$ and a component orthogonal to it. The component parallel to $\hat{\mathbf{V}}_a$ is its orthogonal projection, which we define as the systematic component:

$$\text{Proj}_{\hat{\mathbf{V}}_a}(\mathbf{\Phi}_a^{(n)}) = (\mathbf{\Phi}_a^{(n)} \cdot \hat{\mathbf{V}}_a)\hat{\mathbf{V}}_a. \tag{19}$$

Intuitively, in each modality, non-paired but semantically similar samples ("hard negatives") exert a weak but systematic pull in the same direction as the true cross-modal target; this shared component is modeled by $c_n \hat{\mathbf{V}}_a$. Residual variation due to batch composition, data augmentations, and encoder stochasticity is captured by zero-mean, bounded perturbations $\boldsymbol{\varepsilon}_n$ that are approximately independent across modalities. As the number of modalities increases, the systematic components add coherently while the residuals average out, leading to the observed accumulation of alignment–uniformity conflict.

The conflict metric ζ_a is the cosine similarity between $\mathbf{V}_a$ and the total uniformity force $\mathbf{\Phi}_a$:

$$\zeta_a = \cos(\mathbf{V}_a, \mathbf{\Phi}_a) = \frac{\mathbf{V}_a \cdot \mathbf{\Phi}_a}{\|\mathbf{V}_a\| \|\mathbf{\Phi}_a\|} = \frac{\hat{\mathbf{V}}_a \cdot \mathbf{\Phi}_a}{\|\mathbf{\Phi}_a\|}. \tag{20}$$

Let $N = M - 1$. The total uniformity force is $\mathbf{\Phi}_a = \sum_{n=1}^{N} \mathbf{\Phi}_a^{(n)} = (\sum_{n=1}^{N} c_n)\hat{\mathbf{V}}_a + \sum_{n=1}^{N} \boldsymbol{\varepsilon}_n$. Let $S_c = \sum_{n=1}^{N} c_n$ and $\mathbf{S}_\varepsilon = \sum_{n=1}^{N} \boldsymbol{\varepsilon}_n$. Due to orthogonality, the numerator of ζ_a is $\hat{\mathbf{V}}_a \cdot \mathbf{\Phi}_a = S_c$ and the squared norm of the denominator is $\|\mathbf{\Phi}_a\|^2 = S_c^2 + \|\mathbf{S}_\varepsilon\|^2$. Substituting these back, we obtain a precise expression for ζ_a:

$$\zeta_a = \frac{S_c}{\sqrt{S_c^2 + \|\mathbf{S}_\varepsilon\|^2}} = \frac{1}{\sqrt{1 + \frac{\|\mathbf{S}_\varepsilon\|^2}{S_c^2}}}. \tag{21}$$

The proof now hinges on showing that the ratio $\frac{\|\mathbf{S}_\varepsilon\|^2}{S_c^2}$ converges to zero as $N \to \infty$. The denominator $S_c^2 = (\sum c_n)^2 \geq (Nc_0)^2$ grows at least quadratically. For the numerator, due to the independence and zero-mean properties of $\{\boldsymbol{\varepsilon}_n\}$, its expected value grows at most linearly:

$$\mathbb{E}[\|\mathbf{S}_\varepsilon\|^2] = \sum_{n=1}^{N} \mathbb{E}[\|\boldsymbol{\varepsilon}_n\|^2] \leq NC_\varepsilon, \tag{22}$$

for some constant $C_\varepsilon < \infty$ implied by the bounded covariance. The ratio of the expected numerator to the lower-bounded denominator is of the order $O(N)/O(N^2) = O(1/N)$, which converges to 0. This implies that the random variable $\frac{\|\mathbf{S}_\varepsilon\|^2}{S_c^2}$ converges to 0 in probability.

