TEXTME: TEXT-ONLY TRAINING FOR MODALITY EXPANSION VIA LLM SPACE PIVOTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expanding multimodal representations to novel modalities is constrained by reliance on large-scale paired datasets (e.g., text–image, text–audio, text–3D, text–molecule), which are costly and often infeasible in domains requiring expert annotation such as medical imaging, 3D modeling, and molecular analysis. We introduce TextME, the first framework for text-only modality expansion that removes paired data requirements. Our method leverages the universal geometric properties of pre-trained encoders—consistent modality gaps—which enable zero-shot cross-modal transfer once embedding spaces satisfy these properties. We empirically verify that these hold across audio, 3D, X-ray, and molecular domains, enabling effective cross-modal tasks without paired supervision. Furthermore, we evaluated LLM and multimodal text encoders to determine which is more effective as a unified anchor space. Experiments show that TextME achieves 88.2% of paired-data performance in zero-shot classification and cross-modal retrieval, while also supporting emergent capabilities between unseen modality pairs (e.g., audio-to-3D, molecule-to-image). These results highlight text-only modality expansion as a practical and scalable path toward foundation models spanning arbitrary modalities.

1 INTRODUCTION

Expanding multimodal representations to novel modalities constitutes a fundamental challenge in contemporary representation learning (Baltrušaitis et al., 2018; Manzoor et al., 2023; Liang et al., 2024; Yuan et al., 2025; Liu et al., 2025). Modality expansion aims to align heterogeneous data modalities into a unified embedding space where semantically equivalent content maintains proximity (Zhang et al., 2023a; Han et al., 2023; Zhu et al., 2023; Lyu et al., 2024; Wang et al., 2023a). Large-scale paired datasets such as text–image or text–audio corpora have enabled remarkable progress in vision–language (Radford et al., 2021; Jia et al., 2021) and audio–language (Wu et al., 2023; Manco et al., 2022) modeling, but the construction of such resources proves prohibitively expensive or infeasible. Medical imaging requires costly expert annotations while navigating privacy constraints (Wang et al., 2025; Kitamura et al., 2024; Ziller et al., 2021), molecular analysis demands complex domain-specific representations (Edwards et al., 2024; Xiao et al., 2024), and 3D modeling necessitates labor-intensive curation (Deitke et al., 2023; Sarkar et al., 2025). Consequently, the scalability of modality expansion is constrained not merely by architectural limitations but, more fundamentally, by the availability of paired supervision.

Recent advances (Wang et al., 2023b; Zhang et al., 2024b; Wang et al., 2024a;b) demonstrate that pre-trained multimodal encoders like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) can be effectively reused through lightweight projection networks to integrate multiple modalities into shared representation spaces. However, these approaches still require fully-paired multimodal data during training, demanding simultaneous access to all target modalities with corresponding supervision. This requirement becomes particularly challenging when extending to modalities beyond standard vision-language pairs—such as audio, 3D point clouds, medical X-rays, and molecular structures—where natural correspondences are often absent and domain expertise is scarce.

In this work, we eliminate the paired data requirement by exploiting an inherent geometric property of pre-trained multimodal encoders—the consistent modality gap. Inspired by prior theoretical work demonstrating this phenomenon (Liang et al., 2022; Zhang et al., 2023b), we propose Text-anchored

Modality Expansion (TextME), a framework that leverages this gap for text-only training. Zhang et al. (2024a) demonstrated that when a directionally consistent *offset* vector exists between pre-trained image and text embedding spaces, cross-modal transfer can be achieved via simple vector translation without paired training. As illustrated in Figure 1, we extend this insight by empirically verifying that such modality gaps are a universal property of contrastively-trained encoders, regardless of the specific modality they encode. Since these encoders rely on text for alignment during training, TextME exploits their shared text embedding space as a semantic anchor, applying pre-computed *offset* translation to bridge modality spaces using only text descriptions.

We validate TextME's effectiveness through comprehensive experiments on diverse modalities using zero-shot classification and cross-modal retrieval tasks. Despite training exclusively on text descriptions, TextME achieves an average of 88.2% performance preservation compared to paired-data methods, with specific tasks like molecular retrieval, even surpassing supervised learning baselines. Moreover, our framework enables emergent cross-modal capabilities between modality pairs that have never seen during training, such as audio-to-3D and molecule-to-image retrieval, demonstrating that text-anchored alignment creates meaningful semantic bridges across arbitrary modalities.

In addition, we empirically evaluate two candidate text representation spaces as a semantic anchor:LLM-based embeddings and multimodal text encoders. Experimental results reveal a consistent trend: LLM-anchored embeddings deliver stronger performance on retrieval tasks, while multimodal-anchored embeddings excel in classification. We attribute this distinction to their training paradigms—LLMs learn semantically rich representations well-suited for aligning natural language queries, whereas multimodal encoders trained under contrastive objectives emphasize discriminative boundaries advantageous for categorical separation.

Our contribution is three-fold:

- We provide comprehensive empirical validation that the consistent modality gap—a systematic offset between text and non-text embeddings—exists universally across diverse pre-trained encoders (e.g., audio, 3D, X-ray, and molecule). We demonstrate that this gap operates orthogonally to semantic content, enabling zero-shot cross-modal transfer without requiring paired multimodal data.
- We propose TextME, the first framework that exploits this geometric consistency to achieve modality expansion using only text descriptions. By leveraging LLM embeddings as a unified semantic anchor, our method captures richer semantic relationships across diverse domains.
- We demonstrate that text-only training can achieve 88.2% of paired-data performance across diverse modalities (i.e., audio, 3D, X-ray, molecule) while eliminating the need for target modality data during training. TextME enables emergent cross-modal retrieval capabilities between modality pairs that were never seen during training (e.g., audio-to-3D, molecule-to-image retrieval), demonstrating that text serves as an effective semantic bridge across arbitrary modalities.

2 Theoretical Foundation: Cross-Modal Instance Mapping

We establish the theoretical underpinnings for text-only modality expansion by analyzing the geometric structure of pre-trained multimodal encoders. Our investigation reveals that contrastively-trained encoders exhibit a consistent modality gap—a systematic offset between text and non-text embeddings—enabling zero-shot cross-modal transfer through simple offset translation.

2.1 Geometric Properties of Cross-Modal Alignment

Building on observations of vision-language models (Liang et al., 2022; Zhang et al., 2023b; 2024a), we extend the theoretical analysis to diverse specialized modalities, including audio, 3D, medical imaging, and molecular structures. We identify three critical hypotheses that support offset-based alignment: intra-modal clustering (Hypothesis 0), inter-modal gap consistency (Hypothesis 1), and orthogonality between gap and content variations (Hypothesis 2).

