

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 BARYBIND: BINDING ALL MODALITIES VIA MULTIMODAL WASSERSTEIN BARYCENTER SPACE

Anonymous authors

Paper under double-blind review

## ABSTRACT

Multimodal joint representation, which aligns multiple modalities in a shared latent space, has emerged as the foundation of recent multimodal understanding models. To scale beyond two modalities, existing models typically **treat a specific modality (e.g., text) as the anchor** to bind other modalities via pairwise contrastive losses. However, the learned joint representation space tends to be sub-optimal and imbalanced, **as the modality-specific anchor may inherit the modality bias and insufficiently capture the modality-agnostic semantics and holistic geometric structures within multimodal data**. In this work, we are motivated by the intuition that multimodal representations arise from different shifts from an underlying modality-agnostic representation space. Based on this, we present **BaryBind**, a multimodal framework that aligns modalities in the multimodal Wasserstein barycenter (WB) space, which inherently models a modality-agnostic distribution by minimizing the average of Wasserstein distances to all modalities. We further construct a barycenter polytope, whose volume serves as a geometric metric for quantifying  $n$ -modality alignment. This metric is integrated as a barycenter-anchored volumetric contrastive loss that contrasts the volumes of the  $n$ -dimensional polytopes, encouraging global alignment of non-anchor modalities to the barycenter while reducing inter-modality gaps. Extensive experiments show that BaryBind delivers more balanced zero-shot generalization performance in downstream tasks, e.g., cross-modal text/video retrieval and classification.

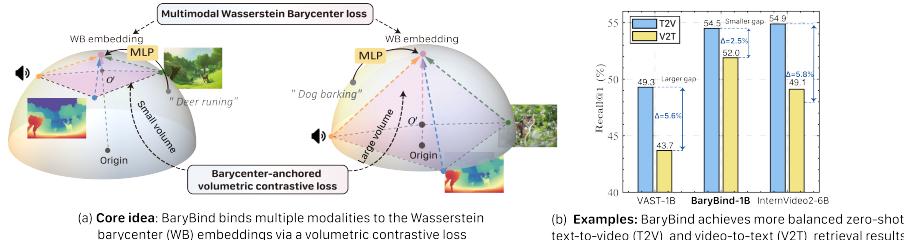


Figure 1: **BaryBind binds multiple modalities to the Wasserstein barycenter (WB)**, which encodes modality-agnostic semantics by minimizing the average of Wasserstein distances to all modalities. By contrasting the barycenter polytope’s volume, BaryBind achieves  $n$ -modality alignment to the WB space while preserving inter-modal interactions. Notably, it delivers more balanced retrieval results, reducing the T2V/V2T gap to 2.5% (vs. 5.6% for our baseline VAST [Chen et al. \(2023\)](#)).

## 1 INTRODUCTION

Multimodal learning ([Baltrušaitis et al., 2018](#)) seeks to integrate and process heterogeneous signals from multiple modalities (e.g., vision, language, audio, depth, *etc.*) to build a coherent perception of the surrounding world. Since multimodal data arise from heterogeneous observations of a shared underlying reality, recent multimodal learning methods ([Radford et al., 2021](#); [Jia et al., 2021](#)) learn a shared representation space for representations from different modalities. In this field, the success of CLIP ([Radford et al., 2021](#)) in aligning unified vision–language representations via contrastive learning has sparked the adoption of contrastive losses as an appealing solution for multimodal representation learning. However, traditional contrastive losses, such as InfoNCE ([Oord et al., 2018](#)) and BYOL ([Grill et al., 2020](#)), are formulated in a pairwise fashion for  $n = 2$  representation spaces typically arising from two modalities such as image–text ([Radford et al., 2021](#); [Jia et al., 2021](#)) or audio–text ([Guzhov et al., 2022](#); [Elizalde et al., 2023](#)) scenarios.

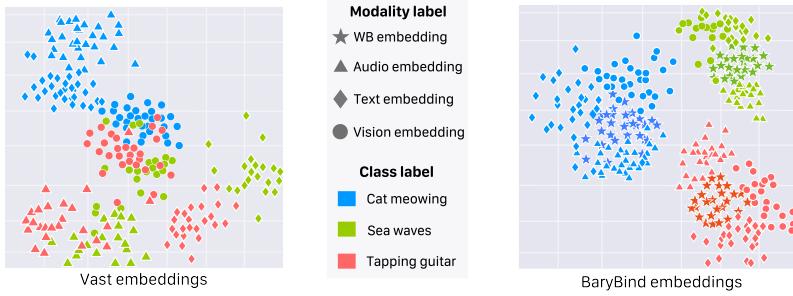


Figure 2: **The t-SNE comparison of embeddings on the zero-shot VGGSound dataset** (Chen et al., 2020). Compared to VAST (Chen et al., 2023), BaryBind induces a latent space where class clusters are clearly separated and multimodal embeddings are grouped around the WB embeddings.

While scaling to  $n$  modalities ( $n \geq 3$ ) poses unique challenges, a series of recent works (Chen et al., 2023; Zhu et al., 2024; Wang et al., 2025) originating from ImageBind (Girdhar et al., 2023) leverage the binding property of an modality-specific anchor (e.g., image or language) to align other modalities to the chosen anchor via pairwise losses. However, treating a specific modality as the alignment center can lead to sub-optimal shared space, as it inherits modality-specific biases while overlooking modality-agnostic semantics shared across all modalities. Consequently, the learned representations tend to be imbalanced, with a certain modality dominating the joint representation space (see the video/text retrieval results in Fig. 1). Moreover, although the pairwise losses align each modality to the anchor, they omit the correlations/interactions among non-anchor modalities, which may undermine the  $n$ -modality global alignment consistency.

To address these challenges, this work is motivated by the notion that *multimodal data are collected from heterogeneous sensors observing a shared underlying reality, thus the multimodal representations arise from different shifts from an underlying modality-agnostic representation space*. With this insight, we propose **BaryBind** that aligns multimodal representations in the multimodal Wasserstein Barycenter (WB) space, which inherently models a modality-agnostic distribution that minimizes the average of optimal transport (OT) distances to all modalities while capturing the OT-grounded geometry. This property inherently filters out modality-specific redundancy and results in a modality-agnostic barycenter space that reduces the divergence caused by multimodal domain shifts. Furthermore, we suggest a barycenter-anchored volumetric contrastive loss defined on the volume of a barycenter polytope, which quantifies the global alignment of  $n$ -dimensional multimodal data. The barycenter polytopes are spanned by the barycenter and non-anchor modality-to-barycenter gap vectors (see Fig. 1 (a)). By contrasting the polytope’s volume, BaryBind binds the modalities to the modality-agnostic barycenter and reduces inter-modality gaps across modalities.

Specifically, we first propose a multimodal Wasserstein barycenter loss, which is optimized to seek the WB space for multimodal joint representation. **This is achieved by training a lightweight map to filter out modality-specific biases, transforming the original modality-specific anchor to the WB that better approximates a modality-agnostic alignment anchor.** (Fig. 1(a)) The barycenter is then leveraged to construct the  $n$ -dimensional barycenter polytope, whose volume serves as a measure of omni-modality alignment. Accordingly, we design a volumetric contrastive loss that encourages a smaller barycenter polytope volume for matched samples and a larger one for unmatched samples, thereby shaping a joint representation space that maintains the holistic geometry and inter-modal interactions within the multimodal data (Fig. 1 (a)). Our experiments show that BaryBind delivers competitive generalization performance in downstream tasks such as text/video retrieval and audio/video classification. Notably, BaryBind substantially alleviates the imbalance across modalities, reducing the T2V/V2T retrieval gap by 3.1% compared to VAST (Chen et al., 2023) (Fig. 1 (b)), and induces an embedding space where modalities of each class are consistently clustered around the WB embeddings, as presented in Fig. 2.

The main contributions of this paper are highlighted as follows:

- We propose BaryBind, a novel framework that aligns multimodal representations to the WB space. The WB space models a modality-agnostic distribution that captures the OT-grounded holistic geometry of multimodal data and inherently filters out modality-specific bias across modalities.
- We build a barycenter polytope with the barycenter and modality gap vectors. Its volume, serving as a global alignment metric, extends the insights of measuring  $n$ -modality alignment.

- 108 • Based on the polytope volume, we introduce a barycenter-anchored volumetric contrastive loss,  
109 which encourages global alignment to the barycenter while retaining inter-modal interactions.
- 110 • Extensive experiments show that BaryBind achieves more balanced representations and improves  
111 zero-shot generalization in downstream tasks such as text/video retrieval and audio/video classifi-  
112 cation. Particularly, BaryBind delivers decent scalability with increasing modality number.

## 114 2 RELATED WORKS

116 **Multimodal representation learning.** Multimodal representation learning seeks to align heterogeneous  
117 modalities into a shared semantic space. CLIP (Radford et al., 2021) initiates this paradigm  
118 with image–text contrastive learning, followed by audio extensions such as AudioCLIP (Guzhov  
119 et al., 2022), CLAP (Elizalde et al., 2023), and LAION-CLAP (Wu et al., 2023). WavCaps (Mei  
120 et al., 2024) constructs a large-scale audio captioning dataset to support audio–language retrieval.  
121 To scale beyond two modalities, ImageBind (Girdhar et al., 2023) introduces a pivot-based strat-  
122 egy, aligning each modality to the vision anchor. LanguageBind (Zhu et al., 2024), UniBind (Lyu  
123 et al., 2024), GRAM (Cicchetti et al., 2025b), and Triangle (Cicchetti et al., 2025a) extend align-  
124 ment to language anchors, adopting unique volume-based contrastive losses. MiCo (Zhang et al.,  
125 2024) effectively expands modalities, data, and model size to learn unified representations. ViT-  
126 Lens (Lei et al., 2024) innovatively adapts a pretrained ViT via modality-specific “lenses” to estab-  
127 lish a shared space. OmniBind (Wang et al., 2025) aligns pre-trained unimodal experts via pairwise  
128 losses. VAST (Chen et al., 2023) fuses modalities into a shared space but still relies on text-centered  
129 supervision. Despite their scalability, these approaches bind each modality to a specific modality  
130 (e.g., text) rather than modeling a modality-agnostic joint space. Consequently, they may suffer  
131 from modality-specific biases, which prevent them from learning a balanced joint space that cap-  
132 tures the shared modality-agnostic semantics within multimodal data. Differently, BaryBind binds  
133 the modalities to the barycenter that encodes modality-agnostic semantics and adopts a barycenter-  
134 anchored volumetric contrastive loss to encourage the global alignment across  $n$  modalities.

135 **Wasserstein barycenter.** The Wasserstein barycenter (Aguech & Carlier, 2011) defines an averaging  
136 distribution that minimizes the weighted sum of Wasserstein distances to input measures, preserving  
137 mass structure and OT-grounded geometry. This formulation has shown effectiveness on heteroge-  
138 neous supports and has been applied in generative modeling (Cuturi & Doucet, 2014) and domain  
139 adaptation (Bonneel et al., 2015). Recent works estimate high-dimensional barycenters via deep  
140 dual formulations, including ICNN-based cycle-consistent models (Korotin et al., 2021), neural OT  
141 maps (Kolesov et al., 2024a; Tang et al., 2025), and energy-guided potentials (Kolesov et al., 2024b).  
142 We extend this perspective to construct a modality-agnostic joint space for multimodal alignment,  
143 bridging optimal transport theory with multimodal representation learning.

## 144 3 BARYBIND: BINDING VIA THE WASSERSTEIN BARYCENTER SPACE

145 In this section, we present BaryBind, a multimodal learning framework that aligns different modal-  
146 ities in the Wasserstein barycenter (WB) space. Our key insight is leveraging the inherent modality-  
147 agnostic nature of WB to build a joint space that mitigates divergence caused by modality-specific  
148 shifts. By aligning modalities to this space that encodes intrinsic modality-agnostic invariance,  
149 BaryBind potentially captures more balanced multimodal representations for downstream tasks.

150 Specifically, as shown in Fig. 3, BaryBind first constructs the multimodal WB space that aligns  
151 features from multimodal latent space via the multimodal WB loss (§3.2), in which an MLP is  
152 learned to transport the anchor features to the WB space as WB embeddings. We then construct a  
153 barycenter polytope defined by the WB embeddings and modality-to-barycenter gap vectors, whose  
154 volume quantifies the degree of  $n$ -modality alignment (§3.3). Based on the polytope, we intro-  
155 duce a barycenter-anchored volumetric contrastive loss, which encourages high-order multimodal  
156 alignment to the WB space while reducing inter-modality gaps (§3.4).