By the Continuous Mapping Theorem, ζ_a converges in probability to 1. As ζ_a is bounded in $[-1, 1]$, the Bounded Convergence Theorem ensures that convergence in probability to a constant implies convergence in expectation. Therefore:

$$\lim_{M\to\infty} \mathbb{E}[\zeta_a] = 1. \tag{23}$$

□

B. Proof of Proposition 2

Proposition 2 (Intra-alignment Conflict). The expected intra-alignment conflict, $\mathbb{E}[\chi_a]$, is governed by M and the average pairwise alignment $\bar{\mu} = \mathbb{E}[\mathbf{z}_i^{(m)\top}\mathbf{z}_i^{(n)}] \in [0, 1]$ for $m \neq n$ between modalities:

$$\mathbb{E}[\chi_a] \;\geq\; 1 - \sqrt{\frac{1 + (M-2)\bar{\mu}}{M-1}}. \tag{24}$$

For imperfect alignment ($\bar{\mu} < 1$), the conflict increases with the number of modalities M and admits a non-zero asymptotic lower bound:

$$\liminf_{M\to\infty} \mathbb{E}[\chi_a] \;\geq\; 1 - \sqrt{\bar{\mu}}. \tag{25}$$

Proof. The intra-alignment conflict for an anchor modality a is defined as:

$$\chi_a \triangleq 1 - \frac{\|\mathbf{V}_a\|_2}{\sum_{n\neq a} w_{an}/\tau}$$

where the alignment force is $\mathbf{V}_a = \sum_{n\neq a} \frac{w_{an}}{\tau}\mathbf{z}_i^{(n)}$. To derive the fundamental scaling relationship with the number of modalities M, we make a simplifying assumption of uniform weighting, i.e., $w_{an}/\tau = 1$ for all $n \neq a$. Under this assumption,

$$\chi_a = 1 - \frac{\|\mathbf{V}_a\|_2}{M-1}, \qquad \mathbf{V}_a = \sum_{n\neq a} \mathbf{z}_i^{(n)}. \tag{26}$$

Let $N = M - 1$. Then

$$\|\mathbf{V}_a\|_2^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} \mathbf{z}_i^{(n)\top}\mathbf{z}_i^{(m)} = N + \sum_{n\neq m} \mathbf{z}_i^{(n)\top}\mathbf{z}_i^{(m)}. \tag{27}$$

Taking expectations and using $\bar{\mu} = \mathbb{E}[\mathbf{z}_i^{(n)\top}\mathbf{z}_i^{(m)}]$ for $n \neq m$,

$$\mathbb{E}\big[\|\mathbf{V}_a\|_2^2\big] = N + N(N-1)\bar{\mu} = (M-1)\big(1 + (M-2)\bar{\mu}\big). \tag{28}$$

By Jensen's inequality,

$$\mathbb{E}\big[\|\mathbf{V}_a\|_2\big] \;\leq\; \sqrt{\mathbb{E}\big[\|\mathbf{V}_a\|_2^2\big]} = \sqrt{(M-1)\big(1 + (M-2)\bar{\mu}\big)}. \tag{29}$$

Hence

$$\mathbb{E}[\chi_a] = 1 - \frac{\mathbb{E}\|\mathbf{V}_a\|_2}{M-1} \;\geq\; 1 - \sqrt{\frac{1 + (M-2)\bar{\mu}}{M-1}}, \tag{30}$$

which proves (24). Finally,

$$\lim_{M\to\infty} \sqrt{\frac{1 + (M-2)\bar{\mu}}{M-1}} = \sqrt{\bar{\mu}} \quad\Rightarrow\quad \liminf_{M\to\infty} \mathbb{E}[\chi_a] \;\geq\; 1 - \sqrt{\bar{\mu}}. \tag{31}$$

□

C. Generalized Hölder Divergence

KDE estimation of the global Hölder divergence. Let $\{p_m(\mathbf{z})\}_{m=1}^M$ be the (unknown) continuous densities of M modalities and

$$\begin{aligned} D_{\text{Hölder}} &= -\log \frac{\int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z}}{\left(\prod_{m=1}^M \int p_m(\mathbf{z})^M\, d\mathbf{z}\right)^{1/M}} \\ &= \frac{1}{M}\sum_{m=1}^M \log\left(\int p_m(\mathbf{z})^M\, d\mathbf{z}\right) - \log\left(\int \prod_{m=1}^M p_m(\mathbf{z})\, d\mathbf{z}\right), \end{aligned} \tag{32}$$

which is nonnegative by Hölder's inequality and equals 0 iff the Hölder inequality is tight.