Definition 1 (Cross-Modal Instance Mapping)**.** *Given a set of modalities $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ with embeddings in a shared space $\mathbb{R}^d$, a cross-modal instance mapping is a transformation Φ_{ij} :*

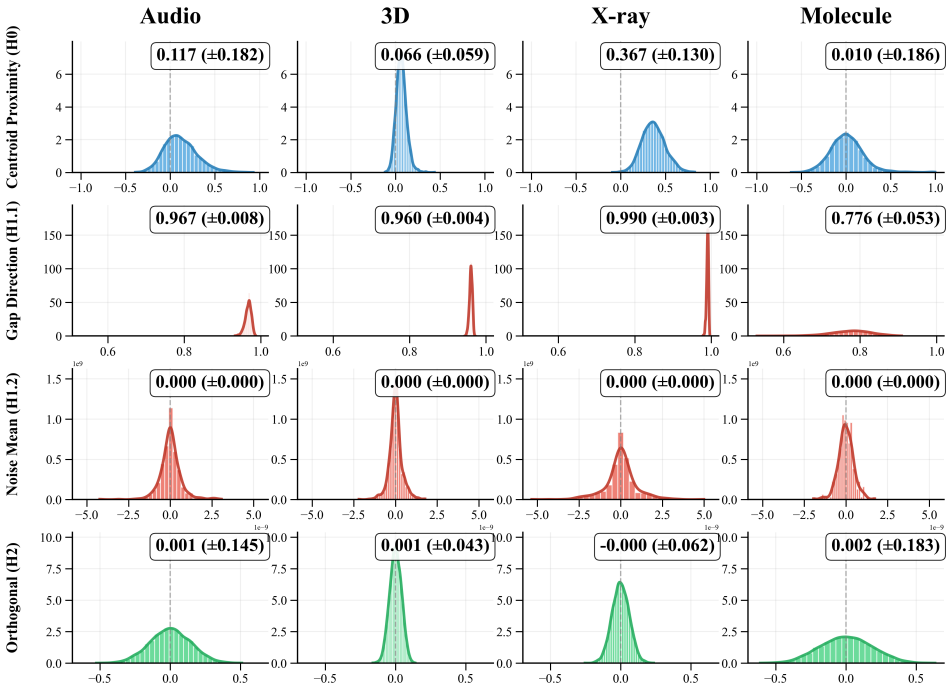

Figure 1: Geometric analysis of modality gaps across five multimodal encoders. Centroid Proximity (top): Mean pairwise distances from embeddings to modality centroids. Gap Direction (second): Cosine similarity between different sample pairs' gap vectors. Noise Mean (third): Distribution of alignment noise mean values. Gap Orthogonality (bottom): Cosine similarity between gap vectors and content variations within modalities.

$\mathbb{R}^d \to \mathbb{R}^d$ *that aligns embeddings from modality* m_i *to modality* m_j*:*

$$\Phi_{ij}(e) = e - \Delta_{ij} \tag{1}$$

where $\Delta_{ij} = \mu_{m_i} - \mu_{m_j}$ *is the inter-modal offset between modality centroids* $\mu_{m_i} = \mathbb{E}[e_{m_i}]$ *and* $\mu_{m_j} = \mathbb{E}[e_{m_j}]$*. The mapping enables zero-shot cross-modal transfer when Hypotheses 0–2 are satisfied, ensuring that semantically corresponding embeddings* e_i, e_j *satisfy* $\|\Phi_{ij}(e_i) - e_j\| < \delta$ *for small* δ*.*

Hypothesis 0 (Intra-Modal Alignment Independence)**.** *For a modality* $m \in \mathcal{M}$*, normalized embeddings concentrate within a bounded region on the unit hypersphere:* $\cos(\hat{e}_i, \hat{e}_j) > \tau_{intra}$ *for all embeddings* e_i, e_j *from modality* m*, where* $\hat{e} = e/\|e\|_2$ *denotes normalization and* τ_{intra} *is determined by the contrastive objective.*

This property emerges from the ℓ_2 normalization applied during contrastive learning, which projects representations onto the unit hypersphere. We verify this through Centroid Proximity statistics, measuring the mean pairwise cosine similarity of normalized embeddings within each modality. This concentration property establishes a well-defined centroid $\mu_m = \mathbb{E}[e_m]$ for each modality, enabling the characterization of inter-modal offsets $\Delta_{ij} = \mu_i - \mu_j$ as meaningful geometric transformations.

Hypothesis 1 (Inter-Modal Gap Consistency)**.** *For modalities* $m_i, m_j \in \mathcal{M}$*, a consistent offset exists between their embedding spaces:*

Hypothesis 1.1 (Space-level Gap Consistency)**.** *Instance-level offsets can be approximated by a single group-level offset:* $\Delta_{ij}^{(k)} \approx \Delta_{ij}$ *for all instance pairs* k*, where* $\Delta_{ij} = \mu_i - \mu_j$ *is the difference between modality centroids.*

The modality gap originates from inherent differences in modality characteristics and architectural properties from initializing separate encoders. We verify this through Gap Direction analysis, measuring an average cosine similarity between instance-level and group-level offsets. Our experiments show $\cos(\Delta_{ij}^{(k)}, \Delta_{ij}) > 0.95$ across all instance pairs, demonstrating strong alignment with the mean gap direction.

Hypothesis 1.2 (Instance-level Gap Consistency). *Deviations from the mean offset follow a bounded distribution:* $\epsilon_k = \Delta_{ij}^{(k)} - \Delta_{ij} \sim \mathcal{N}(0, \sigma^2)$ *where* $\sigma < \gamma \cdot \tau$*, with* τ *being the temperature parameter and* γ *a modality-specific constant.*

The robustness of the offset is related to the temperature parameter in the contrastive objective. We verify this through Alignment Noise analysis, confirming that $\mathbb{E}[\epsilon_k] \approx 0$ with bounded variance, demonstrating robustly predictable instance-level variations.

Hypothesis 2 (Orthogonality of Inter/Intra Variations). *The inter-modal offset* Δ_{ij} *is orthogonal to intra-modal semantic variations:* $\Delta_{ij} \perp r_m^{(p,q)}$ *for any instances* p, q *within modality* $m \in \{m_i, m_j\}$*, where* $r_m^{(p,q)} = e_p - e_q$ *denotes the difference vector between embeddings.*

This orthogonality indicates that Δ_{ij} operates independently of semantic content within each modality. Since the offset is orthogonal to semantic variations $r_m^{(p,q)}$, applying the mapping $\Phi_{ij}(e) = e - \Delta_{ij}$ preserves relative distances and semantic relationships. We verify this through Gap Orthogonality analysis, measuring $|\cos(\Delta_{ij}, r_m^{(p,q)})| < 0.1$ for random within-modality pairs.

2.2 Empirical Modality Gap Validation Across Diverse Modalities

To validate our theoretical framework, we analyzed five pre-trained multimodal encoders spanning diverse modalities: LanguageBind (Zhu et al., 2023) for vision, CLAP (Elizalde et al., 2023) for audio, Uni3D (Zhou et al., 2023) for 3D point clouds, CXR-CLIP (You et al., 2023) for medical X-rays, and MoleculeSTM (Liu et al., 2023) for molecular structures. For each encoder, we randomly sampled $N = 5,000$ text-modal pairs from their training domains to compute the inter-modal offset $\Delta_{ij} = \mathbb{E}[\mathcal{E}_{m_i}(x)] - \mathbb{E}[\mathcal{E}_{m_j}(t)]$ and analyze its geometric properties. Figure 1 presents comprehensive validation results demonstrating that all three hypotheses hold across diverse modalities.

Centroid Proximity (Hypothesis 0) confirms tight intra-modal clustering across most modalities, though X-ray shows notably dispersed distributions that correlate with reduced modality expansion performance as demonstrated in Section 4.2. **Gap Direction** demonstrates $\cos(\Delta_{ij}^{(k)}, \Delta_{ij}) > 0.96$ consistency (Hypothesis 1.1), validating single-vector characterization across modalities. **Noise Mean** confirms zero-centered distributions with $\mathbb{E}[\epsilon_k] \approx 0$ (Hypothesis 1.2), indicating predictable alignment variations. **Gap Orthogonality** shows $|\cos(\Delta_{ij}, r_m^{(p,q)})| < 0.05$ (Hypothesis 2), confirming that modality gaps operate independently of semantic content, enabling effective cross-modal transfer through simple offset operations. These geometric properties establish a unified framework for understanding and exploiting cross-modal relationships in pre-trained encoders.

3 Text-anchored Modality Expansion Framework

Building on the theoretical insights from Section 2, we propose TextME, a framework that exploits the consistent modality gap property to enable text-only training for modality expansion. Our approach leverages the geometric consistency demonstrated in Section 2.1—that pre-trained encoders exhibit a constant offset between text and non-text embeddings orthogonal to semantic content. This property allows us to create an interchangeable coordinate system through simple centering operations, eliminating the need for paired multimodal data.