### 157 3.1 PRELIMINARIES

158 **Notation.** Let  $\bar{K} = \{0, 1, \dots, K\}$  for some  $K \in \mathbb{N}$ . For a sequence  $e_0, \dots, e_K$ , we denote by  $e_{0:K}$   
159 the tuple  $(e_0, \dots, e_K)$ . Let  $\mathcal{X} \subset \mathbb{R}^d$ ,  $\mathcal{Y} \subset \mathbb{R}^{d'}$ , and  $\mathcal{X}_k \subset \mathbb{R}^{d_k}$  be compact subsets of Euclidean  
160 space. Denote by  $\mathcal{C}(\mathcal{X})$  the space of continuous real-valued functions on  $\mathcal{X}$ , and by  $\mathcal{P}(\mathcal{X})$  the set of

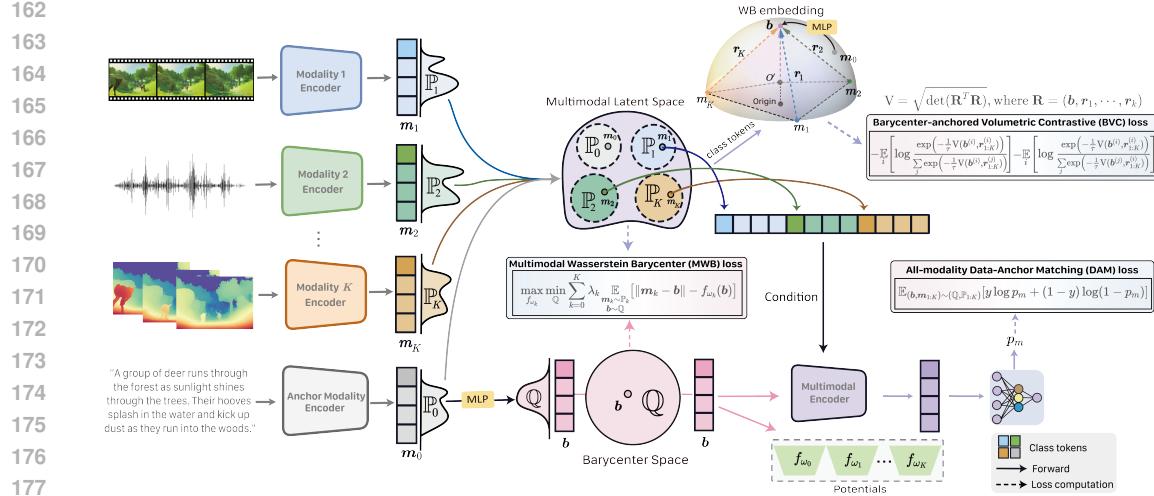


Figure 3: **The pipeline of BaryBind.** Given the multimodal embeddings, BaryBind constructs the WB space via the MWB loss, in which an MLP is learned to transform the anchor  $\mathbf{m}_0$  into the WB embedding  $\mathbf{b}$ . The WB embedding acts as the new anchor and induces a barycenter polytope, whose volume is contrasted in the BVC loss to align other modalities. WB and non-anchor embeddings are then integrated in the multimodal encoder to predict whether the anchor and data are matched.

probability measures supported on  $\mathcal{X}$ . Given  $\mathbb{P} \in \mathcal{P}(\mathcal{X})$  and  $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ , we write  $\Pi(\mathbb{P}, \mathbb{Q})$  for the set of transport plans between them, i.e., all joint distributions on  $\mathcal{X} \times \mathcal{Y}$  whose marginals are  $\mathbb{P}$  and  $\mathbb{Q}$ . The notation  $\langle \cdot, \cdot \rangle$  denotes the cosine similarity that involves the normalization over features.

**Optimal transport.** Given two distributions  $\mathbb{P} \in \mathcal{P}(\mathcal{X})$  and  $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ , along with a cost function  $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ , the classic optimal transport (OT) problem (Kantorovich, 1942) aims to find a joint distribution  $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$  that minimizes the expected transport cost:

$$\text{OT}_c(\mathbb{P}, \mathbb{Q}) \triangleq \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x, y) \sim \pi} [c(x, y)]. \quad (1)$$

The specific choice of  $c(x, y) = \|x - y\|$  yields  $W(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x, y) \sim \pi} \|x - y\|$ , known as the Earth-Mover or Wasserstein-1 distance. In this paper, we refer to it as the Wasserstein distance.

**Wasserstein barycenter (WB).** Given distributions  $\mathbb{P}_k \in \mathcal{P}(\mathcal{X}_k)$  for  $k \in \bar{K}$  and a vector  $\lambda \in \mathbb{R}^{K+1}$  of non-negative weights  $\lambda_k$  summing to 1, the WB problem seeks the distribution  $\mathbb{Q}$  that minimizes the weighted sum of Wasserstein distances to the fixed marginals  $\mathbb{P}_{0:K}$ :

$$\inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{Y})} \sum_{k=0}^K \lambda_k W(\mathbb{P}_k, \mathbb{Q}). \quad (2)$$

We apply the WB formulation to the multimodal latent space  $\mathcal{M}$ , where the encoded features of the  $k$ -th modality lie in a subspace  $\mathcal{M}_k$  and follow a distribution  $\mathbb{P}_k$ . The barycenter inherently encodes modality-agnostic semantics as it models the ‘‘closest’’ distribution to all multimodal distributions.

### 3.2 MULTIMODAL WASSERSTEIN BARYCENTER SPACE

Let  $\mathbb{P}_k$  be the distribution of encoded features  $\mathbf{m}_k \in \mathcal{M}_k \subset \mathcal{M}$  for modality  $k \in \bar{K}$ , defined in the multimodal latent space  $\mathcal{M} \subset \mathbb{R}^D$ . In particular,  $\mathbf{m}_0$  denote the anchor modality features. The WB space is defined as  $\mathcal{M}_B := \text{supp}(\mathbb{Q})$  where  $\mathbb{Q}$  denotes the WB distribution and  $\mathcal{M}_B$  contains the barycenter features  $\mathbf{b}$ . Given the distributions  $\mathbb{P}_{0:K}$ , our goal is to establish the barycenter space  $(\mathcal{M}_B, \mathbb{Q})$  and use it as the joint representation space for multimodal alignment. Based on the WB formulation (2) over multimodal latent space, the multimodal WB problem can be written as

$$\mathcal{L}_{\text{MWB}}^* = \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M})} \sum_{k=0}^K \lambda_k W(\mathbb{P}_k, \mathbb{Q}). \quad (3)$$

However, directly optimizing (3) is highly intractable. To overcome this challenge, we deduce a new dual reformulation (see Appendix A.1), which leads to the following sup-inf objective:

216 **Proposition 1 (Dual form of the multimodal WB problem).** *The infimum value  $\mathcal{L}_{\text{MWB}}^*$  of the*  
 217 *multimodal WB problem (3) can be expressed as*

$$219 \quad \mathcal{L}_{\text{MWB}}^* = \sup_{\sum_k \lambda_k f_k = 0} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sum_{k=0}^K \lambda_k \mathbb{E}_{\substack{\mathbf{m}_k \sim \mathbb{P}_k \\ \mathbf{b} \sim \mathbb{Q}}} [\|\mathbf{m}_k - \mathbf{b}\| - f_k(\mathbf{b})], \quad (4)$$

222 where the supremum is taken over all the dual potentials  $f_k : \mathcal{M}_B \rightarrow \mathbb{R}$ . We aim to learn the  
 223 distribution  $\mathbb{Q}$  by sampling the WB embeddings  $\mathbf{b} = T(\mathbf{m}_0)$  via a trainable projection map  $T_\theta(\cdot) : \mathcal{M}_0 \rightarrow \mathcal{M}_B$ . Here,  $\mathbf{m}_0$  represents the feature of anchor modality, typically selected as text, as it  
 224 retains the broad semantics of multimodal data. We also perform an ablation study on the anchor  
 225 selection, which is included in Appendix B. In practice, we simply parameterize  $T_\theta$  using an MLP.  
 226 To ensure the congruent constraint  $\sum_k \lambda_k f_k = 0$  (Li et al., 2020), we parameterize the family of  
 227 potentials  $f_{\omega_k}$  as  $g_{\omega_k} - \sum_{i=0}^K \lambda_i g_{\omega_i}$  with MLPs  $g_{\omega_k} : \mathbb{R}^D \rightarrow \mathbb{R}$ , which is a common trick used in  
 228 (Li et al., 2020; Kolesov et al., 2024a;b).  
 229

230 **Multimodal Wasserstein Barycenter (MWB) loss.** With this parameterization, we rewrite (4) as a  
 231 max-min objective of the MWB loss  $\mathcal{L}_{\text{MWB}}$ , which can be optimized to compute the map  $T_\theta$

$$233 \quad \mathcal{L}_{\text{MWB}}^*(\omega_{0:K}, \theta) \triangleq \max_{\omega_{0:K}} \min_{\theta} \sum_{k=0}^K \lambda_k \mathbb{E}_{\substack{\mathbf{m}_k \sim \mathbb{P}_k \\ \mathbf{b} \sim \mathbb{Q}}} [\|\mathbf{m}_k - T_\theta(\mathbf{m}_0)\| - f_{\omega_k}(T_\theta(\mathbf{m}_0))]. \quad (5)$$

236 To solve the problem (5), we train the networks  $T_\theta$  and  $f_{\omega_{0:K}}$  by alternately maximizing over  $\omega_{0:K}$   
 237 and minimizing over  $\theta$  in the MWB loss, in which we estimate the expectation using mini-batch data  
 238 at each training step. Given the anchor modality feature  $\mathbf{m}_0$ , the WB embedding is then computed  
 239 as  $\mathbf{b} = T_\theta(\mathbf{m}_0)$ , which will serve as the new anchor and other modalities are then aligned to it.

240 We establish the error bounds for the map  $T$  with the following simplified notations:

$$242 \quad \mathcal{F}(f_{0:K}, T) := \mathcal{L}_{\text{MWB}}(f_{1:K}, T), \quad \mathcal{L}(f_{0:K}) := \inf_{T: \mathcal{M} \rightarrow \mathcal{M}_B} \mathcal{F}(f_{0:K}, T) \quad \text{and} \quad \mathcal{L}^* := \mathcal{L}_{\text{MWB}}^*. \quad (6)$$

244 **Theorem 3.1** (Error analysis via duality gaps for the barycenter distribution). *Let  $C_k$  be any trans-*  
 245 *port costs. Assume that the maps  $\mathbf{b} \mapsto C_k(\mathbf{m}_k, \mathbf{b}) - \hat{f}_k(\mathbf{b})$  are  $\beta$ -strongly convex for  $\mathbf{m}_k \in \mathcal{M}_k$ ,*  
 246  *$k \in \{0, \dots, K\}$ . Consider the duality gaps for an approximate solution  $(\hat{f}_{0:K}, \hat{T})$ :*

$$247 \quad \mathcal{E}_1(\hat{f}_{0:K}, \hat{T}) \triangleq \mathcal{F}(\hat{f}_{0:K}, \hat{T}) - \mathcal{L}(\hat{f}_{0:K}), \quad \mathcal{E}_2(\hat{f}_{0:K}) \triangleq \mathcal{L}^* - \mathcal{L}(\hat{f}_{0:K}),$$

249 Then the following inequality holds:

$$251 \quad W^2(\hat{T}_\# \mathbb{P}_0, \mathbb{Q}^*) \leq \frac{4}{\beta} (\mathcal{E}_1 + \mathcal{E}_2). \quad (7)$$

253 The proof is provided in the Appendix A.2. This theorem ensures that the Wasserstein distance  
 254 between the estimated distribution  $\hat{T}_\# \mathbb{P}_0$  and the true barycenter  $\mathbb{Q}^*$  is upper-bounded by the sum of  
 255 these two errors. This establishes that, as both approximation and estimation errors decrease during  
 256 training, the learned distribution converges toward the true WB in a distributional sense.