For modality m, let $\{\mathbf{z}_k^{(m)}\}_{k=1}^B$ be a batch of embeddings on $\mathbb{R}^d$ and define the kernel density estimator (KDE)

$$\widehat{p}_m(\mathbf{z}) = \frac{1}{B}\sum_{k=1}^B K_\tau\big(\mathbf{z}, \mathbf{z}_k^{(m)}\big), \qquad K_\tau(\mathbf{z}, \mathbf{z}') = \frac{1}{(2\pi\tau^2)^{d/2}} \exp\Big(-\tfrac{\|\mathbf{z}-\mathbf{z}'\|_2^2}{2\tau^2}\Big), \tag{33}$$

with bandwidth $\tau > 0$.[1]

Using $\int p_m^M = \mathbb{E}_{Z\sim p_m}\big[p_m(Z)^{M-1}\big]$ and $\int \prod_{m=1}^M p_m = \mathbb{E}_{Z\sim p_1}\big[\prod_{m=2}^M p_m(Z)\big]$, we obtain Monte-Carlo plug-in estimators by sampling from the empirical p_m via $\{\mathbf{z}_j^{(m)}\}$ and evaluating the KDEs:

$$\int \widehat{p}_m(\mathbf{z})^M\, d\mathbf{z} = \mathbb{E}_{Z\sim\widehat{p}_m}\big[\widehat{p}_m(Z)^{M-1}\big] \approx \frac{1}{B}\sum_{j=1}^B \left(\frac{1}{B}\sum_{k=1}^B K_\tau\big(\mathbf{z}_j^{(m)}, \mathbf{z}_k^{(m)}\big)\right)^{M-1}, \tag{34}$$

$$\int \prod_{m=1}^M \widehat{p}_m(\mathbf{z})\, d\mathbf{z} = \mathbb{E}_{Z\sim\widehat{p}_1}\Big[\prod_{m=2}^M \widehat{p}_m(Z)\Big] \approx \frac{1}{B}\sum_{j=1}^B \prod_{m=2}^M \left(\frac{1}{B}\sum_{k=1}^B K_\tau\big(\mathbf{z}_j^{(1)}, \mathbf{z}_k^{(m)}\big)\right). \tag{35}$$

Equivalently, with the unnormalized Gaussian kernel κ (dropping constants), the formulas above match

$$\int p_m^M \approx \frac{1}{B}\sum_{j=1}^B \Big(\frac{1}{B}\sum_{k=1}^B \kappa\big(\mathbf{z}_j^{(m)}, \mathbf{z}_k^{(m)}\big)\Big)^{M-1}, \qquad \int \prod_{m=1}^M p_m \approx \frac{1}{B}\sum_{j=1}^B \prod_{m=2}^M \Big(\frac{1}{B}\sum_{k=1}^B \kappa\big(\mathbf{z}_j^{(1)}, \mathbf{z}_k^{(m)}\big)\Big). \tag{36}$$

Define the (within-modality) kernel means $s_i^{(m)} \triangleq \frac{1}{B}\sum_{k=1}^B K_\tau(\mathbf{z}_i^{(m)}, \mathbf{z}_k^{(m)})$ and the (cross-modality-to-anchor) kernel means $c_i \triangleq \prod_{m=2}^M \frac{1}{B}\sum_{k=1}^B K_\tau(\mathbf{z}_i^{(1)}, \mathbf{z}_k^{(m)})$. Then the Hölder divergence estimator (up to an additive constant if using κ) is

$$\widehat{D}_{\text{Hölder}} = \frac{1}{M}\sum_{m=1}^M \log\left(\frac{1}{B}\sum_{i=1}^B \big(s_i^{(m)}\big)^{M-1}\right) - \log\left(\frac{1}{B}\sum_{i=1}^B c_i\right). \tag{37}$$

All terms are differentiable; the computation costs $O(MB^2)$ and can be vectorized via kernel matrices $K_{ij}^{(m)} = K_\tau(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(m)})$ and $K_{ij}^{(m\to 1)} = K_\tau(\mathbf{z}_i^{(1)}, \mathbf{z}_j^{(m)})$.

D. Design Space for Uniformity and Alignment

We illustrate some possible designs following our principle in Table 4, showing the generality and flexibility of our framework.