3.1 Problem Formulation

Modality expansion aims to integrate diverse pre-trained encoders into a unified semantic space where similar concepts maintain proximity regardless of their originating modality. Consider a set of pre-trained encoders $\{E_m : \mathcal{X}_m \to \mathbb{R}^{d_m}\}$, where each encoder E_m maps inputs from modality m's input space $\mathcal{X}_m$ to d_m-dimensional embeddings. Given a source modality m_s with established semantic representations and target modalities $\mathcal{M}_T = \{m_1, \ldots, m_k\}$ to be incorporated, the objective is to learn projection networks $P_m : \mathbb{R}^{d_m} \to \mathbb{R}^{d_h}$ that preserve semantic relationships across modalities, where d_h denotes the dimensionality of the shared embedding space.

We consider a practical scenario where only unpaired textual descriptions $\mathcal{D}_{\text{text}} = \{t_i\}_{i=1}^N$ are available for training. Pre-trained multimodal models consist of text encoders E_m^{text} and modal encoders

E_m^{modal} jointly optimized through contrastive learning. Our approach exploits the geometric properties identified in Section 2.1—specifically, the consistent offset between these encoders—to enable alignment without paired multimodal data.

3.2 Framework Overview

TextME operates through three stages that decouple geometric alignment from semantic projection. First, we pre-compute modality-specific offsets $\Delta_m = \mu_m^{\text{modal}} - \mu_m^{\text{text}}$ using Equation 1 to characterize the geometric transformation between text and modal encoders. Second, we train lightweight projection networks P_m exclusively on centered text embeddings, mapping them to a shared representation space using only unpaired text descriptions. Third, at inference, we apply the pre-computed offset to non-text modality embeddings, then project them using the text-trained network. This design exploits the orthogonality property (Hypothesis 1) to preserve semantic relationships while enabling cross-modal transfer without paired supervision. Algorithm 1 formalizes the complete procedure.

3.2.1 Offset Computation

We establish interchangeability between text and modal embedding spaces by pre-computing modality-specific offsets. For each encoder E_m, we compute centroids $\mu_m^{\text{text}} = \mathbb{E}[E_m^{\text{text}}(t)]$ and $\mu_m^{\text{modal}} = \mathbb{E}[E_m^{\text{modal}}(x)]$ over representative samples from each distribution. By centering each modality independently—subtracting μ_m^{text} from text embeddings and μ_m^{modal} from modal embeddings—we create a shared coordinate system where both modalities are aligned at the origin. This enables projection networks trained on centered text embeddings to generalize to centered modal embeddings at inference. The offset computation requires only 5,000 samples—a 99% reduction from typical paired training requirements (Zhu et al., 2023; Zhang et al., 2024b).

3.2.2 Text-to-Text Alignment

Contrastive Learning for Projection Network. Given text descriptions $\mathcal{D}_{\text{text}} = \{t_i\}_{i=1}^N$ from the target modality domain, we train a lightweight projection networks $P_m : \mathbb{R}^{d_m} \to \mathbb{R}^{d_h}$ to map centered text embeddings into a shared representation space. Each projection network consists of a 2-layer MLP with GeLU activation, requiring only ∼10M parameters. The training objective employs contrastive learning with hard negative mining (Lee et al., 2024; Moreira et al., 2024; Rösch et al., 2024):

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(z_i, z_i')/\tau)}{\sum_{j\in\mathcal{N}_i\cup\{i\}} \exp(\text{sim}(z_i, z_j')/\tau)} \tag{2}$$

where $z_i = P_m(E_m^{\text{text}}(t_i) - \mu_m^{\text{text}})$ is the projected centered embedding, $z_i' = E_s(t_i)$ is the shared space embedding, and $\mathcal{N}_i$ contains hard negatives with similarity scores in $[0.1 \cdot \text{sim}(z_i, z_i'), 0.9 \cdot \text{sim}(z_i, z_i')]$. This sampling strategy accelerates convergence by focusing gradients on informative examples near the decision boundary.

Choice of Shared Anchor Space. We empirically validate two candidate text representation spaces as semantic anchors: LLM embeddings (i.e., NV-Embed-v2 (Lee et al., 2024) and Qwen3-Embeddings (Zhang et al., 2025)) and multimodal text encoders (i.e., LanguageBind (Zhu et al., 2023)). To assess whether the embeddings faithfully capture semantic representations, we evaluated their performance on the STS benchmark, a widely used metric for contextual understanding that measures sentence similarity on a 0–5 scale. LLM embeddings achieve 85.79 ∼ 90.40 Spearman correlation on STS benchmarks versus 68.29 ∼ 68.83 for multimodal encoders (Table 3 in Appendix C), reflecting their superior semantic

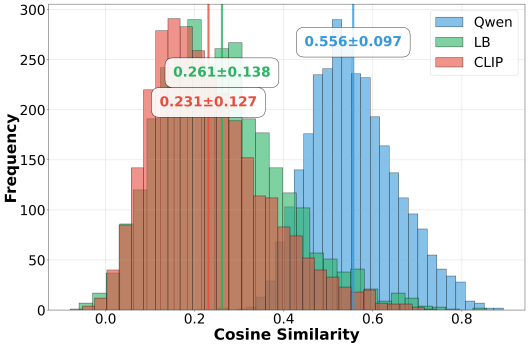

Figure 2: Semantic anchoring comparison between LLM embeddings and multimodal encoders on 3,000 semantically equivalent cross-modal description pairs.

understanding from extensive next-token prediction training. To assess cross-domain alignment capabilities, we analyzed $3,000$ audio-image caption pairs from FlickrNet (Senocak et al., 2018), where we generated linguistically distinct but semantically equivalent descriptions using the Gemini API (Google, 2024). For instance, an image caption "a red sports car speeding on highway" was paired with its audio equivalent "loud engine roar with wind rushing past", testing whether encoders can recognize semantic similarity despite different surface forms. Figure 2 shows LLM embeddings (Qwen) maintain 2.4× higher similarity (0.56 vs 0.23-0.26) for matched pairs compared to multimodal encoders, demonstrating their effectiveness as semantic anchors for text-only training despite lacking cross-modal supervision.

3.2.3 Inference-time Cross-Modal Transfer

At inference, TextME enables zero-shot cross-modal capabilities through offset-based transformation. For a non-text input x from modality m, we compute:

$$e_{\text{final}} = P_m(E_m^{\text{modal}}(x) - \mu_m^{\text{modal}}) \tag{3}$$

The centering operation $(E_m^{\text{modal}}(x) - \mu_m^{\text{modal}})$ transforms the modal embedding into the coordinate system used during text training. Since the offset is orthogonal to semantic variations (Hypothesis 2), this transformation preserves semantic relationships while enabling the text-trained projection network P_m to map modal embeddings to the shared representation space, achieving effective cross-modal retrieval and classification without paired supervision.

4 Experiments

We evaluate TextME's ability to expand multimodal representations through comprehensive experiments across diverse modalities. Our analysis includes quantitative evaluation on the standard benchmarks for cross-modal retrieval and zero-shot classification (Section 4.2), and qualitative examination of emergent cross-modal capabilities between modality pairs never paired during training (Section 4.3).