### 258 3.3 MEASURING $n$ -MODALITY ALIGNMENT WITH BARYCENTER POLYTOPE VOLUME

260 To bind modalities to the WB space while preserving the global structures of  $n$ -dimensional (where  
 261  $n = K + 1$  is the number of modalities) multimodal data, we introduce a geometric structure  
 262 called the *barycenter polytope*. As illustrated in Fig. 3, the polytope takes the WB embedding  $\mathbf{b}$   
 263 as the apex and is spanned by the vectors from the origin to  $\mathbf{b}$  and the modality-to-barycenter gap  
 264 vectors  $\mathbf{r}_k = \mathbf{b} - \mathbf{m}_k$  for all non-anchor modalities  $\mathbf{m}_k$  ( $k \geq 1$ ). Owing to its unique composition,  
 265 the polytope volume quantifies two aspects of multimodal features: (1) how closely non-anchor  
 266 modalities align with the barycenter, and (2) the inter-modality discrepancy across modalities.

267 We use the volume of this barycenter polytope to measure  $n$ -modality alignment. Intuitively, a  
 268 smaller volume indicates that the modalities are more tightly clustered around the WB embedding  
 269 and more consistent with one another, suggesting stronger alignment. Conversely, a larger volume  
 reflects greater distances from the barycenter and increased inter-modality discrepancy.

Given the WB embedding vector  $\mathbf{b}$  (i.e., the vector from the origin to the barycenter) and gap vectors  $\{\mathbf{r}_k\}_{k=1}^K$  (i.e., the vectors from each non-anchor modality to the barycenter), we define the matrix  $\mathbf{R} = (\mathbf{b}, \mathbf{r}_1, \dots, \mathbf{r}_K)$ , representing the set of vectors spanning the barycenter polytope. Then the square of the polytope's volume can be computed according to [Gantmacher \(1959\)](#) as:

$$V^2(\mathbf{b}, \mathbf{r}_{0:K}) := V^2(\mathbf{b}, \mathbf{r}_1, \dots, \mathbf{r}_K) = \det(\mathbf{R}^T \mathbf{R}) = \begin{vmatrix} \langle \mathbf{b}, \mathbf{b} \rangle & \langle \mathbf{b}, \mathbf{r}_1 \rangle & \dots & \langle \mathbf{b}, \mathbf{r}_K \rangle \\ \langle \mathbf{r}_1, \mathbf{b} \rangle & \langle \mathbf{r}_1, \mathbf{r}_1 \rangle & \dots & \langle \mathbf{r}_1, \mathbf{r}_K \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{r}_K, \mathbf{b} \rangle & \langle \mathbf{r}_K, \mathbf{r}_1 \rangle & \dots & \langle \mathbf{r}_K, \mathbf{r}_K \rangle \end{vmatrix}. \quad (8)$$

**Remark.** The volume offers a metric for measuring global alignment to the barycenter while retaining inter-modal interactions among non-anchor modalities, going beyond simple pairwise comparisons. The volume is computed as the determinant of a  $n \times n$  matrix, where the number of modalities  $n$  is typically small (e.g., 3 or 4) and much less than the embedding dimension  $D$ , introducing negligible computational overhead, which is further validated in our experiments.

### 3.4 BARYCENTER-BASED LOSS FUNCTIONS

To enforce tight and balanced multimodal binding in the Wasserstein barycenter space, we propose a barycenter-anchored volumetric contrastive (BVC) loss that leverages the volume of the barycenter polytope defined by the barycenter and modality-to-barycenter gap vectors.

Given multimodal inputs  $\{\mathbf{m}_0^{(i)}, \mathbf{m}_1^{(i)}, \dots, \mathbf{m}_K^{(i)}\}_{i=1}^B$ , where  $B$  is the batch size and  $\mathbf{m}_0^{(i)}$  is the anchor modality. The barycenter is computed by  $\mathbf{b}^{(i)} = T_\theta(\mathbf{m}_0^{(i)})$ , and the modality-to-barycenter gap vectors are defined as  $\mathbf{r}_k^{(i)} = \mathbf{b}^{(i)} - \mathbf{m}_k^{(i)}$  for  $k = 1, \dots, K$ , aggregated as  $\mathbf{r}_{0:K}^{(i)} = \{\mathbf{r}_1^{(i)}, \dots, \mathbf{r}_K^{(i)}\}$ . We treat each  $(\mathbf{b}^{(i)}, \mathbf{r}_{0:K}^{(i)})$  as a *positive pair*, and construct two sets of *negative pairs*: (i) by fixing  $\mathbf{b}^{(i)}$  and pairing it with  $\mathbf{r}_{0:K}^{(j)}$  for  $j \neq i$ , and (ii) by fixing  $\mathbf{r}_{0:K}^{(i)}$  and pairing it with  $\mathbf{b}_j$  for  $j \neq i$ .

**Barycenter-anchored volumetric contrastive (BVC) loss.** The BVC loss contrasts small volumes for positive pairs against large volumes for negative ones, and is formulated as

$$\mathcal{L}_{\text{BVC}} = -\frac{1}{2} \mathbb{E}_i \left[ \log \frac{\exp(-V(\mathbf{b}^{(i)}, \mathbf{r}_{0:K}^{(i)})/\tau)}{\sum_j \exp(-V(\mathbf{b}^{(i)}, \mathbf{r}_{0:K}^{(j)})/\tau)} \right] - \frac{1}{2} \mathbb{E}_i \left[ \log \frac{\exp(-V(\mathbf{b}^{(i)}, \mathbf{r}_{0:K}^{(i)})/\tau)}{\sum_j \exp(-V(\mathbf{b}^{(j)}, \mathbf{r}_{0:K}^{(i)})/\tau)} \right], \quad (9)$$

where  $V(\mathbf{b}, \mathbf{r}_{0:K})$  denotes the volume of the polytope formed by the barycenter and its modality gap vectors, whose square is computed as (8) and  $\tau$  is the temperature parameter.

The BVC loss encourages non-anchor modalities to move closer to the common barycenter, ensuring tight binding around a shared modality-agnostic representation. Simultaneously, it provides a geometric guarantee that the inter-modality gap vectors converge, reducing discrepancies among modalities. To prevent trivial collapse of the polytope volume caused by shrinking vector magnitudes, we normalize each gap vector  $\mathbf{r}_k$  to unit length. This normalization ensures that volume minimization effectively promotes alignment of vector directions rather than their norms, leading to a stable and balanced multimodal binding in the Wasserstein barycenter space.

**Data-anchor matching (DAM) loss.** We also introduce an auxiliary data-anchor matching (DAM) loss, which encourages the model to infer whether a pair of data and anchor is matched or not. This loss is commonly used in multimodal retrieval ([Li et al., 2021b; Chen et al., 2023](#)) to enforce fine-grained semantic alignment distinguishing matched and mismatched data pairs. To enable such matching, we can reuse the anchor-modality encoder as the multimodal encoder and integrate it with cross-attention layers, which take the WB embedding  $\mathbf{b}$  as input and are conditioned on the unpooled concatenated embeddings  $\mathbf{m}_{1:K}$  from non-anchor modalities. The output feature from the multimodal encoder is then passed through a two-layer MLP to produce binary predictions  $p_m$ . To construct informative in-batch negative pairs, we adopt a hard negative mining strategy following ([Li et al., 2021a; Chen et al., 2023](#)). The DAM loss is formulated as follows, where  $y = 1$  if the barycenter anchor and non-anchor modalities are matched, and  $y = 0$  otherwise:

$$\mathcal{L}_{\text{DAM}} = \mathbb{E}_{(\mathbf{b}, \mathbf{m}_{1:K}) \sim (\mathbb{Q}, \mathbb{P}_{1:K})} [y \log p_m(\mathbf{b}, \mathbf{m}_{1:K}) + (1 - y) \log(1 - p_m(\mathbf{b}, \mathbf{m}_{1:K}))] \quad (10)$$

324 **The overall pre-training loss.** The overall pre-training objective  $\mathcal{L}$  combines the three proposed  
 325 losses,  $\mathcal{L}_{\text{MWB}}$ ,  $\mathcal{L}_{\text{BVC}}$ , and  $\mathcal{L}_{\text{DAM}}$ , as follows:  
 326

$$\mathcal{L} := \mathcal{L}_{\text{MWB}} + \mathcal{L}_{\text{BVC}} + \alpha \mathcal{L}_{\text{DAM}}, \quad (11)$$

328 Particularly, the potentials  $f_{\omega_{0:K}}$  in  $\mathcal{L}_{\text{MWB}}$  are optimized via maximization, and are trained in an  
 329 alternating manner against the minimization of the remaining network parameters.  
 330

## 331 4 EXPERIMENTS

### 332 4.1 SETUP

333 We adopt VAST framework (Chen et al., 2023) as backbone, with BERT-B for text, BEATs for  
 334 audio, and EVA-CLIP-ViT-G (Sun et al., 2023) for visual encoding. In total, the model comprises  
 335 approximately 1B parameters. Unlike the VAST framework, we discard its modality fusion layers  
 336 and introduce lightweight MLPs for barycenter optimization. **We continue pretraining for one epoch**  
 337 **based on VAST using our proposed loss functions on the VAST150k dataset (Chen et al., 2023), a**  
 338 **subset of the VAST27M comprising 150k samples.** The temperature factor  $\tau$  is 0.07 and the trade-  
 339 off parameter in (11) is set as  $\alpha = 0.1$ , following (Chen et al., 2023). The barycenter weights  $\lambda_k$   
 340 ( $k \in \bar{K}$ ) are uniformly set as  $1/n$ , where  $n$  is the number of modalities.  
 341

342 We use benchmarks spanning diverse modalities to evaluate the BaryBind’s capability in multimodal  
 343 understanding across retrieval and classification tasks. These benchmarks include: (i) three-modality  
 344 datasets such as DiDeMo (Anne Hendricks et al., 2017) and ActivityNet (Caba Heilbron et al., 2015),  
 345 where video serves as the primary modality while audio and text provide auxiliary cues; (ii) four-  
 346 modality datasets like MSR-VTT (Xu et al., 2016) and VATEX (Wang et al., 2019), covering video  
 347 (V), audio (A), text (T), and subtitles (S); and (iii) audio-centered dataset VGGSound (Chen et al.,  
 348 2020), where audio plays the dominant role and complementary information is also available in  
 349 visual and textual forms. T-VAS denotes that the WB embedding is derived from the text modality,  
 350 while the concatenated VAS embeddings serve as the condition for the multimodal encoder.  
 351

### 352 4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

353 **Zero-shot video/audio classification.** We first  
 354 evaluate BaryBind on VGGSound5K to as-  
 355 sess its multimodal understanding ability in  
 356 the zero-shot setting, particularly when jointly  
 357 modeling video and audio signals. As shown  
 358 in Tab. 1, BaryBind achieves the best top-1  
 359 and top-5 accuracy across all modality con-  
 360 figurations, reaching 55.6% Acc@1 and 83.4  
 361 % Acc@5 under the A+V setting. It signifi-  
 362 cantly outperforms the VAST baseline (48.1%  
 363 / 79.6%) and other strong competitors such as  
 364 OmniBind (45.4% / 73.2%). This demon-  
 365 strates its strong generalization ability across modal-  
 366 ities. The substantial gains with A+V highlight  
 367 the enhanced performance of cross-modal un-  
 368 derstanding, enabled by the BVC loss design  
 369 which inherently reduces inter-modality gaps  
 370 and preserves the global geometry of multi-  
 371 modal data by aligning multimodal representation based on the barycenter polytope volume.  
 372

Table 1: **Zero-shot video/audio classification** re-  
 373 sults on VGGSound5K. Results from our baseline  
 374 VAST and BaryBind are highlighted accordingly.