[1] If one uses the *unnormalized* kernel $\kappa(\mathbf{z}, \mathbf{z}') = \exp(-\|\mathbf{z}-\mathbf{z}'\|_2^2/(2\tau^2))$, then $\widehat{p}_m$ is scaled by a constant depending on (d, τ). This yields an *additive constant* in $D_{\text{Hölder}}$ that does not affect optimization; we drop such constants in practice.

Table 4. **Design space for uniformity and alignment**. Uniformity can be instantiated in Euclidean or manifold geometries; alignment can incorporate geometric constraints beyond pairwise distance.

Principle	Space	Kernel/Metric	Notes
Uniformity (repulsion)	Euclidean ($\mathbb{R}^d$)	$\exp\Big(-\|\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)}\|_2^2/2\tau^2\Big)$	Gaussian kernel in $\mathbb{R}^d$; encourages spread.
	Unit Hypersphere ($\mathbb{S}^{d-1}$)	$\exp\Big(-d_{\mathbb{S}}(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(m)})^2/2\tau^2\Big)$	Geodesic (Riemannian) Gaussian on $\mathbb{S}^{d-1}$.
Alignment (attraction)	Euclidean ($\mathbb{R}^d$)	$\|\mathbf{z}_i^{(m)} - \mathbf{z}_i^{(n)}\|_2^2$	Pairwise matching per sample.
	Unit Hypersphere ($\mathbb{S}^{d-1}$)	$\big[d_{\mathbb{S}}(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)})\big]^2$	Geodesic pairwise alignment.
	Area/volume preservation	$\sqrt{\det \mathbf{G}(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}}$	Penalizes global volume; $\mathbf{G}$ is the Gram Matrix.

E. Computational Complexity

Our framework is formulated upon the core principles of feature alignment and uniformity, which are fundamental to the contrastive learning objective. The implementation of our method operates within the standard computational pipeline of modern contrastive learning. For a given batch of size B with d-dimensional representations, the dominant computational cost remains the construction of the $B \times B$ pairwise similarity matrix, an operation with $O(B^2 d)$ time complexity. The objectives of alignment and uniformity are then computed based on this matrix, inheriting the same computational profile as the loss calculation and gradient backpropagation stages of standard contrastive methods.

Crucially, our formulation does not require any operations beyond those already present in the baseline. As such, our method introduces no additional computational overhead and shares an identical time and memory complexity profile with widely-used InfoNCE-based frameworks. This efficiency ensures our approach is scalable and readily applicable to large-scale training regimes.

F. Experimental Details and Additional Experiments

F.1. Implementation Details

All experiments use 4× NVIDIA A6000 GPUs. We train with AdamW using a learning rate of $\mathbf{2 \times 10^{-5}}$ and a batch size of 128 per GPU (global batch = 512). For zero-shot video retrieval, we use the GRAM (Cicchetti et al., 2025) codebase and only replace GRAM's volume-based InfoNCE with our uniformity and alignment losses. We follow their settings: each video clip is sampled with 8 frames during training, and the model is trained for 5 epochs. For cross-modal generation, we train on VGGSound for 50 epochs. For hyperparameters, we use $\lambda_{\text{align}} = 1$ and set the temperature of uniformity loss as $\tau_{\text{ctr}} = 0.07$, the same as in standard CLIP. We use a separate temperature $\tau_{\text{ctr}} = 0.07$ for tuple-level uniformity $U(\mathbf{C})$.

F.2. Empirical evidence of InfoNCE conflict

To further support the theoretical analysis of the InfoNCE conflict and to motivate our method, we provide an empirical observation from training dynamics. Specifically, we train a lightweight 5-layer Transformer adapter on top of the CLIP ViT-L, optimized with the standard InfoNCE objective. Following the ImageBind-style setup, we treat the CLIP image embeddings as fixed targets and learn to bind text embeddings to the image embedding space. We train on the MSCOCO training set and visualize the learned representations using t-SNE on 5K image-text pairs from the validation split, saved every 50 epochs. As shown in Fig. 4, the two modalities become closer at early stages of training, but the clusters later separate again. While t-SNE is qualitative, this phenomenon suggests that InfoNCE may not consistently reduce the distributional mismatch between modalities and can even widen the modality gap at later stages. This observation is consistent with the analysis in Yin et al. (2025), which links weakened alignment to the evolution of the (learnable) temperature during training.

