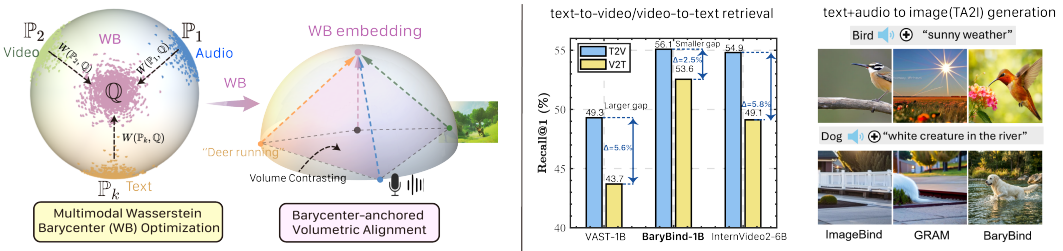
Binding Multiple Modalities via Multimodal Wasserstein Barycenter

Xiaole Tang **Jiayi Xu** **Xiang Gu** **Yan Yang** **Jian Sun***

Xi'an Jiaotong University

Abstract

Multimodal learning beyond two modalities commonly leverages a specific modality (e.g., text) to bind other modalities. However, how to establish a more balanced representation space that approximates shared semantics while respecting the holistic geometry of n-modal data remains challenging. In this work, we present BaryBind, which aims to transport the specific modality towards the Wasserstein barycenter (WB) optimized across all modalities and introduces a volumetric alignment objective to establish a unified semantic space around the WB embedding. Specifically, we project specific modalities to the WB, which minimizes the average Wasserstein distances to multimodal distributions and serves as the anchor for subsequent alignment. We then construct a barycenter simplex, whose volume is taken as a similarity metric for global alignment centered at the WB. Experiments show that BaryBind achieves competitive performance in text-video-audio retrieval, classification, videoQA, and cross-modal generation tasks, along with robustness under modality absence and scalability to more than three modalities.



(a) **Core idea**: BaryBind learns the WB of multimodal distributions, and then utilizes it as the new anchor for volumetric alignment.

(b) **Examples:** BaryBind achieves more balanced retrieval results and enables TA2I generation that preserves multimodal semantics.

Figure 1: BaryBind learns the multimodal Wasserstein Barycenter (WB) of all modalities via WB optimization, and the WB serves as the anchor to bind modalities via volumetric alignment based on the barycenter simplex volume. Notably, BaryBind yields more balanced multimodal retrieval results and enables TA2I generation that well preserves both text and audio semantics.

1 Introduction

Multimodal learning [3] 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 are heterogeneous observations of a shared underlying reality/object, recent multimodal learning methods [46, 21] learn a shared representation from different modalities. In this field, the success of CLIP [46] in aligning unified vision–language representations via contrastive learning has sparked the adoption of contrastive losses as an appealing solution for multimodal

*Corresponding authors.

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

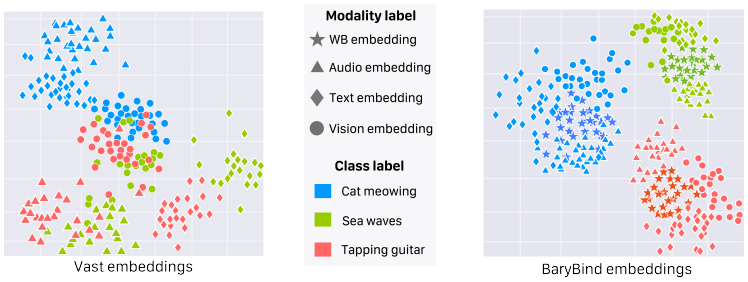


Figure 2: **The t-SNE comparison of multimodal embeddings on the zero-shot VGGSound dataset** [8]. Compared to VAST [9], BaryBind induces a latent space where class clusters are clearly separated, and multimodal embeddings are grouped around the WB embeddings.

representation learning. However, traditional contrastive losses, e.g., InfoNCE [44] and BYOL [18], are formulated in a pairwise fashion for $n = 2$ cases typically arising from two modalities such as image–text [46, 21] or audio–text [19, 14] scenarios.

While scaling to n modalities ($n \geq 3$) poses significant challenges, a series of recent works [73, 64] originating from ImageBind [17] leverage the binding property of a modality-specific anchor (e.g., image or language) to align other modalities via pairwise losses. Recently, VAST [9], GRAM [12], Triangle [11], and Symile [48] relied on global similarity objectives that align n modalities to a designated central modality, also known as the modality-specific anchor. However, such anchors can introduce modality biases, inducing an imbalanced representation space that deviates from the shared underlying semantics, with one modality dominating the representation space. This can be observed in Fig. 1b, where the VAST model exhibits a large gap in Recall@1 between the text-to-video and video-to-text tasks, indicating potential asymmetry in cross-modal alignment.