4.1 Experimental Setup

Source and Target Modalities. We conducted experiments to verify the modality expansion capability of TextME. For the source representation space, we select LanguageBind (Zhu et al., 2023), an image-text aligned standard multimodal foundation model. As target modalities, we integrate four specialized domains that lack natural multimodal correspondences: **Audio** using CLAP (Elizalde et al., 2023) trained on AudioCaps (Kim et al., 2019) descriptions, **3D** using Uni3D (Zhou et al., 2023) trained on Cap3D-Objaverse (Luo et al., 2023) captions, **X-ray** using CXR-CLIP (You et al., 2023) trained on CheXpert (Irvin et al., 2019) reports, and **Molecule** using MoleculeSTM (Liu et al., 2023) trained on PubChem (Kim et al., 2025) descriptions. For fair comparisons across modalities, we sample $100K$ text descriptions from each modality-specific training dataset.

Text Anchor Space. To establish a unified text embedding space, we utilize three distinct models: LanguageBind (LB; Zhu et al. 2023), NV-Embed-v2 (NV; Lee et al. 2024), and Qwen3-Embedding-4B (Qwen; Zhang et al. 2025). These models are carefully selected to evaluate the efficacy of our proposed framework across diverse shared anchor spaces, each exhibiting different representational capabilities as Section 3.2.2. For clarity, we denote the corresponding implementations as **Ours**$_{LB}$, **Ours**$_{NV}$, and **Ours**$_{Qwen}$.

Baselines. We compare against two categories of methods to evaluate TextME's effectiveness. *Paired*-data methods include LanguageBind (Zhu et al., 2023), which trains modality-specific encoders from scratch, and Ex-MCR (Zhang et al., 2024b), which adapts frozen pre-trained encoders. For direct comparisons, we implement **Ours**$_{upper\text{-}bound}$, a variant of TextME with the same architecture but trained on paired multimodal data, representing the performance upper bound. *Unpaired*-data methods lack established baselines for our zero-shot setting. We therefore compare our approach with COX (Huang et al., 2025), which fine-tunes target modalities without instance-level pairing, although it requires labeled target data, unlike our approach. We re-implemented COX following the specification of the paper; details are given in Appendix E.2. We also include a Naïve

Table 1: Zero-shot cross-modal retrieval performance. Highlighted rows share identical architecture but differ only in training data type (paired multimodal vs. text-only) and LLM anchoring. *Avg. Preservation* represents the average percentage of the supervised upper bound (**Ours**$_{upper\text{-}bound}$) achieved by each TextMEvariant, computed across R@1 and R@5 metrics. † indicates our reproduction due to unavailable public code. **Bold** indicates best among unsupervised methods.

	Text → Audio				Text → Molecule		Audio → Image	
	AudioCaps		Clotho		DrugBank		FlickrNet	
Method	R@1	R@5	R@1	R@5	MRR@4	MRR@20	R@1	R@5
Paired								
LanguageBind	12.42	36.70	11.32	31.03	–	–	1.52	6.36
Ex-MCR	19.07	47.05	7.01	22.04	–	–	1.57	5.94
Ours$_{upper\text{-}bound}$	19.79	51.48	9.53	26.56	27.97	22.03	–	–
Unpaired								
Naïve	0.02	0.35	0.04	0.23	10.17	4.24	0.02	0.06
COX†	0.08	0.64	0.11	0.78	7.63	2.54	0.02	0.10
Ours$_{LB}$	14.54	41.02	6.93	22.33	29.66	20.34	0.92	**3.42**
Ours$_{NV}$	**16.20**	**45.15**	7.75	23.73	26.27	22.88	0.74	3.28
Ours$_{Qwen}$	15.35	43.88	**7.81**	**23.81**	**31.36**	**26.27**	**1.06**	3.14
Avg. Preservation	77.6%	84.2%	78.7%	87.7%	104.0%	105.1%	–	–

Table 2: Zero-shot classification performance across diverse modalities.

	Audio			3D				X-ray
	AudioSet	ESC-50		ModelNet40		ScanObjectNN		RSNA
Method	mAP	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1
Paired								
LanguageBind	18.33	94.00	99.70	–	–	–	–	–
Ex-MCR	6.67	71.20	96.80	66.53	93.60	40.31	77.20	–
Ours$_{upper\text{-}bound}$	6.67	70.55	94.25	81.85	97.00	61.56	88.44	52.71
Unpaired								
Naïve	1.14	2.90	8.45	0.81	8.95	3.32	30.52	**26.36**
COX†	1.26	2.00	10.00	4.05	13.70	2.84	26.68	23.18
Ours$_{LB}$	**6.42**	74.65	94.60	**81.12**	**97.49**	**54.81**	**84.88**	26.03
Ours$_{NV}$	5.13	**79.40**	**97.20**	76.30	94.37	40.24	75.95	22.26
Ours$_{Qwen}$	5.80	77.25	96.85	70.86	92.14	42.15	77.89	22.46
Avg. Preservation	86.7%	109.3%	102.1%	93.0%	97.6%	74.3%	90.0%	44.7%

baseline using PCA projection to the source embedding space (i.e., 768 dimensions for LanguageBind) with standard normalization, demonstrating that simple dimensionality reduction without learned alignment is insufficient.

Evaluation Tasks. To verify the effectiveness of TextME, we evaluate its performance on two categories of cross-modal downstream tasks. For **cross-modal retrieval**, we evaluate: (i) *Text*→*X* retrieval on AudioCaps (Kim et al., 2019), Clotho (Drossos et al., 2020), and DrugBank (Knox et al., 2024) (using MRR@k for molecules following MoleculeSTM (Liu et al., 2023)); (ii) *X*→*X* retrieval on Flickr30k (Plummer et al., 2015) for Audio→Image, demonstrating emergent cross-modal capabilities between modalities never paired during training. For **zero-shot classification**, we test on ModelNet40 (Qiu et al., 2021) and ScanObjectNN (Uy et al., 2019) for 3D point clouds, AudioSet (Gemmeke et al., 2017) and ESC-50 (Piczak) for audio, and RSNA Pneumonia Detection (RSNA, 2018) for X-ray images, reporting top-k accuracy and mAP.

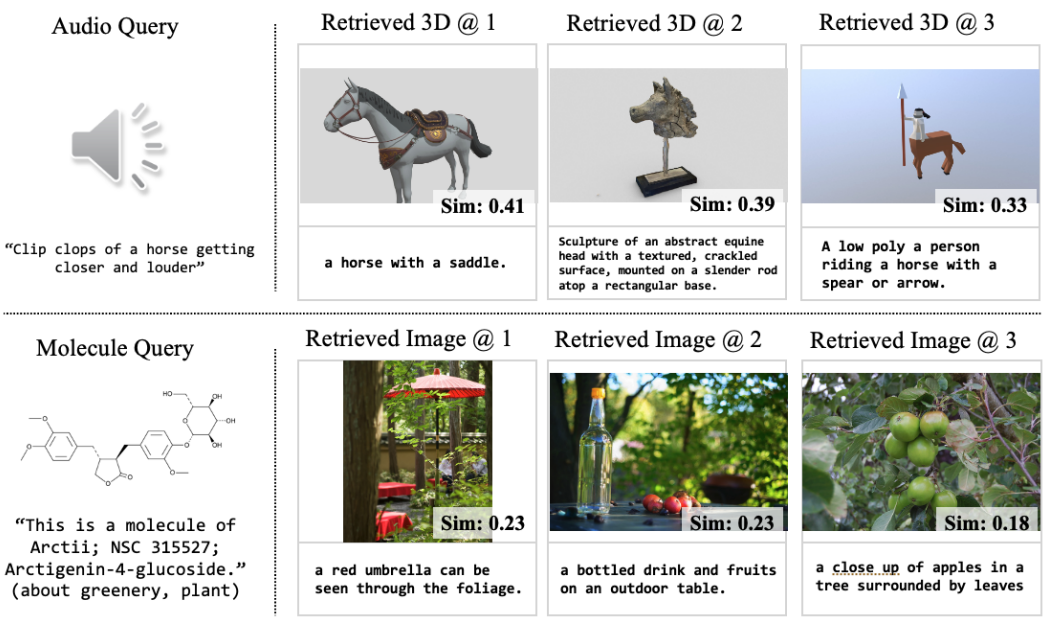

Figure 3: Emergent cross-modal retrieval without paired supervision. Audio queries retrieve semantically related 3D objects (top), and molecular structures retrieve contextually appropriate images (bottom). Results obtained by sampling $5,000$ instances per modality and performing cosine similarity-based retrieval in the learned embedding space.