F.3. More Ablation Studies

Sensitivity of temperature in $U(\mathbf{C})$. As we mentioned above, we use the temperature $\tau = 0.07$ for the uniformity loss. However, a separate temperature τ_{ctr} for the tuple-level uniformity loss $U(\mathbf{C})$ could control the global separability. Hence,

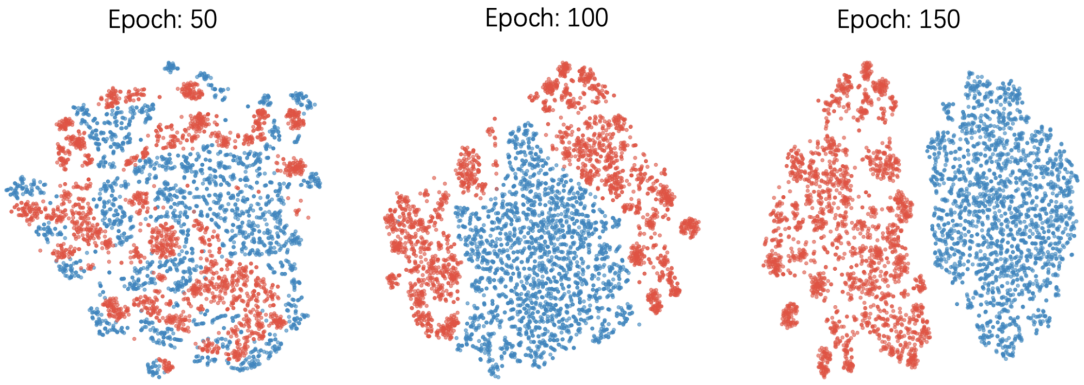


Figure 4. **t-SNE visualization of image and text embeddings across training epochs when optimizing an InfoNCE-based text-to-image binding.** The two modalities move closer at early stages but later separate again, suggesting inconsistent distributional alignment.

Table 5. Ablation on centroid uniformity temperature τ_{ctr} on MSR-VTT retrieval (Recall@1, %).

τ_{ctr}	T2V R@1	V2T R@1	Avg R@1
0.01	56.0	54.4	55.2
0.03	56.8	53.2	55.0
0.07	57.4	53.2	55.3

we perform an ablation study on this parameter. Table 5 shows that the average performance is robust to the temperature τ, while controlling centroid τ may affect the subtask performance (T2V and V2T).

Sensitivity to λ_{align} and from-scratch training. We study the sensitivity of the alignment weight λ_{align} on MSR-VTT using our basic instantiation, which combines modality-wise uniformity with an explicit alignment regularizer:

$$\mathcal{L} = \sum_{m=1}^{M} U\big(Z^{(m)}\big) + \lambda_{\text{align}}\, L_{\text{align}}. \tag{38}$$

In this ablation, we keep all other training settings fixed and vary only λ_{align} to isolate its effect. In addition, we conduct this experiment with a small-scale training-from-scratch setup, demonstrating that our objective remains effective without large-scale pretraining. For evaluation, we perform cross-modal retrieval using cosine similarity between normalized embeddings and report standard MSR-VTT retrieval metrics. Overall, the results show that the proposed objective is stable across a broad range of λ_{align} and exhibits the expected trade-off: increasing λ_{align} strengthens cross-modal alignment, while overly large values can reduce within-modality uniformity and degrade retrieval performance.

Table 6. **Sensitivity of λ_{align}.**

λ_{align}	T2V	V2T	Avg
0.2	23.8	26.8	25.3
0.5	26.2	29.4	27.8
1.0	23.1	28.6	25.85
1.5	19.6	19.3	19.45
2.0	13.1	13.6	13.35

$D_{\text{Hölder}}$ curve during training. To demonstrate that our principle serves as an efficient proxy for global distribution estimation, we additionally track the evolution of the cross-modal divergence $D_{\text{Hölder}}$ on MSR-VTT throughout training. Concretely, we train the model under the same setting as in the cross-modal generation experiments, and every 5 epochs we compute $D_{\text{Hölder}}$ between the image, text, and audio embedding distributions using the KDE estimator in Eq. (37). Fig. 5 shows that the global Hölder divergence decreases rapidly during training as our method effectively closes the distribution gaps among the three modalities. Notably, $D_{\text{Hölder}}$ quickly approaches zero, suggesting that directly optimizing

this divergence can be problematic in practice (e.g., due to early saturation and weak gradients), motivating our proxy objective instead.