Method	Modality	Acc@1	Acc@5
ImageBind (Girdhar et al., 2023)	A	31.6	58.7
ImageBind (Girdhar et al., 2023)	V	37.9	65.9
LanguageBind (Zhu et al., 2024)	A	34.1	62.8
LanguageBind (Zhu et al., 2024)	V	39.6	64.5
GRAM (Cicchetti et al., 2025b)	V	43.1	71.8
GRAM (Cicchetti et al., 2025b)	A+V	42.3	74.5
Triangle (Cicchetti et al., 2025a)	A+V	44.8	80.0
OmniBind (Wang et al., 2025)	A	41.7	70.8
OmniBind (Wang et al., 2025)	V	45.4	73.2
OmniBind (Wang et al., 2025)	A+V	46.2	76.2
VAST (Chen et al., 2023)	A	40.3	71.7
VAST (Chen et al., 2023)	V	46.3	72.7
VAST (Chen et al., 2023)	A+V	48.1	79.6
BaryBind (Ours)	A	45.7	75.2
BaryBind (Ours)	V	48.3	76.4
BaryBind (Ours)	A+V	<b>55.6</b>	<b>83.4</b>

375 On the other hand, BaryBind improves the performance of the weaker audio-only modality clas-  
 376 sification (A) from 40.3% (VAST) to 45.7% for Acc@1, indicating its ability to alleviate under-  
 377 optimization of weaker modalities. These results confirm that BaryBind learns more balanced mul-  
 378 timodal representations by aligning modalities via our barycenter-based binding strategy.  
 379

380 **Cross-modal retrieval.** We evaluate BaryBind on zero-shot text-to-video (T2V) and video-to-text  
 381 (V2T) retrieval across four benchmarks. As shown in Tab. 2, BaryBind consistently achieves state-  
 382 of-the-art results under all modality configurations. For instance, under the T-VA setting, it achieves  
 383

378  
379  
380  
381 Table 2: **Zero-shot** text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall  
382 at 1 score (R@1). Results from our baseline VAST and BaryBind are highlighted accordingly.  
383  
384  
385  
386  
387  
388  
389  
390  
391

Methods	Modality	MSR-VTT		DiDeMo		ActivityNet		VATEX	
		T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
VideoCoCa (Yan et al., 2022)	T-V	34.3	64.7	-	-	34.5	33.0	53.2	73.6
X-CLIP (Ma et al., 2022)	T-V	46.1	46.8	45.2	42.3	44.3	42.6	-	-
ImageBind (Girdhar et al., 2023)	T-V	36.8	-	-	-	-	-	-	-
ViCLIP (Wang et al., 2024c)	T-V	42.4	41.3	18.4	27.9	15.1	24.0	-	-
VideoPrism-b (Zhao et al., 2024)	T-V	51.4	50.2	-	-	49.6	47.9	62.5	77.1
LanguageBind (Zhu et al., 2024)	T-V	44.8	40.9	39.9	39.8	41.0	39.1	-	-
InternVL (Chen et al., 2024)	T-V	46.3	42.4	43.7	42.2	45.1	42.4	66.8	69.3
OmniBind (Wang et al., 2025)	T-V	47.4	45.2	43.5	42.6	44.3	40.8	-	-
NarVid (Hur et al., 2025)	T-V	51.8	50.3	52.4	50.5	51.8	46.6	73.8	76.3
Video-ColBERT (Reddy et al., 2025)	T-V	51.9	48.8	51.7	50.1	52.7	47.8	72.4	73.7
<b>GRAM</b> (Cicchetti et al., 2025b)	T-VAS	54.2	51.6	-	-	-	-	83.2	81.9
VAST (Chen et al., 2023)	T-VA	49.3	43.7	49.5	48.2	51.4	46.8	80.0	77.3
<b>VAST</b> (Chen et al., 2023)	T-VAS	50.9	47.9	-	-	-	-	82.1	78.7
BaryBind (Ours)	T-V	53.4	51.3	54.3	52.7	59.3	51.6	82.3	79.8
BaryBind (Ours)	T-VA	54.5	52.0	<b>55.3</b>	<b>52.8</b>	<b>59.6</b>	<b>53.9</b>	84.2	81.3
BaryBind (Ours)	T-VAS	<b>56.3</b>	<b>53.6</b>	-	-	-	-	<b>84.6</b>	<b>83.5</b>
InternVideo2-6B (Wang et al., 2024d)	T-VA	54.9	49.1	55.7	51.6	61.2	52.8	82.7	76.4

392  
393  
394  
395  
396  
397  
398  
399  
400 Table 3: **Finetuning** text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall  
401 at 1 score (R@1). Results from our baseline VAST and BaryBind are highlighted accordingly.  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416

Methods	Modality	MSR-VTT		DiDeMo		ActivityNet		VATEX	
		T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
CLIP4Clip (Luo et al., 2021)	T-V	45.6	45.9	43.0	43.6	40.3	41.6	63.0	78.3
InternVideo-L (Wang et al., 2022)	T-V	53.1	54.4	57.9	59.1	62.2	62.8	69.8	80.6
HiTeA (Ye et al., 2022)	T-V	46.8	-	56.5	-	-	-	-	-
mPLUG-2 (Xu et al., 2023)	T-V	53.1	-	56.4	-	-	-	-	-
TEFAL (Ibrahim et al., 2023)	T-VA	52.0	-	-	-	-	-	61.0	-
ViCLIP (Wang et al., 2024c)	T-V	52.5	51.8	49.4	50.2	49.8	48.1	-	-
T-MASS (Wang et al., 2024a)	T-VA	52.7	-	53.3	-	-	-	65.6	-
VALOR-L (Liu et al., 2024)	T-VAS	54.4	-	57.6	-	63.4	-	76.9	-
VideoCLIP-XL (Wang et al., 2024b)	T-V	54.6	54.0	62.3	62.7	58.4	59.2	-	-
TempMe (Shen et al., 2025)	T-V	49.0	47.6	48.0	48.4	44.9	45.3	69.6	71.8
VAST (Chen et al., 2023)	T-VA	55.8	57.6	65.6	62.0	68.8	66.7	86.9	84.1
<b>VAST</b> (Chen et al., 2023)	T-VAS	<b>56.6</b>	<b>57.6</b>	-	-	-	-	87.5	84.0
BaryBind (Ours)	T-V	57.4	57.8	67.2	64.6	67.3	65.2	85.0	82.4
BaryBind (Ours)	T-VA	60.3	60.8	<b>68.5</b>	<b>64.4</b>	<b>72.1</b>	<b>69.4</b>	87.4	<b>84.8</b>
BaryBind (Ours)	T-VAS	<b>64.6</b>	<b>65.2</b>	-	-	-	-	<b>88.4</b>	84.6

420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
55.3 and 52.8 R@1 on DiDeMo (T2V/V2T), outperforming all prior methods. With T-VAS, BaryBind  
432 further sets new records on MSR-VTT (56.3/53.6) and VATEX (84.6/83.5), demonstrating  
433 strong retrieval performance across both directions. In addition to these overall gains, BaryBind  
434 significantly reduces the T2V/V2T performance gap (e.g., 5.7 on ActivityNet vs. 9.6 for the VAST  
435 baseline), indicating improved bidirectional alignment. This suggests that BaryBind learns more  
436 generalizable and balanced cross-modal representations, which can be attributed to the WB binding  
437 mechanism that geometrically unifies diverse modalities into a modality-agnostic latent space.

438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220

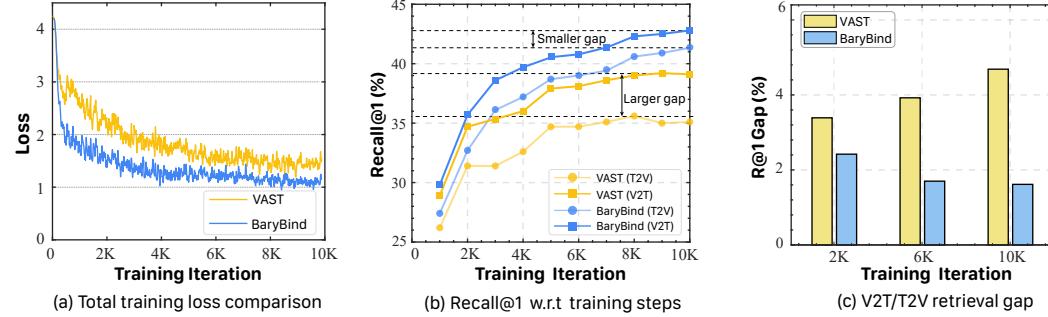


Figure 4: **Training dynamics comparison with cosine-based VAST:** (a) Total training loss showing faster convergence for BaryBind. (b) Recall@1 (R@1) evolution on MSR-VTT under T2V and V2T retrieval tasks. BaryBind achieves higher retrieval accuracy and better modal balance (smaller R@1 gap). (c) V2T/T2V retrieval gaps throughout training.

#### 4.3 ABLATION STUDIES

**Training dynamics comparison.** We compare the training (from scratch) behaviors of BaryBind and the baseline VAST to assess the impact of our binding strategy and loss design. As shown in Fig. 4, BaryBind exhibits consistently superior training dynamics. In (a), the total loss drops faster and stabilizes earlier, suggesting that our MWB and BVC losses facilitate more efficient optimization. In (b), BaryBind consistently achieves higher R@1 accuracy for both T2V and V2T retrieval, while maintaining a smaller performance gap between the two directions, suggesting more symmetric multimodal representations. (c) demonstrates that BaryBind rapidly reduces the gap between V2T and T2V performance, converging toward a balanced retrieval behavior. This improvement stems from the joint effect of MWB and the BVC loss, which together promote consistent alignment of modality embeddings toward a unified barycenter. Such training behavior reflects the effectiveness of our barycentric modeling in encouraging modality-agnostic representation learning.

Table 4: **Ablation study on loss functions.** We report top-1 classification accuracy (Acc@1) on VGGSound5K and top-1 retrieval recall at 1 score (R@1) on MSR-VTT. Key improvements from MWB, BVC, and their combination are highlighted correspondingly.

TV+TA CL	MWB	BVC	DAM	Classification			Retrieval	
				VGGSound			MSR-VTT	
				A	V	V+A	T2V	V2T
✓	✗	✗	✗	38.1	44.5	46.8	46.8	40.1
✓	✗	✗	✓	40.3	46.3	48.1	49.3	43.7
✓	✓	✗	✗	43.6	45.6	47.6	48.8	46.2
✓	✓	✗	✓	44.3	46.8	49.8	49.7	48.3
✗	✗	✓	✗	42.9	46.3	50.3	50.6	46.4
✗	✗	✓	✓	43.4	47.2	52.6	51.5	46.8
✗	✓	✓	✗	45.2	47.8	54.2	53.2	50.6
✗	✓	✓	✓	45.7	48.3	55.6	54.5	52.0

**Comparison of loss functions.** To evaluate the contribution of each loss component, we conduct an ablation study on VGGSound (Chen et al., 2020) for audio classification and MSR-VTT (Xu et al., 2016) for text-video retrieval under a tri-modal setting (text, video, audio). The baseline adopts pairwise cosine contrastive losses (TV+TA CL), while VAST (Chen et al., 2023) additionally includes the data-anchor matching (DAM) loss. As summarized in Tab. 4, introducing the multimodal Wasserstein barycenter (MWB) loss consistently improves performance across tasks—for example, boosting audio-only classification from 40.3% to 44.3% and V2T retrieval from 43.7% to 48.3%. MWB constructs a barycentric semantic anchor to filter out modality-specific biases from the original anchor (e.g., text), leading to a more balanced and modality-agnostic anchor. The barycenter volume contrastive (BVC) loss further enhances global geometric consistency by preserving the relative structures among modalities, enabling full-modality gains such as 55.6% Acc@1 in audio classification and 54.5/52.0 R@1 for T2V/V2T retrieval. Complementarily, the DAM loss leverages

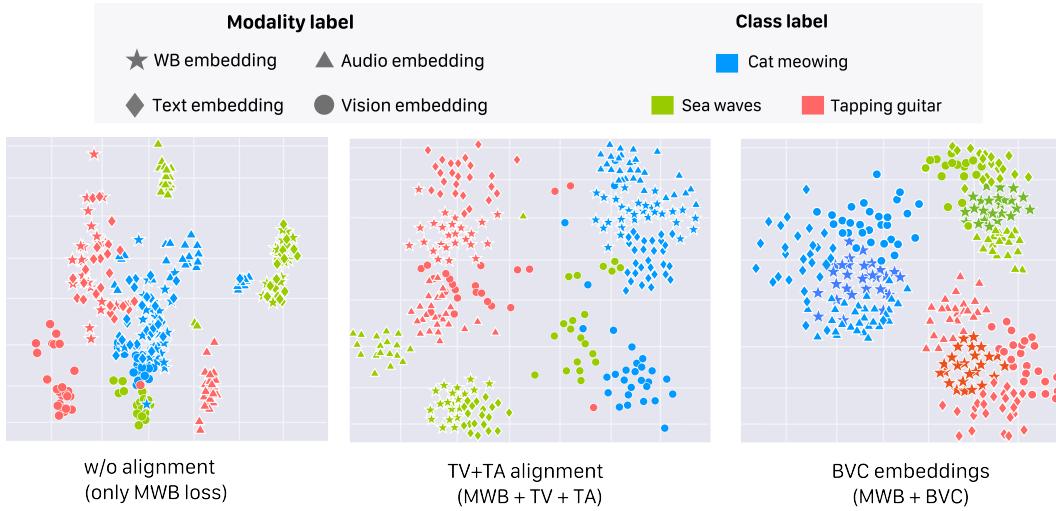


Figure 5: **Visualization of multimodal before and after alignment.** Embeddings before (left) and after (right) applying the proposed BVC loss. The BVC loss promotes convergence of modality-specific embeddings toward the Wasserstein barycenter, reducing inter-modality gaps and forming compact, modality-agnostic clusters, which highlights improved multimodal alignment.

instance-level supervision to distinguish matched and mismatched cross-modal pairs. Overall, the combination of MWB, BVC, and DAM facilitates a collaborative effect by aligning all modalities toward a shared semantic barycenter, yielding more balanced and generalizable multimodal representations.