4.2 Zero-shot Cross-modal Task Performance

Tables 1 and 2 demonstrate that TextME achieves competitive zero-shot performance across diverse modalities. To understand what drives this efficacy, we examine two key factors: the effectiveness of cross-modal mapping mechanisms and the impact of shared anchor space selection.

Effectiveness of Cross-Modal Mapping. While all three hypotheses are empirically validated in Section 2.2, Hypothesis 0 emerges as the critical determinant of cross-modal transfer performance, as centroid proximity directly governs the effectiveness of the constant offset approximation central to TextME's approach. The experimental results demonstrate a clear inverse correlation between centroid proximity and performance preservation: Molecule, 3D, and Audio modalities exhibit relatively low centroid proximity values, enabling an average preservation rate of 91.56% that approaches or exceeds paired-data performance. In contrast, X-ray modality shows significantly elevated centroid proximity, creating dispersed embeddings that undermine the constant offset approximation and result in substantially degraded preservation. Overall, TextME achieves strong preservation rates compared to paired-data methods while significantly outperforming unpaired baselines like COX and the naïve approaches, demonstrating the effectiveness of exploiting consistent modality gap.

Effectiveness of Shared Anchor Selection. We evaluate the performance of TextME across three different shared anchor spaces—LanguageBind (Zhu et al., 2023) for multimodal models' text encoder and NV-Embed-v2 (Lee et al., 2024), Qwen3-Embedding-4B (Zhang et al., 2025) for LLMs. Our evaluation reveals a clear pattern: LLM-anchored methods excel on retrieval tasks, while multimodal anchoring performs better on classification tasks. This distinction likely reflects the fundamental differences in their training objectives. Specifically, LLM embeddings capture rich semantic relationships crucial for matching natural language queries in retrieval, while multimodal encoders trained with contrastive objectives develop discriminative boundaries better suited for categorical classification. These complementary strengths demonstrate that optimal anchor selection depends on downstream task requirements, validating our framework's flexibility in accommodating various semantic pivot spaces. We leave the exploration of unified anchoring strategies that leverage these complementary strengths as future work.

4.3 Qualitative Analysis of Cross-modal Capabilities

To evaluate emergent cross-modal transfer capabilities, we conducted qualitative analysis through retrieval experiments involving modality pairs not present in the training data. Due to the absence of established benchmarks for these novel cross-modal tasks, we designed an evaluation protocol sampling $5,000$ instances from AudioCaps (Kim et al., 2019) for audio, Objaverse (Deitke et al., 2023) for 3D, PubChem (Kim et al., 2025) for Molecule, and COCO (Lin et al., 2014) for image, performing nearest-neighbor retrieval based on cosine similarity in the learned embedding space. Results in Figure 3 demonstrate semantic coherence across modalities. The audio-to-3D retrieval correctly associates acoustic signatures with corresponding semantics—equestrian sounds retrieve morphologically appropriate horse models, indicating a preserved semantic understanding of auditory features that have been transformed across the modality gap. Similarly, molecule-to-image retrieval reveals that chemical compounds described with pharmaceutical terminology to retrieve semantically related visual scenes. These observations suggest that text-only training with a simple offset operation successfully preserves semantic information from non-text modalities during inference, despite the projection networks being trained exclusively on textual representations.

5 Related Work

Modality Expansion. Contrastive learning has emerged as the dominant paradigm for multimodal alignment, pioneered by CLIP (Radford et al., 2021) for vision-language tasks. This success motivated extensions to multiple modalities: ImageBind (Girdhar et al., 2023) uses images as a central hub to align co-occurring modalities, while LanguageBind (Zhu et al., 2023) leverages text as a semantic pivot. To reduce computational costs, recent methods connect frozen pre-trained encoders through lightweight projectors—C-MCR (Wang et al., 2023b) and Ex-MCR (Zhang et al., 2024b) learn adapters between encoders, while FreeBind (Wang et al., 2024a) and OmniBind (Wang et al., 2024b) ensemble multiple encoders per modality. However, all these approaches require instance-level correspondence between modalities through paired supervision. This requirement becomes prohibitive in specialized domains (e.g., medical imaging, molecular analysis) where paired data is scarce or infeasible, and the computational complexity scales quadratically with the number of modalities.

Text-only Training. Recent work has explored cross-modal alignment using only unimodal data to circumvent the requirement for paired data. DeCap (Li et al., 2023) learns image captioning without image-text pairs by training a decoder to reconstruct sentences from CLIP text embeddings, then projecting image embeddings into text space at inference. LinCIR (Gu et al., 2024) distills sentence semantics into token-like representations for composed image retrieval. Separately, theoretical analyses have revealed that contrastively-trained encoders exhibit a consistent modality gap—a systematic offset between text and non-text embeddings that can be eliminated through mean-centering (Liang et al., 2022; Zhang et al., 2023b; 2024a). While these approaches demonstrate the feasibility of text-only training, they remain limited to specific tasks or modality pairs. TextME unifies these insights by combining modality gap correction with LLM-anchored semantic alignment, providing the first framework for expanding to arbitrary specialized modalities.

6 Conclusion

We presented TextME, a text-only training framework leveraging the consistent modality gap in pre-trained encoders to enable zero-shot cross-modal transfer using only text descriptions. Through experiments across audio, 3D, medical X-ray, and molecular domains, TextME achieved 88.2% average performance preservation compared to paired-data methods, while reducing data requirements by 95%, demonstrating that text-anchored text embeddings can effectively serve as semantic bridges between arbitrary modalities. Our framework addresses the critical bottleneck of paired dataset scarcity in specialized domains, establishing a scalable paradigm for multimodal representation learning in resource-constrained and data-limited settings where conventional paired data acquisition remains computationally prohibitive or infeasible in real-world scenarios. Future work will explore the relationship between modality-specific characteristics and alignment quality, examining how inherent properties of different data types influence the effectiveness of text-based bridging and developing adaptive strategies that better account for these variations.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

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

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. L+ m-24: Building a dataset for language+ molecules@ acl 2024. *arXiv preprint arXiv:2403.00791*, 2024.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 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.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 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.

Google. Gemini api, 2024. URL `https://ai.google.dev/gemini-api`. Accessed: January 2025.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Zhuo Huang, Gang Niu, Bo Han, Masashi Sugiyama, and Tongliang Liu. Towards out-of-modal generalization without instance-level modal correspondence. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2025 update. *Nucleic acids research*, 53(D1):D1516–D1525, 2025.

Felipe C Kitamura, Luciano M Prevedello, Errol Colak, Safwan S Halabi, Matthew P Lungren, Robyn L Ball, Jayashree Kalpathy-Cramer, Charles E Kahn Jr, Tyler Richards, Jason F Talbott, et al. Lessons learned in building expertly annotated multi-institution datasets and hosting the rsna ai challenges. *Radiology: Artificial Intelligence*, 6(3):e230227, 2024.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*, 2025.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36:75307–75337, 2023.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26752–26762, 2024.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Contrastive audio-language learning for music. *arXiv preprint arXiv:2208.12208*, 2022.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL `http://dl.acm.org/citation.cfm?doid=2733373.2806390`.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Shi Qiu, Saeed Anwar, and Nick Barnes. Geometric back-projection network for point cloud classification. *IEEE Transactions on Multimedia*, 24:1943–1955, 2021.