To alleviate this issue, we are motivated by the view that multimodal data, though captured by heterogeneous sensors, reflect a shared underlying reality/object. Therefore, multimodal representations can be interpreted as modality-specific transformations of a shared representation that encodes semantic consensus. With this insight, we propose **BaryBind** that aligns multiple modalities to the Wasserstein Barycenter (WB) anchor, which defines a distribution that minimizes the average of Wasserstein distances to all multimodal distributions. As shown in Fig. 1a, the WB is optimized across all modalities via the multimodal WB optimization process and then utilize as the anchor for subsequent alignment, aiming to mitigate modality-specific bias and encourages semantic consensus (Fig. 2). Based on the WB anchor, we further introduce a volumetric contrastive loss defined on the volume of a barycenter simplex, which quantifies the global alignment of n-modal data. By contrasting the simplex's volume (Fig. 1a), BaryBind aligns modalities to the barycenter while preserving the holistic geometry within n-modal data. Interestingly, with the barycenter-based alignment, we observe that BaryBind substantially alleviates the imbalance across modalities, reducing the T2V/V2T retrieval gap by 3.1% compared to VAST [9] (Fig. 1b), and induces a latent space where modalities of each class are consistently clustered around the WB embeddings (Fig. 2).

The main contributions are highlighted as follows:

- **Method:** We propose BaryBind, which transports the specific anchor to the WB that mitigates modality bias as the geometric semantic center. Multimodal embeddings are aligned to the WB via a volumetric alignment objective, enabling balanced multimodal understanding.
- **Theory:** We develop an optimal transport (OT) duality-based optimization framework and error bounds for approximating the Wasserstein barycenter of multimodal representations. Meanwhile, we establish a barycenter simplex, whose volume serves as a global similarity metric and extends the insights of measuring n-modal alignment.
- **Performance:** BaryBind achieves balanced state-of-the-art performance in cross-modal retrieval, generation, multimodal classification, and videoQA tasks across diverse datasets. Notably, BaryBind shows unique robustness under modality absence and scalability to three or more modalities.

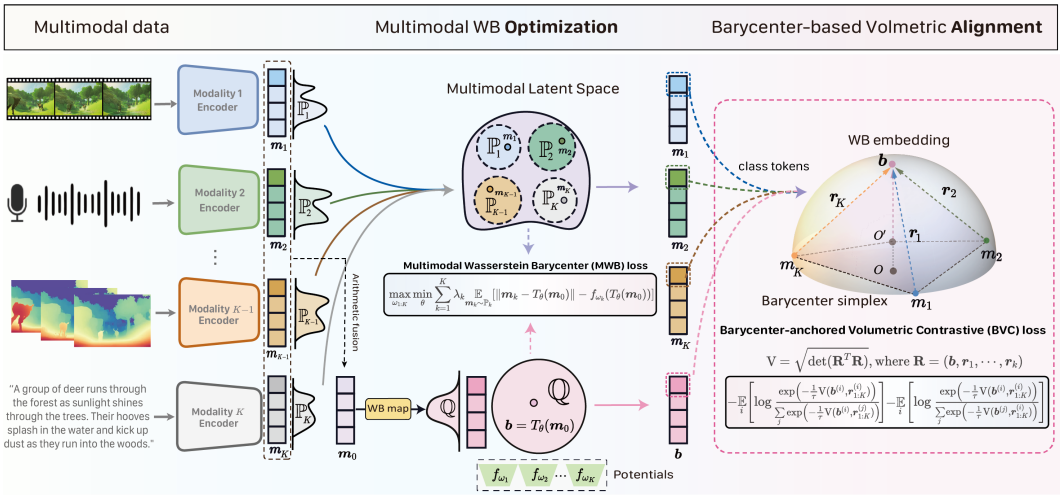


Figure 3: **BaryBind establishes unified representation space through two processes:** 1) The *WB optimization* process transforms modality-specific initializer $\boldsymbol{m_0}$ into the WB anchor $\boldsymbol{b}$, in which the WB map T_θ is optimized by minimizing the average of Wasserstein distances to all multimodal distributions $\mathbb{P}_k$. 2) The *volumetric alignment* process utilizes the WB anchor to induce a barycenter simplex, whose volume is contrasted to achieve global alignment to the WB embedding.

2 Related Works

Multimodal representation learning. Multimodal representation learning seeks to align heterogeneous modalities into a shared semantic space. CLIP [46] initiates this paradigm with image–text contrastive learning, followed by audio extensions such as AudioCLIP [19], CLAP [14], and LAION-CLAP [65]. WavCaps [40] constructs a large-scale audio captioning dataset to support audio–language retrieval. To scale beyond two modalities, ImageBind [17] introduces a pivot-based strategy, aligning each modality to the vision anchor. LanguageBind [73], UniBind [37], GRAM [12], Triangle [11], and LCO-EMB [66] extend alignment to language anchors, adopting unique volume-based contrastive losses. OmniBind [64] aligns pre-trained unimodal experts via pairwise losses. VAST [9] fuses modalities into a shared space but still relies on text-centered supervision.