**Visualization of multimodal embeddings before and after alignment.** To intuitively illustrate the effect of our barycenter-based alignment, we visualize the embeddings of three categories (cat meowing, sea waves, tapping guitar) across audio, text, and vision modalities from VGGSound (Chen et al., 2020), as shown in Figure 5. Without alignment (left), embeddings of different modalities are scattered, exhibiting large inter-modality gaps even within the same class. Incorporating pairwise alignment (middle) improves class clustering but still reveals modality-specific separations. In contrast, with our proposed BVC loss (right), embeddings of all modalities converge tightly around their class-wise Wasserstein barycenters, leading to compact intra-class structures and reduced inter-modality discrepancies. This clearly demonstrates that the volumetric constraint effectively promotes modality-agnostic alignment around a unified barycenter space.

## 5 CONCLUSION

We introduced BaryBind, a novel multimodal learning framework that aligns multiple modalities to a shared multimodal Wasserstein barycenter space. Unlike traditional anchor-based alignment strategies, BaryBind leverages the barycenter to model a modality-agnostic semantic distribution, providing a principled and geometry-aware alignment target. By constructing a barycenter polytope and leveraging its volume as a global alignment metric, BaryBind captures higher-order interactions among modalities and quantifies alignment quality beyond pairwise similarities. The proposed volumetric contrastive loss further encourages all modalities to converge toward the barycenter while preserving inter-modal structure. Extensive experiments on retrieval and classification tasks demonstrate that BaryBind learns more balanced and generalizable multimodal representations, outperforming state-of-the-art approaches across diverse benchmarks. We hope this work opens new directions for scalable multimodal learning and inspires the development of more interpretable and geometry-grounded multimodal methods.

## REFERENCES

Martial Aguech and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

540 Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell.  
 541 Localizing moments in video with natural language. In *IEEE/CVF International Conference on*  
 542 *Computer Vision*, pp. 5803–5812, 2017.

543

544 Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning:  
 545 A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):  
 546 423–443, 2018.

547

548 Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein  
 549 barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

550

551 Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet:  
 552 A large-scale video benchmark for human activity understanding. In *IEEE/CVF Conference on*  
 553 *Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

554

555 Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-  
 556 visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*,  
 557 pp. 721–725. IEEE, 2020.

558

559 Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. VAST:  
 560 A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural*  
 561 *Information Processing Systems*, 2023.

562

563 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
 564 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning  
 565 for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern*  
 566 *Recognition*, pp. 24185–24198, 2024.

567

568 Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal  
 569 alignment beyond cosine similarity. *arXiv preprint arXiv:2509.24734*, 2025a.

570

571 Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multi-  
 572 modal representation learning and alignment. In *International Conference on Learning Repre-*  
 573 *sentations*, 2025b.

574

575 Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International*  
 576 *Conference on Machine Learning*, pp. 685–693. PMLR, 2014.

577

578 Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning au-  
 579 dio concepts from natural language supervision. In *IEEE International Conference on Acoustics,*  
 580 *Speech and Signal Processing*, pp. 1–5, 2023.

581

582 Felix R Gantmacher. Matrix theory. *Chelsea, New York*, 21:6, 1959.

583

584 Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand  
 585 Joulin, and Ishan Misra. ImageBind one embedding space to bind them all. In *IEEE/CVF Con-*  
 586 *ference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

587

588 Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena  
 589 Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar,  
 590 et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural*  
 591 *Information Processing Systems*, 33:21271–21284, 2020.

592

593 Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to  
 594 image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Pro-*  
 595 *cessing*, pp. 976–980. IEEE, 2022.

596

597 Chan Hur, Jeong hun Hong, Dong hun Lee, Dabin Kang, Semin Myeong, Sang hyo Park, and  
 598 Hyeyoung Park. Narrating the video: Boosting text-video retrieval via comprehensive utilization  
 599 of frame-level captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*,  
 600 2025.

594 Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mo-  
 595 hamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment.  
 596 *IEEE/CVF International Conference on Computer Vision*, pp. 12020–12030, 2023.

597

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

602 Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37,  
 603 pp. 199–201, 1942.

604

605 Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generat-  
 606 ing captions for audios in the wild. In *Annual Meeting of the Association for Computational  
 607 Linguistics*, pp. 119–132, 2019.

608

609 Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny  
 610 Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal  
 611 transport. *arXiv preprint arXiv:2402.03828*, 2024a.

612 Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Anas-  
 613 tasis Kratsios, Evgeny Burnaev, and Aleksandr Korotin. Energy-guided continuous entropic  
 614 barycenter estimation for general costs. *Advances in Neural Information Processing Systems*,  
 615 37:107513–107546, 2024b.

616

617 Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2  
 618 barycenter estimation without minimax optimization. In *International Conference on Learning  
 619 Representations*, 2021.

620

621 Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan,  
 622 and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the  
 623 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26647–26657, 2024.

624

625 Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven  
 626 Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum  
 627 distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021a.

628

629 Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong,  
 630 and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with  
 631 momentum distillation. In *Advances in Neural Information Processing Systems*, 2021b.

632

633 Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M Solomon. Continuous regularized  
 634 wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pp.  
 17755–17765, 2020.

635

636 Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang.  
 637 Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions  
 638 on Pattern Analysis and Machine Intelligence*, 2024.

639

640 Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip:  
 641 An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304,  
 642 2021.

643

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

647

648 Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end  
 649 multi-grained contrastive learning for video-text retrieval. In *ACM International Conference on  
 650 Multimedia*, pp. 638–647, 2022.

648 Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumb-  
 649 ley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio caption-  
 650 ing dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech,*  
 651 *and Language Processing*, 2024.

652 Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predic-  
 653 tive coding. *arXiv preprint arXiv:1807.03748*, 2018.

654 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-  
 655 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya  
 656 Sutskever. Learning transferable visual models from natural language supervision. In *Interna-  
 657 tional Conference on Machine Learning*, 2021.

658 Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno  
 659 Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextu-  
 660 alized late interaction for text-to-video retrieval. In *IEEE/CVF Conference on Computer Vision*  
 661 *and Pattern Recognition*, 2025.

662 Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang,  
 663 Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion  
 664 priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22841–  
 665 22852, 2025.

666 Legi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao,  
 667 and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. In  
 668 *International Conference on Learning Representations*, 2025.

669 Maurice Sion. On general minimax theorems. 1958.

670 Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training  
 671 techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

672 Xiaole Tang, Xiaoyi He, Xiang Gu, and Jian Sun. Baryir: Learning multi-source unified representa-  
 673 tion in continuous barycenter space for generalizable all-in-one image restoration. *arXiv preprint*  
 674 *arXiv:2505.21637*, 2025.

675 Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail A. Dianat, Majid Rabbani, Raghu-  
 676 veer M. Rao, and Zhiqiang Tao. Text Is MASS: Modeling as stochastic embedding for text-video  
 677 retrieval. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16551–16560,  
 678 2024a.

679 Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advanc-  
 680 ing long description understanding for video clip models. In *Conference on Empirical Methods*  
 681 *in Natural Language Processing*, pp. 16061–16075, 2024b.

682 Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex:  
 683 A large-scale, high-quality multilingual dataset for video-and-language research. In *IEEE/CVF*  
 684 *International Conference on Computer Vision*, pp. 4581–4591, 2019.

685 Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan  
 686 Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and  
 687 discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

688 Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan  
 689 Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal under-  
 690 standing and generation. In *International Conference on Learning Representations*, 2024c.

691 Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng,  
 692 Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang,  
 693 Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling video foundation models for mul-  
 694 timodal video understanding. *ArXiv preprint: arXiv:2403.15377*, 2024d.

702 Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang  
 703 Zhao, and Zhou Zhao. Omnidbind: Large-scale omni multimodal representation via binding  
 704 spaces. In *International Conference on Learning Representations*, 2025.

705

706 Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov.  
 707 Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption  
 708 augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*,  
 709 2023.

710 Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi,  
 711 Qiuchen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren  
 712 Zhou. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In  
 713 *International Conference on Machine Learning*, 2023.

714 Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging  
 715 video and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.  
 716 5288–5296, 2016.

717

718 Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui  
 719 Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*  
 720 preprint: *arXiv:2212.04979*, 2022.

721 Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA:  
 722 Hierarchical temporal-aware video-language pre-training. *IEEE/CVF International Conference*  
 723 *on Computer Vision*, pp. 15359–15370, 2022.

724

725 Yiyuan Zhang, Handong Li, Jing Liu, and Xiangyu Yue. Explore the limits of omni-modal pretrain-  
 726 ing at scale. *arXiv preprint arXiv:2406.09412*, 2024.

727 Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun,  
 728 Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming  
 729 Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing  
 730 Gong. Videoprism: A foundational visual encoder for video understanding. In *International*  
 731 *Conference on Machine Learning*, 2024.

732 Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang,  
 733 Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. LanguageBind:  
 734 Extending video-language pretraining to n-modality by language-based semantic alignment. In  
 735 *International Conference on Learning Representations*, 2024.

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A THEORETICAL RESULTS**  
 757

758 **A.1 PROOF OF THEOREM 1**  
 759

760 *Proof.* Beyond the WB problem, we deduce based on the duality of with general costs  $C_k(x, y)$ . In  
 761 this sense, we write the dual reformulation of the multimodal OT barycenter problem (3):  
 762

$$763 \mathcal{L}^* = \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sup_{f_0, \dots, f_K \in \mathcal{C}(\mathcal{M}_B)} \underbrace{\sum_{k=1}^K \lambda_k \left\{ \int_{\mathcal{M}_k} f_k^{C_k}(\mathbf{m}_k) d\mathbb{P}_k(\mathbf{m}_k) + \int_{\mathcal{M}_B} f_k(\mathbf{b}) d\mathbb{Q}(\mathbf{b}) \right\}}_{\triangleq \tilde{\mathcal{F}}(\mathbb{Q}, f_{0:K})}. \quad (12)$$

764  
 765  
 766  
 767  
 768  
 769

770 where  
 771

$$772 f_k^{C_k}(\mathbf{m}_k) = \inf_{\mathbf{b} \in \mathcal{M}_B} [C_k(\mathbf{m}_k, \mathbf{b}) - f_k(\mathbf{b})]. \quad (13)$$

773  
 774

775 We denote the expression under the inf and inf sup in (12) as functionals  $\mathcal{L} : \mathcal{C}(\mathcal{M}_B)^K$  and  $\tilde{\mathcal{F}} : \mathcal{P}(\mathcal{M}_B) \times \mathcal{C}(\mathcal{M}_B)^K$ , respectively. For simplicity, we also introduce the following notation  
 776

$$777 \bar{f} \triangleq \sum_{k=1}^K \lambda_k f_k \quad \text{and} \quad M \triangleq \inf_{\mathbf{b} \in \mathcal{M}_B} \bar{f}(\mathbf{b}) = \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \int_{\mathcal{M}_B} \bar{f}(\mathbf{b}) d\mathbb{Q}(\mathbf{b}), \quad (14)$$