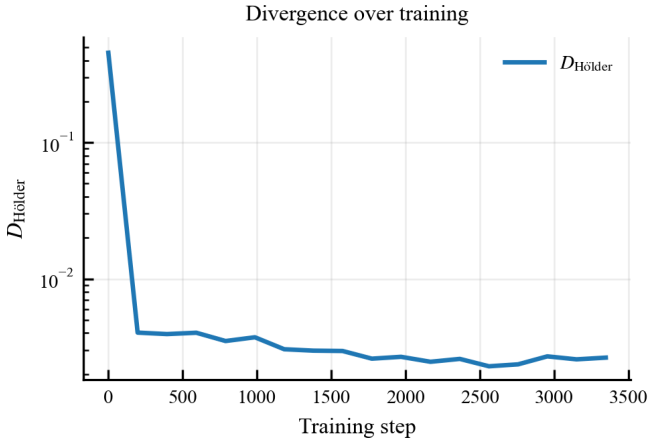


Figure 5. $D_{\text{Hölder}}$ **curve during training.**

F.4. Experimental Results on Modality Interpolation

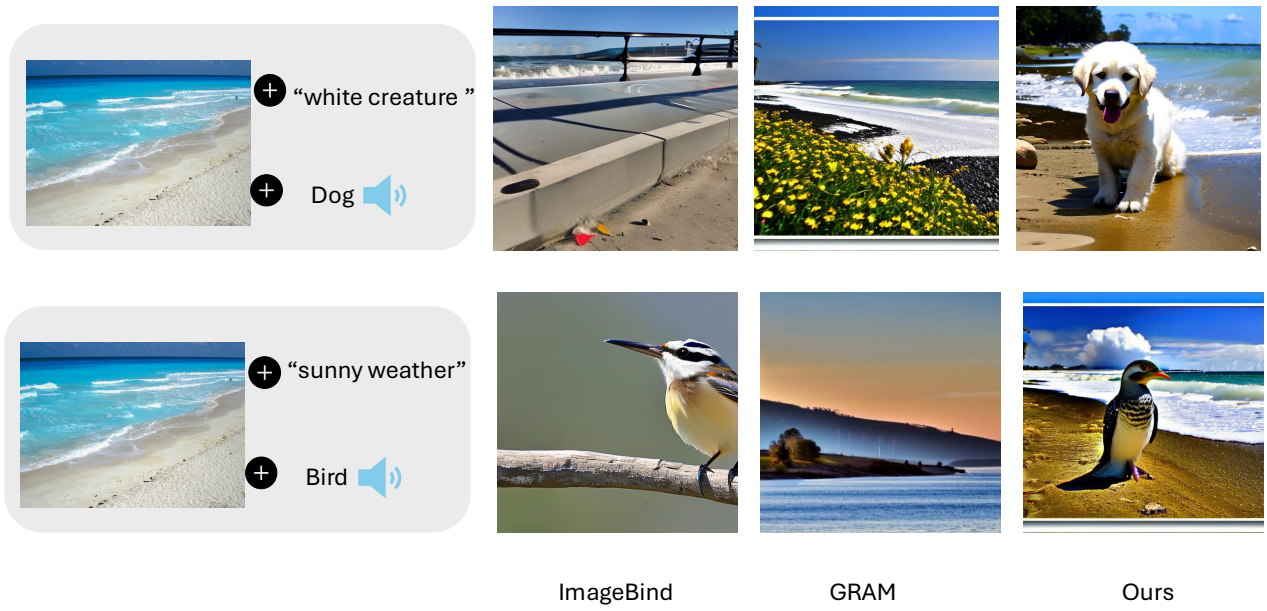


Figure 6. **Modality-interpolation generation results (V+T+A) → I.** When interpolating, vision, text, and audio representations, our method has a better ability to fuse the semantic information across modalities, leading to better generation.