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

Philipp J Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovickỳ. Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples. *arXiv preprint arXiv:2403.02875*, 2024.

RSNA. RSNA Pneumonia Detection Challenge. `https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge`, 2018. [Online; accessed 28-Aug-2018].

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8985–8994, 2025.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4358–4366, 2018.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1588–1597, 2019.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023a.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5123–5133, 2025.

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

Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng Huang, Yang Zhao, Tao Jin, Peng Gao, et al. Freebind: Free lunch in unified multimodal space via knowledge fusion. *arXiv preprint arXiv:2405.04883*, 2024a.

Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024b.

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

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G Honavar. Molbind: Multimodal alignment of language, molecules, and proteins. *arXiv preprint arXiv:2403.08167*, 2024.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111. Springer, 2023.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023a.

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023b.

Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. *arXiv preprint arXiv:2401.08567*, 2024a.

Ziang Zhang, Zehan Wang, Luping Liu, Rongjie Huang, Xize Cheng, Zhenhui Ye, Huadai Liu, Haifeng Huang, Yang Zhao, Tao Jin, et al. Extending multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 37:91880–91903, 2024b.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1): 13524, 2021.

A GENERATIVE AI USAGE DISCLOSURE

During the preparation of this manuscript, the following generative AI tools were used:

- **GitHub Copilot** was used solely for code completion and code snippet generation during the development of experimental pipelines and auxiliary scripts. All generated code was manually reviewed and, where necessary, modified by the authors.
- **Grammarly, Perplexity and ChatGPT** were used only for grammar checking, typo correction, and minor language editing of author-written text. No sections of the paper were written or generated entirely by generative AI models; all scientific content, analysis, and claims were produced by the authors.

No generative AI tool was used to produce any scientific content, experimental results, or substantive text in the manuscript. The use of generative AI tools was strictly limited to the above purposes, in accordance with ICLR 2026 policy.

B STATISTICS OF MODALITY GAP

To analyze the geometric structure of pre-trained multimodal encoders, we compute several statistics that characterize the separation between text and non-text embedding spaces.

B.1 STATISTICAL DEFINITIONS

For each encoder E_m, we randomly sample $N = 5000$ text-modal pairs (t_i, x_i) and compute:

- **Individual gap**: $\mathbf{d}_j^{(i)} = \mathbf{e}_{x_j}^{(i)} - \mathbf{e}_{t_j}^{(i)}$ — the vector difference between paired embeddings.
- **Group gap**: $\mathbf{d}^{(i)} = \mathbb{E}_j[\mathbf{d}_j^{(i)}] = \frac{1}{N}\sum_{j=1}^{N} \mathbf{d}_j^{(i)}$ — the average gap across all pairs.
- **Gap length**: $\|\mathbf{d}^{(i)}\|_2$ — the magnitude of the average gap vector.
- **Gap direction consistency**: $\cos(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}) = \frac{\mathbf{d}^{(i)} \cdot \mathbf{d}^{(j)}}{\|\mathbf{d}^{(i)}\|_2 \|\mathbf{d}^{(j)}\|_2}$ — cosine similarity between gap vectors from different sample sets.
- **Content orthogonality**: $\cos(\mathbf{d}^{(i)}, \mathbf{r}_{j,k}^{(i)})$ where $\mathbf{r}_{j,k}^{(i)} = \mathbf{e}_{x_j}^{(i)} - \mathbf{e}_{x_k}^{(i)}$ — cosine between gap and content variations.
- **Alignment noise**: $\boldsymbol{\epsilon}_j^{(i)} = \mathbf{d}_j^{(i)} - \mathbf{d}^{(i)}$ — deviation from the average gap.

B.2 INTERPRETATION OF STATISTICS

The statistics in Figure 4 reveal three key properties:

High directional consistency ($\cos(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}) > 0.96$ for most encoders) indicates that the modality gap is nearly constant across the entire dataset, suggesting a systematic geometric separation rather than instance-specific variations.

Near-zero orthogonality ($\cos(\mathbf{d}^{(i)}, \mathbf{r}_{j,k}^{(i)}) \approx 0$) confirms that the gap direction is perpendicular to semantic content variations, meaning the modality gap represents a pure coordinate shift independent of semantic information.

Zero-mean alignment noise ($\mathbb{E}[\boldsymbol{\epsilon}_j^{(i)}] \approx 0$) validates that individual gaps cluster tightly around the mean, supporting our approximation of the modality gap as a single constant vector.

These properties justify our centering-based approach. Since the modality gap is consistent and orthogonal to content, we can create an interchangeable space through independent centering without losing semantic information.

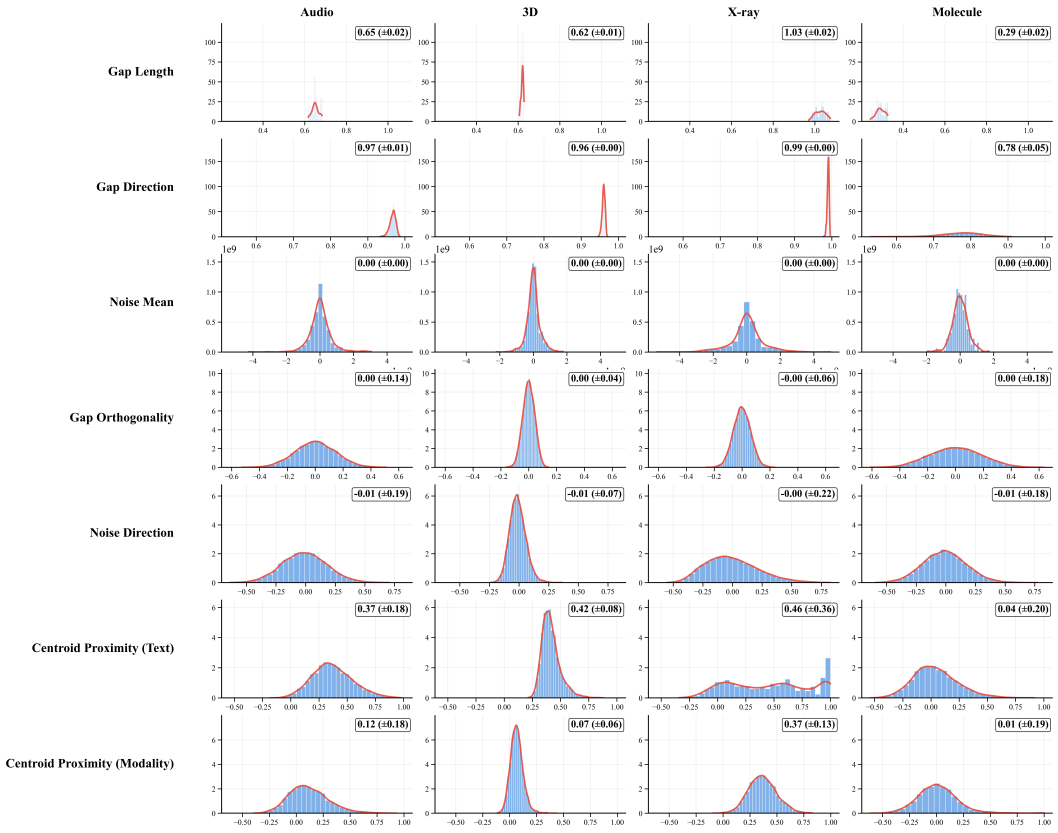

Figure 4: Distribution of modality gap statistics across five multimodal encoders. From top to bottom: gap length, gap direction consistency, gap orthogonality, alignment noise, and noise direction. Red curves show fitted normal distributions. The consistent patterns across all encoders demonstrate the systematic nature of the modality gap.