778  
 779  
 780

781 where the equality follows from two fundamental observations: (a)  $M \leq \int \bar{f}(\mathbf{b}) d\mathbb{Q}(\mathbf{b})$  for any  $\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)$ , and (b)  $\bar{f}(\mathbf{b}) = \int \bar{f}(\mathbf{b}') d\delta_{\mathbf{b}}(\mathbf{b}')$  where  $\delta_{\mathbf{b}}$  represents the Dirac mass at  $\mathbf{b} \in \mathcal{M}_B$ .  
 782

783 Firstly, due to the compactness of  $\mathcal{M}_B$ , the space  $\mathcal{P}(\mathcal{M}_B)$  is compact with respect to the weak  
 784 topology. For fixed potentials  $f_{0:K} \in \mathcal{P}(\mathcal{M}_B)^K$  we have that  $\tilde{\mathcal{F}}(\cdot, f_{0:K})$  is linear, convex and  
 785 continuous. Secondly, for a fixed  $\mathbb{Q}$ , the functional  $\tilde{\mathcal{F}}(\mathbb{Q}, \cdot)$  is a concave due to the concavity of  
 786  $C$ -transform. These properties enable the application of Sion's minimax theorem (Sion (1958),  
 787 Theorem 3.4), which allows the interchange of the sup and inf in (12). Thus with (14) we obtain  
 788

$$789 \mathcal{L}^* = \sup_{f_0, \dots, f_K \in \mathcal{C}(\mathcal{M}_B)} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sum_{k=1}^K \lambda_k \left\{ \int_{\mathcal{M}_k} f_k^{C_k}(\mathbf{m}_k) d\mathbb{P}_k(\mathbf{m}_k) + \int_{\mathcal{M}_B} f_k(\mathbf{b}) d\mathbb{Q}(\mathbf{b}) \right\}$$

790  
 791  
 792  
 793  
 794  
 795  
 796  
 797  
 798  
 799  
 800  
 801

$$= \sup_{f_0, \dots, f_K \in \mathcal{C}(\mathcal{M}_B)} \left\{ \sum_{k=1}^K \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\mathbf{m}_k) d\mathbb{P}_k(\mathbf{m}_k) + \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \int_{\mathcal{M}_B} \bar{f}(\mathbf{b}) d\mathbb{Q}(\mathbf{b}) \right\}$$

$$= \sup_{f_0, \dots, f_K \in \mathcal{C}(\mathcal{M}_B)} \underbrace{\left\{ \sum_{k=1}^K \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\mathbf{m}_k) d\mathbb{P}_k(\mathbf{m}_k) + \inf_{\mathbf{b} \in \mathcal{M}_B} \bar{f}(\mathbf{b}) \right\}}_{\triangleq \tilde{\mathcal{L}}(f_{0:K})}. \quad (15)$$

802 Now we show that the sup in (15) can be restricted to potentials  $\tilde{f}_{0:K}$  which satisfy the congruence  
 803 condition  $\sum_{k=1}^K \lambda_k \tilde{f}_k = 0$ . It is enough to show that for every tuple  $f_{0:K}$  there exists a congruent  
 804 tuple  $\tilde{f}_{0:K} \in \mathcal{C}(\mathcal{M}_B)^K$  such that  $\tilde{\mathcal{L}}(\tilde{f}_{0:K}) \geq \tilde{\mathcal{L}}(f_{0:K})$ .  
 805

806 For this, we consider the congruent potentials given any tuple  $f_{0:K}$   
 807

$$808 (\tilde{f}_0, \dots, \tilde{f}_K) = \left( f_0, \dots, f_{K-1}, f_K - \frac{\bar{f}}{\lambda_K} \right). \quad (16)$$

809

Since  $\tilde{M} \triangleq \inf_{\mathbf{b} \in \mathcal{M}_B} \sum_{k=1}^K \lambda_k \tilde{f}_k = 0$ , we obtain

$$\begin{aligned}
\tilde{\mathcal{L}}(\tilde{f}_{0:K}) - \tilde{\mathcal{L}}(f_{0:K}) &= \lambda_K \int_{\mathcal{M}_k} \left( \tilde{f}_K^{C_K}(\mathbf{m}_K) - f_K^{C_K}(\mathbf{m}_K) \right) d\mathbb{P}_K(\mathbf{m}_K) - M \\
&= \lambda_K \int_{\mathcal{M}_k} \left[ \left( f_K - \frac{\bar{f}}{\lambda_K} \right)^{C_K}(\mathbf{m}_K) - f_K^{C_K}(\mathbf{m}_K) \right] d\mathbb{P}_K(\mathbf{m}_K) - M \\
&\geq \lambda_K \int_{\mathcal{M}_k} \left[ \left( f_K - \frac{M}{\lambda_K} \right)^{C_K}(\mathbf{m}_K) - f_K^{C_K}(\mathbf{m}_K) \right] d\mathbb{P}_K(\mathbf{m}_K) - M \\
&= \lambda_K \int_{\mathcal{M}_k} \frac{M}{\lambda_K} d\mathbb{P}_K(\mathbf{m}_K) - M = 0,
\end{aligned} \tag{17}$$

where the first inequality arises from the monotonicity of the  $C$ -transform, along with the fact  $\tilde{f}_K = f_K - \frac{\tilde{f}}{\lambda_K} \leq f_K - \frac{M}{\lambda_k}$ . The last equality follows from the definition of  $C$ -transform.

Finally, since  $\tilde{\mathcal{L}}(f_{0:K}) = \mathcal{L}(f_{0:K})$  for congruent potentials  $f_{0:K}$ , with (13) we obtain

$$\begin{aligned} \mathcal{L}^* &= \sup_{\sum_k \lambda_k f_k = 0} \mathcal{L}(f_{0:K}) = \sup_{\sum_k \lambda_k f_k = 0} \sum_{k=1}^K \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\mathbf{m}_k) d\mathbb{P}_k(\mathbf{m}_k), \\ &= \sup_{\sum_k \lambda_k f_k = 0} \sum_{k=1}^K \lambda_k \inf_{\mathbf{b} \in \mathcal{M}_B} \int_{\mathcal{M}_k} [C_k(\mathbf{m}_k - \mathbf{b}) - f_k(\mathbf{b})] d\mathbb{P}_k(\mathbf{m}_k). \quad (18) \end{aligned}$$

In practice, we replace each integral with an empirical expectation over the distribution  $\mathbb{P}_k$ , and adopt  $C_k(\mathbf{m}_k - \mathbf{b}) = \|\mathbf{m}_k - \mathbf{b}\|$ , which corresponds to the Wasserstein distance between the modality feature and the barycenter. We exchange the summation and the infimum since  $\mathbf{b}$  is shared across all terms and the objective is linear. We then sample the barycenter  $\mathbf{b}$  from a distribution  $\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)$ , and express the integral as an expectation over both  $\mathbf{m}_k \sim \mathbb{P}_k$  and  $\mathbf{b} \sim \mathbb{Q}$ :

$$\mathcal{L}^* = \sup_{\sum_k \lambda_k f_k = 0} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sum_{k=0}^K \lambda_k \mathbb{E}_{\substack{\mathbf{m}_k \sim \mathbb{P}_k \\ \mathbf{b} \sim \mathbb{Q}}} [\|\mathbf{m}_k - \mathbf{b}\| - f_k(\mathbf{b})], \quad (19)$$

which completes the proof.  $\square$

## A.2 PROOF OF THEOREM 3.1

*Proof.* Let  $\mathbb{P}_{0:K}$  denote the joint distribution over the multimodal tuples  $(\mathbf{m}_0, \dots, \mathbf{m}_K) =: \mathbf{m}_{0:K}$ , where  $\mathbf{m}_k \in \mathcal{M}_k$ . The barycenter map is a function  $T : \mathcal{M}_0 \rightarrow \mathcal{M}_B$  that acts on the anchor modality. For a set of potential functions  $f_{0:K}$  and a map  $T$ , we rewrite the total cost functional  $\mathcal{F}$  and its corresponding minimal cost functional  $\mathcal{L}$  as follows:

$$\mathcal{F}(f_{0:K}, T) \triangleq \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k (C_k(\mathbf{m}_k, T(\mathbf{m}_0)) - f_k(T(\mathbf{m}_0))) \right], \quad (20)$$

$$\mathcal{L}(f_{0:K}) \triangleq \inf_{T: \mathcal{M}_0 \rightarrow \mathcal{M}_B} \mathcal{F}(f_{0:K}, T). \quad (21)$$

The minimizer of functional  $\mathcal{E}$  is denoted as:

$$T^f \in \arg \inf_{T: \mathcal{M} \rightarrow \mathcal{M}_D} \mathcal{F}(\widehat{f}_{0:K}, T). \quad (22)$$

Given  $T^f : \mathcal{M}_0 \rightarrow \mathcal{M}_B$ , the functional  $\mathcal{L}$  in (21) can be written as

$$f(\hat{f}_{0;K}) \equiv \mathcal{F}(\hat{f}_{0;h}, T^f). \quad (23)$$

We can observe that the first gap  $\mathcal{E}_1$  is the difference between (20) and (23):

$$S(\hat{f} - \hat{T}) = T(\hat{f} - \hat{T}) - C(\hat{f} - \hat{T}) - T(\hat{f} - \hat{T}) - T(\hat{f} - T^f) \quad (24)$$

864 Before looking into the second gap  $\mathcal{E}_2$ , we recall the optimal value  $\mathcal{L}^*$  of the OT barycenter problem  
 865 and express the OT cost with Monge's formulation:  
 866

$$867 \quad \mathcal{L}^* \triangleq \sum_{k=0}^K \lambda_k \text{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}^*). \quad (25)$$

870 where  $\mathbb{Q}^*$  is the true barycenter distribution. By introducing the Monge's OT formulation with the  
 871 true OT map  $T^*$  that satisfies  $T_\#^* \mathbb{P}_0 = \mathbb{Q}^*$ , the expression can be rewritten as:  
 872

$$873 \quad \mathcal{L}^* = \sum_{k=0}^K \lambda_k \int_{\mathcal{M}_k} C_k(\mathbf{m}_k, T^*(\mathbf{m}_0)) d\mathbb{P}_k(\mathbf{m}_k). \quad (26)$$

875 Due to the congruence condition on the potentials  $\hat{f}_{0:K}$  and the property  $T_\#^* \mathbb{P}_0 = \mathbb{Q}^*$  for all  $k$ , we  
 876 have:  
 877

$$878 \quad \sum_{k=0}^K \lambda_k \mathbb{E}_{\mathbf{m}_k \sim \mathbb{P}_k} [\hat{f}_k(T^*(\mathbf{m}_0))] = \mathbb{E}_{\mathbf{b} \sim \mathbb{Q}^*} \left[ \sum_{k=0}^K \lambda_k \hat{f}_k(\mathbf{b}) \right] = 0. \quad (27)$$

881 This allows us to reformulate the optimal value  $\mathcal{L}^*$ . Using the definition of  $\mathcal{F}$  in (20), we find:  
 882

$$883 \quad \mathcal{L}^* = \sum_{k=0}^K \lambda_k \mathbb{E}_{\mathbf{m}_k \sim \mathbb{P}_k} [C_k(\mathbf{m}_0, T^*(\mathbf{m}_k))] - \underbrace{\mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k \hat{f}_k(T^*(\mathbf{m}_0)) \right]}_{=0 \text{ from (27)}} \\ 887 \quad = \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k (C_k(\mathbf{m}_k, T^*(\mathbf{m}_0)) - \hat{f}_k(T^*(\mathbf{m}_0))) \right] = \mathcal{F}(\hat{f}_{0:K}, T^*).$$

889 With (23) we derive the second gap  $\mathcal{E}_2$  can be written as  
 890

$$891 \quad \mathcal{E}_2 = \mathcal{L}^* - \mathcal{L}(\hat{f}_{0:K}) = \mathcal{F}(\hat{f}_{0:K}, T^*) - \mathcal{F}(\hat{f}_{0:K}, T^f). \quad (28)$$