Beyond the bimodal cases in Fig. 3, we present tri-modal interpolation results. Conditioning jointly on an image embedding, a text prompt, and an audio embedding, our model synthesizes images that integrate complementary semantics from all three modalities, demonstrating effective cross-modal fusion.

F.5. Generation Results of VGGSound

We present more generated samples from VGGSound in Fig. 7. Note that the image quality of VGGSound's videos is quite noisy, making the generation results similar. Also, we adopt a raw generation process for this demonstration, where the embeddings are directly passed to the decoder without additional conditions (e.g., negative prompts or quality-enhancing

constraints). This can directly reflect the goodness of the multimodal alignment without external factors.

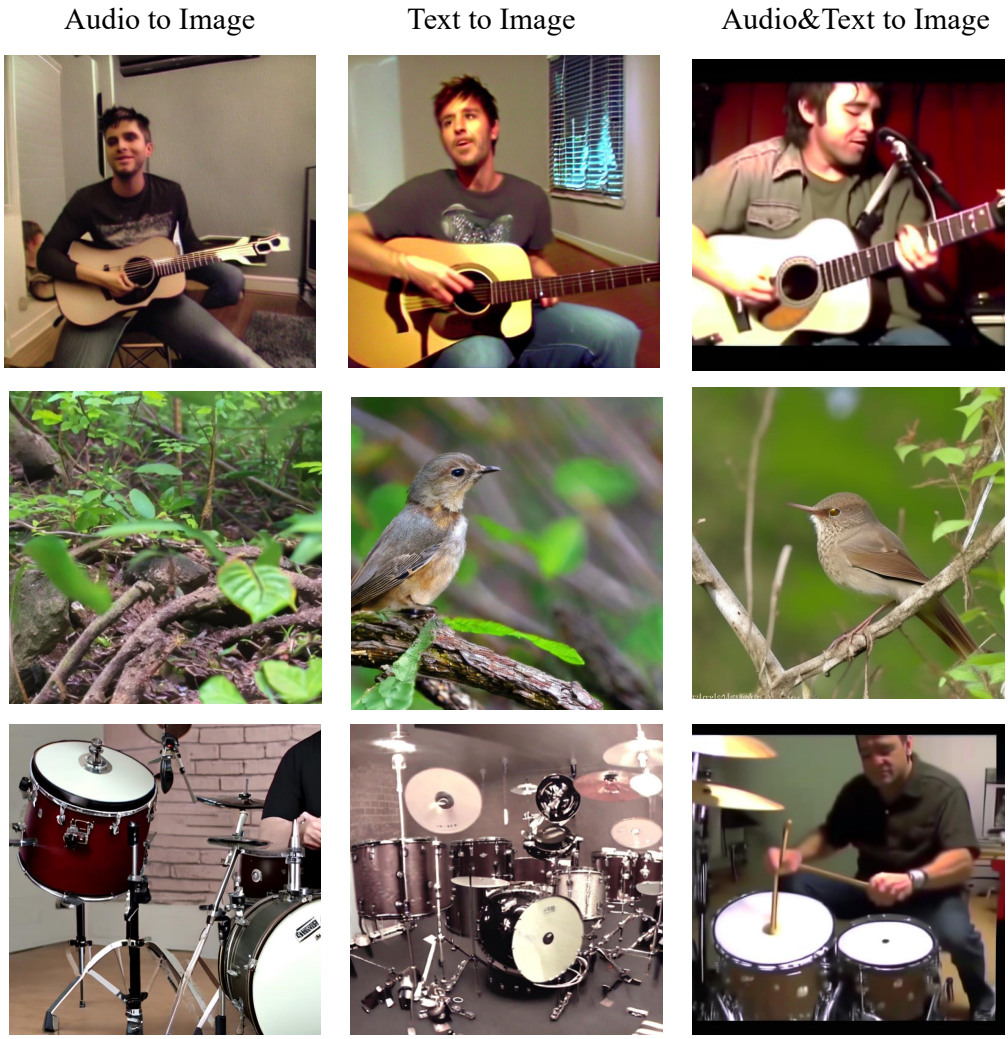


Figure 7. **More generated results from VGGsound.** We adopt a raw generation process for demonstrating the multimodal alignment ability, where the embeddings are directly passed to the decoder without additional conditions (e.g., negative prompts or quality-enhancing constraints)