C Semantic Textual Similarity Benchmark Analysis

To validate our choice of LLM embeddings as the semantic anchor space, we conducted comprehensive evaluation on the Semantic Textual Similarity (STS) benchmark suite. Table 3 presents the Spearman correlation scores across six STS tasks (STS12-16 and STSBenchmark) comparing multimodal encoders (CLIP (Radford et al., 2021), LanguageBind (Zhu et al., 2023)) against LLM embedding models (NV-Embed-v2 (Lee et al., 2024), Qwen3-Embed variants (Zhang et al., 2025)).

Table 3: Semantic Textual Similarity (STS) benchmark performance comparison between multimodal encoders and LLM embedding models. Spearman correlation scores across six STS tasks.

Model	STS Tasks (Spearman ρ)						Avg.
	STS12	STS13	STS14	STS15	STS16	STSBenchmark	
Multimodal Encoders							
CLIP	61.87	63.83	62.09	76.82	72.89	72.26	68.29
LanguageBind	63.12	67.46	63.27	73.82	73.73	71.60	68.83
LLM Embedding Models							
NV-Embed-v2	77.89	88.30	84.30	89.04	86.77	88.41	85.79
Qwen3-Embed-0.6b	79.35	87.31	79.81	87.28	87.07	86.51	84.56
Qwen3-Embed-4b	**84.31**	**93.20**	**88.61**	**92.31**	**92.07**	**91.92**	**90.40**

The results demonstrate a substantial performance gap, with LLM embedding models achieving average correlations of 84.56-90.40 compared to 68.29-68.83 for multimodal encoders. This 22-point difference in correlation scores indicates that LLM embeddings capture more nuanced semantic relationships in textual data. The superior performance stems from their distinct training objectives: while multimodal encoders optimize for cross-modal alignment through contrastive losses, LLMs undergo extensive next-token prediction on diverse text corpora, learning complex linguistic patterns and semantic nuances. These findings provide empirical justification for employing LLM embeddings as the shared anchor space in TextME, particularly when training exclusively on text descriptions without paired multimodal supervision.

D Algorithm Details

This section provides a comprehensive description of the TextME framework's algorithmic details. We describe the three-stage pipeline that enables modality expansion using only text data: (1) precomputing centering offsets for text and non-text embedding alignment, (2) training lightweight projectors on centered text embeddings, (3) and performing inference-time adaptation for non-text modalities. Algorithm 1 formalizes this procedure, showing how we leverage the inherent structure of pre-trained multimodal encoders to achieve zero-shot cross-modal transfer without paired supervision.

Algorithm 1 LLM-anchored Modality Expansion (LLaME)

Require: Encoders $\{E_m\}_{m\in\mathcal{M}}$, LLM encoder E_{LLM}, texts $\mathcal{D}_{\text{text}}$
Ensure: Projections $\{P_m\}_{m\in\mathcal{M}}$, offsets $\{\mu_m^{\text{text}}, \mu_m^{\text{modal}}\}_{m\in\mathcal{M}}$

Stage 1: Compute Centering Offsets
for each modality $m \in \mathcal{M}$ **do**
  $\mu_m^{\text{text}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} E_m^{\text{text}}(t_i)$ *// Text centroid*
  $\mu_m^{\text{modal}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} E_m^{\text{modal}}(x_i)$ *// Modal centroid*
end for
Stage 2: Text-to-Text Alignment
for each modality $m \in \mathcal{M}$ **do**
  Initialize $P_m : \mathbb{R}^{d_m} \to \mathbb{R}^{d_{\text{LLM}}}$ as 2-layer MLP
  while not converged **do**
    Sample batch $\{t_i\}_{i=1}^{B} \sim \mathcal{D}_{\text{text}}$
    $\mathbf{z}_i \leftarrow P_m(E_m^{\text{text}}(t_i) - \mu_m^{\text{text}})$ for $i \in [1, B]$
    $\mathbf{z}'_i \leftarrow E_{\text{LLM}}(t_i)$ for $i \in [1, B]$
    Select hard negatives: $\mathcal{N}_i = \{j : \text{sim}(\mathbf{z}_i, \mathbf{z}'_j) \in [0.1s_i, 0.9s_i]\}$
      where $s_i = \text{sim}(\mathbf{z}_i, \mathbf{z}'_i)$
    $\mathcal{L} \leftarrow -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}'_i)/\tau)}{\sum_{j\in\mathcal{N}_i\cup\{i\}} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}'_j)/\tau)}$
    $P_m \leftarrow P_m - \eta\nabla_{P_m}\mathcal{L}$
  end while
end for
Stage 3: Cross-Modal Inference
Given input x of modality m:
$\mathbf{e} \leftarrow E_m^{\text{modal}}(x)$ *// Encode modal input*
$\mathbf{e}' \leftarrow \mathbf{e} - \mu_m^{\text{modal}}$ *// Apply offset*
$\mathbf{e}_{\text{final}} \leftarrow P_m(\mathbf{e}')$ *// Project to LLM space*

Implementation details and pre-computed offsets will be available in our open-source release upon publication.

D.1 Centering-based Interchangeability

The core insight of TextME's algorithm is that pre-trained multimodal encoders trained with contrastive objectives naturally separate text and non-text embeddings into distinct subspaces. Rather than attempting to bridge this modality gap directly, we create an interchangeable coordinate system through independent centering, following previous works Zhang et al. (2023b; 2024a).

Given an encoder E_m with text branch E_m^{text} and modal branch E_m^{modal}, we compute centering offsets:

$$\mu_m^{\text{text}} = \mathbb{E}[E_m^{\text{text}}(t)] \quad \text{(mean of text embeddings)} \tag{4}$$

$$\mu_m^{\text{modal}} = \mathbb{E}[E_m^{\text{modal}}(x)] \quad \text{(mean of modal embeddings)} \tag{5}$$

By centering each modality independently ($e' = e - \mu$), we align their coordinate origins, enabling the projection network trained on centered text embeddings to generalize to centered modal embeddings at inference.

D.2 Implementation Notes

Stage 1: Offset Computation. The centering offsets require only 5,000 samples per modality, as the mean embeddings converge quickly (see Section **??**). These offsets are computed once before training and remain fixed. Importantly, text and modal samples need not be paired—we simply need representative samples from each distribution.

Stage 2: Text Alignment Training. The projection network P_m is implemented as a 2-layer MLP with GeLU activation and hidden dimension of $2,560$. We train exclusively on centered text embeddings, learning to map from the encoder's text space to the LLM embedding space. Hard negative mining improves convergence by focusing gradients on challenging examples where $\text{sim}(z_i, z'_j)$ falls within $[0.1, 0.9]$ of the positive pair similarity.

Stage 3: Cross-Modal Inference. At inference, modal inputs are processed through three steps:

1. Encoding: $\mathbf{e} = E_m^{\text{modal}}(x)$
2. Centering: $\mathbf{e}' = \mathbf{e} - \mu_m^{\text{modal}}$
3. Projection: $\mathbf{e}_{\text{final}} = P_m(\mathbf{e}')$

The centering step transforms the modal embedding into the same coordinate system used during text training, enabling zero-shot cross-modal transfer.

D.3 Computational Complexity

The algorithm's efficiency stems from its minimal requirements:

- **Memory**: Store only centering offsets ($2 \times d_m$ floats per encoder) and projection networks ($\sim$10M parameters per modality).
- **Training**: Process only text data, reducing data requirements by $>$99% compared to paired multimodal training.
- **Inference**: Add only one vector subtraction to standard encoder inference.

This design enables practical deployment even in resource-constrained settings while maintaining competitive performance with fully-supervised methods.