893 We introduce the function  $g_k(\mathbf{m}_k, \mathbf{b}) \triangleq C_k(\mathbf{m}_k, \mathbf{b}) - \hat{f}_k(\mathbf{b})$ , which is assumed to be  $\beta$ -strongly  
 894 convex with respect to  $\mathbf{b}$ . Using this, we can rewrite our total cost functional  $\mathcal{F}$  from (20) as:  
 895

$$896 \quad \mathcal{F}(f_{0:K}, T) = \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k g_k(\mathbf{m}_k, T(\mathbf{m}_0)) \right]. \quad (29)$$

898 As a result of convexity, it follows that a necessary condition for  $T^f$  to minimize  $\mathcal{F}(f_{0:K}, T)$  is the  
 899 vanishing of its first variation, yielding  
 900

$$902 \quad \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k \nabla_{\mathbf{b}} g_k(\mathbf{m}_k, T^f(\mathbf{m}_0)) \right] = 0. \quad (30)$$

905 Now, we analyze the gap  $\mathcal{E}_1$  by applying the  $\beta$ -strong convexity of  $g_k(\mathbf{m}_k, \cdot)$ :  
 906

$$907 \quad \begin{aligned} \mathcal{E}_1 &= \mathcal{F}(\hat{f}_{0:K}, \hat{T}) - \mathcal{F}(\hat{f}_{0:K}, T^f) \\ &= \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k (g_k(\mathbf{m}_k, \hat{T}(\mathbf{m}_0)) - g_k(\mathbf{m}_k, T^f(\mathbf{m}_0))) \right] \\ &\geq \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \sum_{k=0}^K \lambda_k \left( \langle \nabla_{\mathbf{b}} g_k(\mathbf{m}_k, T^f(\mathbf{m}_0)), \hat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0) \rangle + \frac{\beta}{2} \|\hat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0)\|^2 \right) \right] \\ &= \mathbb{E}_{\mathbf{m}_{0:K} \sim \mathbb{P}_{0:K}} \left[ \left\langle \sum_{k=0}^K \lambda_k \nabla_{\mathbf{b}} g_k(\mathbf{m}_k, T^f(\mathbf{m}_0)), \hat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0) \right\rangle \right] + \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|\hat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0)\|^2] \\ &\stackrel{(30)}{=} 0 + \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|\hat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0)\|^2]. \end{aligned} \quad (31)$$

918 For the second gap  $\mathcal{E}_2$ , we conduct the same analysis and obtain  
 919

$$920 \quad 921 \quad \mathcal{E}_2 \geq \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|T^f(\mathbf{m}_0) - T^*(\mathbf{m}_0)\|^2]. \quad (32)$$

922 Now we sum the inequalities for  $\mathcal{E}_1$  (31) and  $\mathcal{E}_2$  (32):  
 923

$$924 \quad 925 \quad \mathcal{E}_1 + \mathcal{E}_2 \geq \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|\widehat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0)\|^2] + \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|T^f(\mathbf{m}_0) - T^*(\mathbf{m}_0)\|^2] \\ 926 \quad 927 \quad = \frac{\beta}{2} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|\widehat{T}(\mathbf{m}_0) - T^f(\mathbf{m}_0)\|^2 + \|T^f(\mathbf{m}_0) - T^*(\mathbf{m}_0)\|^2] \\ 928 \quad 929 \quad \geq \frac{\beta}{4} \mathbb{E}_{\mathbf{m}_0 \sim \mathbb{P}_0} [\|\widehat{T}(\mathbf{m}_0) - T^*(\mathbf{m}_0)\|^2] = \frac{\beta}{4} W_2^2(\widehat{T}_\# \mathbb{P}_0, T^*_\# \mathbb{P}_0) = \frac{\beta}{4} W_2^2(\widehat{T}_\# \mathbb{P}_0, \mathbb{Q}^*). \quad (33)$$

□

### 934 A.3 ANALYSIS ON BARYCENTER POLYTOPE VOLUME

935 Given the matrix containing vectors spanning the barycenter polytope  
 936

$$937 \quad \mathbf{R} = (\mathbf{b}, \mathbf{r}_1, \dots, \mathbf{r}_K),$$

938 we discuss the geometric meaning of barycenter polytope volumes for  $n = 2$  and  $n = 3$  modalities.  
 939  $\langle \cdot, \cdot \rangle$  denotes the cosine similarity between two normalized vectors, so that  
 940

$$941 \quad \langle \mathbf{b}, \mathbf{b} \rangle = \langle \mathbf{r}_1, \mathbf{r}_1 \rangle = \langle \mathbf{r}_2, \mathbf{r}_2 \rangle = 1. \quad (34)$$

942 We define the cosine of the angles between these vectors as  
 943

$$944 \quad \cos \theta := \langle \mathbf{b}, \mathbf{r}_1 \rangle, \quad \cos \beta := \langle \mathbf{b}, \mathbf{r}_2 \rangle, \quad \cos \gamma := \langle \mathbf{r}_1, \mathbf{r}_2 \rangle. \quad (35)$$

945 For the special case  $n = 2$ , the squared volume can be expressed as the determinant  
 946

$$947 \quad 948 \quad \text{Vol}_{n=2}^2 = \det(\mathbf{R}^\top \mathbf{R}) = \begin{vmatrix} \langle \mathbf{b}, \mathbf{b} \rangle & \langle \mathbf{r}_1, \mathbf{b} \rangle \\ \langle \mathbf{b}, \mathbf{r}_1 \rangle & \langle \mathbf{r}_1, \mathbf{r}_1 \rangle \end{vmatrix} = 1 - \cos^2 \theta$$

949 Therefore, the polytope volume in the bimodal case reduces to:  
 950

$$951 \quad 952 \quad \text{Vol}_{n=2} = \sqrt{1 - \cos^2 \theta} = \sqrt{\sin^2 \theta} = \sin \theta,$$

953 which quantifies the spatial deviation between the WB embedding vector  $\mathbf{b}$  and the vector  $\mathbf{b} - \mathbf{m}_1$ ,  
 954 reflecting how well the modality aligns with the overall barycenter direction. A smaller volume  
 955 indicates stronger alignment, while a larger value implies greater directional discrepancy.  
 956

957 For the special case  $n = 3$ , the squared volume can be expressed as the determinant  
 958

$$959 \quad 960 \quad \text{Vol}_{n=3}^2 = \det(\mathbf{R}^\top \mathbf{R}) = \begin{vmatrix} \langle \mathbf{b}, \mathbf{b} \rangle & \langle \mathbf{r}_1, \mathbf{b} \rangle & \langle \mathbf{r}_2, \mathbf{b} \rangle \\ \langle \mathbf{b}, \mathbf{r}_1 \rangle & \langle \mathbf{r}_1, \mathbf{r}_1 \rangle & \langle \mathbf{r}_2, \mathbf{r}_1 \rangle \\ \langle \mathbf{b}, \mathbf{r}_2 \rangle & \langle \mathbf{r}_1, \mathbf{r}_2 \rangle & \langle \mathbf{r}_2, \mathbf{r}_2 \rangle \end{vmatrix}.$$

961 Substituting (34) and (35) into the determinant, the expression simplifies to  
 962

$$963 \quad 964 \quad \text{Vol}_{n=3}^2 = 1 - \cos^2 \theta - \cos^2 \beta - \cos^2 \gamma + 2 \cos \theta \cos \beta \cos \gamma, \\ 965 \quad 966 \quad = \sin^2 \theta + \sin^2 \beta + \sin^2 \gamma + 2 \cos \theta \cos \beta \cos \gamma - 2.$$

967 This form reveals the geometric interpretation more clearly: 1) The  $\sin^2$  terms quantify the angular  
 968 deviation of each vector pair. The  $2 \cos \theta \cos \beta \cos \gamma$  term reflects the angular coupling between all  
 969 three direction. 2)  $\gamma$  is the angel between the two modality gap vectors  $\mathbf{r}_1$  and  $\mathbf{r}_2$ , and thus directly  
 970 reflects the structural inter-modality gaps between non-anchor modalities.  
 971

The volume becomes small when each modality vector closely follows the barycenter direction (small angles  $\alpha$  and  $\beta$ ). In addition, if the gap vectors between modalities,  $r_1$  and  $r_2$ , are nearly aligned (small inter-modality angle  $\gamma$ ), the volume is further reduced. In this case, all three angles are small, their sines are close to 0, and their cosines are close to 1. Conversely, the volume increases when the directions are mutually orthogonal, which maximizes the total dispersion among them. Fig. 6 plots the R@1 values of zero-shot retrieval on MSR-VTT with respect to the volume, showing a clear negative correlation between volume and performance.

Unlike pairwise cosine similarity, which only captures alignment between two modalities, the barycenter polytope volume offers a rich higher-order understanding of multimodal structure. It not only measures the holistic geometric alignment of each modality toward the barycenter, but also preserves inter-modal interactions, yielding a more compact multimodal joint representation space. In this sense, the volume serves as a non-trivial metric for measuring  $n$ -modality alignment.

## B EXPERIMENTAL DETAILS AND MORE RESULTS

## B.1 EXPERIMENTAL DETAILS

Table 5: Overview of multimodal benchmarks used for downstream evaluation.

Benchmark	Modalities	Train	Val	Test	# Frames (train)	# Frames (test)
DiDeMo	Video + Text + Audio	8,394	1,065	1,003	8	32
ActivityNet	Video + Text + Audio	10,009	—	4,917	8	32
MSR-VTT	Video + Text + Audio + Subtitle	9,000	—	1,000	8	8
VATEX	Video + Text + Audio + Subtitle	14,060	—	431	8	16
VGGSound	Audio + Video + Text	—	—	5000	-	8

Tab. 5 summarizes the modality configuration, data statistics, and frame settings across benchmarks.

**MSR-VTT** (Xu et al., 2016) is a widely used benchmark for video-text retrieval. It contains 10,000 video clips, each paired with approximately 20 textual captions, totaling around 200,000 captions. We use 9,000 videos for training and 1,000 for testing, with 8 frames sampled per video.

**DiDeMo** (Anne Hendricks et al., 2017) consists of 10,000 long-form videos, each annotated with 4 temporally ordered paragraph descriptions. It is mainly used for moment localization retrieval. The official split includes 8,394/1,065/1,003 videos for training/validation/testing, and 12 frames are sampled per video.

**ActivityNet** (Caba Heilbron et al., 2015) contains about 20,000 YouTube videos with a total duration of approximately 180 hours, annotated with multiple temporal sentence descriptions. We use the official training set (10,009 videos) for training and the validation set (4,917 videos) for downstream testing. The number of sampled frames per video is 8.

**VATEX** (Wang et al., 2019) includes around 25,000 English videos, each annotated with 10 English captions. It is commonly used for four-modality text retrieval tasks. We adopt 14,060 videos for training and 431 for testing, with 8 frames sampled per video.

**VGGSound5K** (Chen et al., 2020) is a 5,000-video subset of VGGSound, containing diverse audio event categories (typically 310 classes) along with corresponding video frames and subtitles. Each video clip in the dataset has a duration of 10 seconds and is annotated with a single label corresponding to the predominant sound event occurring within the clip. The dataset covers a wide spectrum of audio events, including human actions, animal vocalizations, natural phenomena, and mechanical sounds. It is commonly used for audio-video multimodal classification tasks.

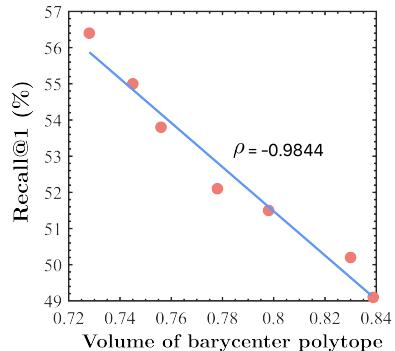


Figure 6: The polytope's volume is correlated ( $\rho = -0.9844$ ) with the downstream performance.

1026 B.2 DISCUSSIONS ON THE ANCHOR SELECTION  
1027

1028 To explore the impact of different anchor modalities to BaryBind, we conduct experiments using  
1029 text and video as the original anchor and evaluate on the VGGSound dataset (Chen et al., 2020)  
1030 for video classification, and on MSR-VTT for V-T/T-V retrieval tasks. Note that the DAM loss is  
1031 excluded in these settings, as it is inherently designed for text-based supervision.