E Implementation Details

E.1 TextME Implementation

Model Architecture. Each projection network P_m is implemented as a 2-layer MLP with a hidden dimension of each pre-trained encoders and GeLU activation. The input dimension d_m varies according to the source encoder's embedding dimension, while the output dimension is fixed at $d_{\text{LLM}} = 2560$ to match the Qwen3-Embedding-0.6B anchor space.

Training Configuration. We train each projection network with the following hyperparameters:

- Batch size: 512
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01

- Learning rate: 5×10^{-4} with cosine annealing schedule
- Training epochs: 50
- Temperature parameter: $\tau = 0.07$
- Hard negative mining: Select examples where similarity falls within $[0.1 \cdot \text{sim}(z_i, z_i'), 0.9 \cdot \text{sim}(z_i, z_i')]$
- Mixed precision training: fp16

Data Preprocessing. For each target modality, we sample 100K text descriptions from the modality-specific training dataset. Text inputs are tokenized using the respective encoder's tokenizer with a maximum sequence length of 77 tokens. Centering offsets are pre-computed using 5,000 randomly sampled text-modal pairs per modality and remain fixed throughout training.

Computational Resources. All experiments are conducted on a single NVIDIA A6000 GPU with 48GB memory. Training time per modality averages 2 hours, with peak memory usage of approximately 8GB.

E.2 COX Baseline Implementation

Since the original COX (Huang et al., 2025) codebase was not publicly available, we reimplemented the method following the paper. However, we adapted the approach to an zero-shot setting to match the evaluation constraints. Our implementation adheres to the following configuration.

Architecture. COX trains target modality encoders from scratch using a unified architecture across modalities. We employ Vision Transformer Tiny (ViT-T/16) as the encoder backbone, consisting of 12 layers with 3 attention heads and an embedding dimension of 192. The final embedding dimension is projected to 768 to align with LanguageBind's representation space. Following the original design, we incorporate a Variational Information Bottleneck (VIB) layer (Alemi et al., 2016) that applies stochastic dimensionality reduction to 256 dimensions to enforce information compression.

Training Protocol. The training protocol follows the original paper's two-stage methodology. In the first stage, we perform supervised pre-training on labeled target data for 10 epochs to establish basic feature representations. The second stage applies information bottleneck fine-tuning for 50 epochs to learn generalizable features through information compression. We use a batch size of 256 with the Adam optimizer configured with a learning rate of 1×10^{-3} and weight decay of 1×10^{-5}. The learning rate follows a step decay schedule with reduction at predetermined epochs. Critically, COX requires labeled data from the target modality, using approximately 10% of the labeled dataset (roughly $10K$ samples) for training. Specifically, we utilize labeled datasets for each modality: COCO with 80 object classes for visual tasks, ESC-50 with 50 environmental sound classes for audio, Objaverse with 1,000 object classes for 3D point clouds, PubChem with 100 molecular classes for chemical structures, and SIIM with 2 classes for medical X-ray classification.

The key differences from TextME are substantial. COX requires labeled data for the target modality, necessitates training encoders from scratch with over 300M parameters, and demands architectural alignment between source and target encoders. In contrast, TextME leverages pre-trained encoders with only text descriptions, requires merely 10M trainable parameters, and imposes no architectural constraints on the target encoders. These fundamental differences highlight the efficiency and flexibility advantages of our approach over traditional modality generalization methods.

F Ablation Study on Number of Samples for Offset

We investigate the impact of the number of samples used to compute the centering offset in our method. The centering offset is a crucial component that helps align representations from different modalities by estimating and removing systematic biases in the embedding space. To understand how sensitive our approach is to the sample size used for offset computation, we conduct experiments with varying numbers of samples ranging from 100 to 10,000.

Table 4: Impact of sample size for computing centering offsets on performance.

# Samples	AudioCaps R@1	ModelNet40 Acc.	DrugBank R@1	RSNA Acc.	Relative Perf.
100	14.91	70.66	34.75	23.01	90%
500	14.77	70.58	33.05	22.08	95%
1,000	14.89	70.62	**36.44**	22.56	97%
5,000 (default)	**15.35**	**70.86**	31.36	22.46	100%
10,000	14.95	70.58	32.20	**22.73**	100%
Std.	0.21	0.11	2.02	0.36	-

Table 4 presents the results across four diverse tasks: AudioCaps (audio-text retrieval), ModelNet40 (3D shape classification), DrugBank (molecular retrieval), and RSNA (medical image classification). We report Recall@1 (R@1) for retrieval tasks and accuracy for classification tasks, along with the relative performance compared to our default setting of 5,000 samples. Our results reveal several important findings.

Stability across sample sizes. The method demonstrates remarkable stability across different sample sizes. Even with as few as 100 samples, we achieve 90% of the performance obtained with 5,000 samples, indicating that our centering approach is robust and does not require extensive sampling to estimate reliable offsets.

Optimal range. Performance plateaus between 1,000 and 10,000 samples, with our default choice of 5,000 samples providing a good balance between computational efficiency and performance. The relative performance reaches 97% with just 1,000 samples and remains stable at 100% for both 5,000 and 10,000 samples. **Task-specific variations.** While AudioCaps and ModelNet40 show consistent improvements with increased sample size up to 5,000, DrugBank exhibits more variance (std = 2.02), with the best performance surprisingly achieved at 1,000 samples (36.44 R@1). This suggests that for some domains, particularly those with inherently more diverse or noisy representations, the optimal sample size may vary.

Diminishing returns. Doubling the sample size from 5,000 to 10,000 provides no significant improvement and even shows slight degradation in some metrics (AudioCaps R@1: 15.35 $\rightarrow$ 14.95, ModelNet40 Acc: 70.86 $\rightarrow$ 70.58), indicating that beyond a certain threshold, additional samples do not contribute to better offset estimation and may introduce noise.

These findings have important practical implications for deployment. The robustness to small sample sizes means our method can be effectively applied even in low-resource scenarios where obtaining large numbers of samples might be challenging.

G Ablation Study on Domain-Specific Training Data Requirements

Different modalities exhibit distinct representational characteristics that may require tailored training data. We hypothesize that modalities with specialized vocabularies and domain-specific concepts benefit more from targeted text descriptions than those with representations already aligned with general language. To test this, we compare training with general-purpose text (Wiki1M) versus modality- specific captions.

Table 5 quantifies this intuition through the distance ratio $\rho = d_{intra}/d_{inter}$, where d_{intra} and d_{inter} represent average pairwise distances within and between datasets, respectively. Lower ratios indicate specialized, sparse distributions distinct from general language, while higher ratios suggest denser distributions with greater linguistic overlap. Our results validate that specialized modalities ($\rho < 1.7$) like molecules and audio show dramatic improvements with domain-specific data training (181.9% and 170.7%) gains respectively), as their technical vocabularies require precise terminology. Conversely, modalities closer to general language ($\rho > 2.4$) like images and 3D show modest gains (19.7% and 23.4%), as their semantic concepts largely overlap with general linguistic corpora. This pattern extends to cross-modal retrieval—Audio$\rightarrow$Image improves by 55.9% with specialized

Table 5: Impact of training data quality on performance across modalities.

	Dense		Sparse		Cross-modal
	Image Flickr	3D ScanObj.	Molecule DrugBank	Audio AudioCaps	Audio→Image FlickrNet
Distance ratio (ρ)	2.93	2.44	1.65	1.55	–
Wiki1M (general)	43.16	34.15	9.32	5.67	0.68
Modality-specific Captions	**51.66**	**42.15**	**26.27**	**15.35**	**1.06**
Δ (%)	+19.7%	+23.4%	+181.9%	+170.7%	+55.9%

captions, demonstrating that matching training data to modality-specific representational characteristics enhances semantic alignment even without paired supervision.