1032 Table 6: **Ablation on the anchor selection.**  
1033

Anchor Modality	VGGSound				MSR-VTT			
	Video		Audio		T2V		V2T	
	Acc@1	Acc@5	Acc@1	Acc@5	R@1	R@5	R@1	R@5
Text Anchor	<b>47.8</b>	<b>74.2</b>	<b>45.2</b>	<b>72.4</b>	<b>53.2</b>	<b>77.0</b>	50.6	73.8
Video Anchor	46.7	69.2	43.5	66.1	50.8	75.5	<b>51.2</b>	<b>76.4</b>

1041 As shown in Tab. 6, the performance exhibits a slight decline when using the video modality as the  
1042 anchor compared to the text modality. The result is consistent with the fact that text is often the most  
1043 informative anchor in many multimodal tasks, as it retains the broad semantics of multimodal data.  
1044

1045 B.3 HYPERPARAMETER SENSITIVITY ANALYSIS  
1046

1047 We study the sensitivity of the loss weights in the total objective on the MSR-VTT validation set:

$$\mathcal{L} = \mathcal{L}_{\text{MWB}} + \alpha_1 \mathcal{L}_{\text{BVC}} + \alpha_2 \mathcal{L}_{\text{DAM}}. \quad (36)$$

1050 As reported in Tab. 7, adjusting  $\alpha_1$  on MSR-VTT shows that increasing the weight of the BVC loss  
1051 gradually improves retrieval performance, and the best results are achieved when  $\alpha_1 = 1$ . A further  
1052 increase leads to a performance decline, indicating that over-restricting barycentric geometry may  
1053 weaken instance-level discrimination.

1054 Table 7: Sensitivity analysis of  $\alpha_1$  on MSR-  
1055 VTT validation set.  
1056

$\alpha_1$	0.1	0.5	1	2	3
T2V (R@1)	54.2	54.9	<b>56.5</b>	55.3	54.5
V2T (R@1)	55.9	56.7	<b>58.3</b>	56.5	55.6

1057 Table 8: Sensitivity analysis of  $\alpha_2$  on  
1058 MSR-VTT validation set.  
1059

$\alpha_2$	0.02	0.05	0.1	0.15	0.2
T2V (R@1)	54.6	55.2	<b>56.5</b>	55.1	54.8
V2T (R@1)	55.2	55.8	<b>58.3</b>	56.2	55.1

1062 Tab. 8 further reveals that  $\alpha_2$  also has an optimal operating range, with  $\alpha_2 = 0.1$  yielding the  
1063 strongest retrieval performance. Excessively small or large weights cause suboptimal alignment due  
1064 to either under-constrained or modality-biased instance matching.

1065 Overall,  $\alpha_1 = 1$  and  $\alpha_2 = 0.1$  deliver the most stable and balanced multimodal alignment on the  
1066 MSR-VTT validation set, demonstrating the complementary strengths of BVC and DAM.  
1067

1068 B.4 NORMALIZATION STRATEGIES FOR POLYTOPE VOLUME METRIC  
1069

1070 While the raw barycenter polytope volume  $V$  serves  
1071 as a global alignment metric, it depends on the num-  
1072 ber of modalities and embedding dimensionality,  
1073 which limits interpretability and cross-setting com-  
1074 parability. To address this, we consider normaliza-  
1075 tion strategies such as  $V^{1/n}$ , which scale the volume  
1076 to better reflect per-modality contributions and make  
1077 the metric more comparable across different modal-  
1078 ity configurations.

1079 Table 9 reports zero-shot retrieval results on MSR-  
VTT under different normalization strategies. We

1079 Table 9: Zero-shot generalization on MSR-  
VTT under different normalization strategies  
for the barycenter polytope volume.

Normalization	Setting	T2V	V2T
$V^{1/2}$	T-V	52.8	51.0
$V^{1/3}$	T-VA	53.6	52.2
$V^{1/4}$	T-VAS	55.4	52.7
$V$	T-V	53.4	51.3
$V$	T-VA	54.5	52.0
$V$	T-VAS	56.3	53.6

1080 observe that applying normalization slightly reduces  
 1081 the absolute retrieval scores compared to the raw  
 1082 volume  $V$ , likely because normalization reduces the  
 1083 magnitude of the gradient signal from the volume metric, slightly weakening the alignment supervi-  
 1084 sion. Nonetheless, the overall performance remains stable, indicating that  $V^{1/n}$  provides a reliable  
 1085 and interpretable metric without significantly sacrificing retrieval accuracy.  
 1086

### 1087 B.5 ROBUSTNESS TO MISSING MODALITIES

1088 BaryBind effectively captures modality-agnostic semantics from arbitrary subsets of multimodal in-  
 1089 puts, enabling the aligned representations of available modalities to serve as proxy features when  
 1090 others are missing. We evaluate two settings: (1) training-time missing-audio for cross-modal re-  
 1091 trieval (T2A/A2T) on AudioCaps (Kim et al., 2019), and (2) missing-video inference for multimodal  
 1092 event classification on VGGSound 5K, where the model is trained with videos and audios. In both  
 1093 cases, the barycenter is optimized using only the accessible modalities. Results show that Bary-  
 1094 Bind preserves robust and graceful degradation under missing-modality conditions, which can be  
 1095 attributed to the barycenter modeling of modality-agnostic semantics.  
 1096

1097 Table 10: Text-to-audio retrieval on AudioCaps  
 1098 w/ and w/o audio during training.

Training Setting	T2A R@1	T2A R@10
VAST (w/ audio)	32.1	65.4
GRAM (w/ audio)	33.2	75.3
BaryBind (w/ audio)	35.5	81.2
VAST (w/o audio)	10.4	32.8
GRAM (w/o audio)	12.8	35.1
BaryBind (w/o audio)	21.1	56.2

1099 Table 11: Multimodal event classification on  
 1100 VGGSound5K w/o video during inference.

	Input modality	Acc@1	Acc@5
VAST	A+V	48.1	79.6
GRAM	A+V	42.3	74.5
BaryBind	A+V	55.6	83.4
VAST	A	40.8	71.6
GRAM	A	38.5	70.1
BaryBind	A	49.4	78.3

### 1108 B.6 SYSTEM EFFICIENCY COMPARISON DURING TRAINING

1109 We provide a comparison of system efficiency among VAST, GRAM, and BaryBind. All experi-  
 1110 ments are conducted on 2×NVIDIA A100 80GB GPUs with mixed-precision (FP16/AMP).

1111 Table 12: System efficiency comparison during training.

Model	Params	Batch size	Forward+Backward	Steps/Epoch	Time/Epoch
VAST	1.28B	64	~8.8s	2344	~5.7h
GRAM	1.30B	64	~9.4s	2344	~6.1h
BaryBind	1.34B	64	~9.7s	2344	~6.3h

1112 The additional computation in BaryBind mainly comes from the Wasserstein barycenter optimiza-  
 1113 tion and auxiliary alignment losses. The increase in per-step time remains moderate while providing  
 1114 improved multimodal alignment performance.

### 1115 B.7 MORE QUANTITATIVE RESULTS ON SCALING TO MORE MODALITIES

1116 To systematically evaluate the scalability of multimodal models with respect to the number of input  
 1117 modalities, we conduct experiments under four configurations, progressively increasing from two  
 1118 modalities (text and video) to five (text, video, audio, subtitle, and depth). This stepwise setup  
 1119 enables a controlled analysis of how each additional modality affects performance and alignment  
 1120 quality. For consistency and interpretability, text is used as the anchor modality throughout.

1121 As shown in Table 13, both VAST and BaryBind benefit from the inclusion of additional modalities,  
 1122 demonstrating improved performance on MSR-VTT and VATEX in terms of both T2V and V2T  
 1123 retrieval. Notably, BaryBind consistently outperforms VAST across all modality configurations and  
 1124 datasets, highlighting its stronger scalability and generalization capacity.

Table 13: Downstream performance with increasing number of modalities.

1134	1135					1136				1137			
	1138					1139				1140			
	1141					1142				1143			
Text	Video	Audio	Sub.	Depth		VAST	BaryBind		VAST	BaryBind		VATEX	
					T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	
✓	✓	✗	✗	✗	48.7	43.2	53.4	51.3	78.8	77.0	82.3	79.8	
✓	✓	✓	✗	✗	49.3	43.7	54.5	52.0	80.0	77.3	84.2	81.3	
✓	✓	✓	✓	✗	50.9	47.9	56.3	53.6	82.1	78.7	84.6	83.5	
✓	✓	✓	✓	✓	51.2	49.3	57.0	54.4	82.4	79.2	84.9	83.8	

## B.8 VISUALIZATION OF TOP-1 RETRIEVAL RESULTS

To qualitatively assess the retrieval performance of BaryBind, we visualize top-1 text-to-video results compared with two strong baselines in Fig. 8. BaryBind integrates text, audio, video, and subtitle modalities during both training and inference. Each row shows five frames from the top-retrieved video along with the query subtitle. In the first example, BaryBind retrieves a beach party scene aligned with the query’s semantic and acoustic mood, while baselines return less relevant results. In the second case, BaryBind accurately matches a gameplay scene described by both visual and subtitle cues, whereas baselines retrieve unrelated content. The third example features a simple emotional phrase, “I’m scared.”, where BaryBind selects an animated video of a fearful cat-dog chase, while others fail to reflect the emotional context or core entities. These results demonstrate BaryBind’s ability to leverage complementary multimodal cues for precise and context-aware retrieval, effectively grounding both semantics and affect across diverse scenarios.

## B.9 SCALABILITY OF BARYBIND WITH INCREASING MODALITY NUMBER

Table 14: Computation time (in seconds) of similarity metrics vs. number of modalities, measured on an NVIDIA A100 GPU with batch size  $B = 64$  and embedding dimension  $D = 512$ .

1165	Number of modality $n$	2	3	4	5	10	20
1166	Pairwise cosine similarity	$3.0 \times 10^{-7}$	$7.0 \times 10^{-7}$	$1.0 \times 10^{-6}$	$1.3 \times 10^{-6}$	$2.9 \times 10^{-6}$	$4.9 \times 10^{-6}$
1167	Barycenter polytope volume	$4.9 \times 10^{-6}$	$6.8 \times 10^{-6}$	$5.9 \times 10^{-6}$	$9.8 \times 10^{-6}$	$3.6 \times 10^{-5}$	$8.8 \times 10^{-5}$

We evaluate the computation efficiency of different similarity metrics with varying modality counts  $n$ . For each metric, we randomly sample  $B = 64$  sets of  $n$  vectors in  $\mathbb{R}^D$  with  $D = 512$ , simulating multimodal embeddings. As shown in Tab. 14, the barycenter polytope volume remains computationally efficient and scales reasonably as the number of modalities increases. The negligible overhead incurred when extending to more modalities highlights the polytope volume as a non-trivial and scalable metric for assessing  $n$ -modality alignment.

To evaluate how BaryBind scales from bimodal (e.g., text-vision) setups to richer multimodal settings, we progressively expand the input from a basic text-video (T-V) pair to more complex configurations on the MSR-VTT dataset: text-video-audio (T-VA), text-video-audio-subtitle (T-VAS), and finally text-video-audio-subtitle-depth (T-VASD). The depth modality is derived using ChronoDepth (Shao et al., 2025) and integrated via an additional lightweight head attached to the vision encoder. As shown in Fig. 7, BaryBind consistently improves Recall@1 as the number of modalities increases, significantly outperforming the VAST baseline on MSR-VTT across all configurations. This highlights the scalability of BaryBind in practical multimodal understanding as it effectively integrates more modalities to shape a richer semantic space.

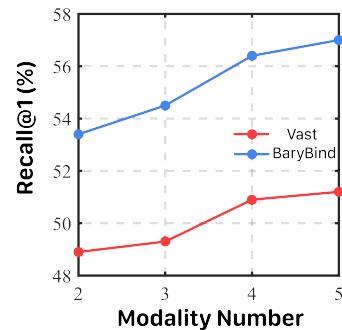


Figure 7: Zero-shot text-to-video retrieval results as scaling from 2 (T-V) to 5 (T-VASD) modalities.

## USE OF LARGE LANGUAGE MODELS

We acknowledge that Large Language Models (LLMs) were used after the completion of the draft, solely to correct grammar and improve sentence fluency.

Figure 8: Visual results of text-to-video retrieval. We display 5 frames from the top-1 video.