In contrast, BaryBind alleviates the dominance of the specific anchor modality by explicitly transforming it into the shared WB, which is optimized as the barycenter of all multimodal distributions and acts as the refined anchor. It further leverages a barycenter-anchored volumetric contrastive loss that enables global alignment to the WB embedding, thereby establishing a more balanced semantic space centered at the WB embeddings.

Wasserstein barycenter. The WB [1, 4] defines an averaging distribution that minimizes the weighted sum of Wasserstein distances [56, 7, 45] to input measures, preserving mass structure and OT-grounded geometry. This formulation has shown effectiveness on heterogeneous supports and has been applied in generative modeling [13] and domain adaptation [5]. Recent works estimate high-dimensional barycenters via deep dual formulations, including ICNN-based cycle-consistent models [26], neural OT maps [24, 54, 55, 53], and energy-guided potentials [25]. We extend this perspective to construct a barycenter-aligned joint space for multimodal alignment, bridging optimal transport with multimodal representation learning.

3 Method

Overview. In this work, we introduce BaryBind that aligns multiple modalities to the WB and suggest a volumetric alignment objective to establish a unified representation space. The overview is presented in Fig. 3, where the WB optimization process (§3.2) learns a transformation from specific modality to the WB, which serves as the alignment center. The volumetric alignment process (§3.3) enables global alignment to the WB anchor by contrasting the volume of the barycenter simplex.

3.1 Preliminaries

Notation. Let $\bar{K} = \{1, \dots, K\}$ for some $K \in \mathbb{N}$. For a sequence $e_0, \dots, e_K$, we denote by $e_{1:K}$ 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 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 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 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 $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the classic optimal transport (OT) problem [22] aims to find a joint distribution $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ that minimizes the expected transport cost:

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

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 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 λ_k summing to 1, the WB problem seeks the distribution $\mathbb{Q}$ that minimizes the weighted sum of Wasserstein distances to the fixed marginals $\mathbb{P}_{1:K}$:

$$\inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{Y})} \sum_{k=1}^{K} \lambda_k W(\mathbb{P}_k, \mathbb{Q}). \tag{2}$$

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 WB approximates the geometric consensus of semantics as it models the "closest" distribution to all multimodal distributions.

3.2 Multimodal Wasserstein Barycenter Optimization

Let $\mathbb{P}_k$ be the distribution of encoded features $\boldsymbol{m}_k \in \mathcal{M}_k \subset \mathcal{M}$ for modality $k \in \bar{K}$, defined in the multimodal latent space $\mathcal{M} \subset \mathbb{R}^D$. $\boldsymbol{m}_0$ denotes the anchor modality feature. The WB space is defined as $\mathcal{M}_B := \mathrm{supp}(\mathbb{Q})$ where $\mathbb{Q}$ denotes the WB distribution and $\mathcal{M}_B$ contains the barycenter features $\boldsymbol{b}$. Given the distributions $\mathbb{P}_{1:K}$, 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

$$\mathcal{L}^*_{\mathrm{MWB}} = \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sum_{k=1}^{K} \lambda_k W(\mathbb{P}_k, \mathbb{Q}). \tag{3}$$

This formulation induces a barycenter distribution that minimizes the average of Wasserstein distances to all modalities. However, directly optimizing (3) is highly intractable. Therefore, we deduce a new dual reformulation (see Appendix A.1) with sup-inf objective as follows.

Proposition 3.1 (Duality of multimodal WB problem). *The infimum value $\mathcal{L}^*_{\mathrm{MWB}}$ of the multimodal WB problem (3) can be expressed as*

$$\mathcal{L}^*_{\mathrm{MWB}} = \sup_{\sum_k \lambda_k f_k = 0} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{M}_B)} \sum_{k=1}^{K} \lambda_k \mathop{\mathbb{E}}_{\substack{\boldsymbol{m}_k \sim \mathbb{P}_k \\ \boldsymbol{b} \sim \mathbb{Q}}} \left[\|\boldsymbol{m}_k - \boldsymbol{b}\| - f_k(\boldsymbol{b}) \right], \tag{4}$$

where the supremum is taken over the potentials $f_k : \mathcal{M}_B \to \mathbb{R}$. We aim to learn the distribution $\mathbb{Q}$ by sampling WB embedding $\boldsymbol{b} = T(\boldsymbol{m_0})$ via a trainable WB map $T_\theta(\cdot) : \mathcal{M}_0 \to \mathcal{M}_B$. Here, $\boldsymbol{m_0}$ represents the WB initializer, empirically selected as the arithmetic mean of all available multimodal features, i.e., $\boldsymbol{m_0} = \bar{\boldsymbol{m}}_i$ and can also be chosen as any modality (see ablation in **Appendix B.7**). We parameterize T_θ using an MLP. To ensure the congruence $\sum_k \lambda_k f_k = 0$ [28], we parameterize the potentials f_{ω_k} as $g_{\omega_k} - \sum_{i=1}^{K} \lambda_i g_{\omega_i}$ with two-layer MLPs $g_{\omega_k} : \mathbb{R}^D \to \mathbb{R}$, which is a common trick used in [28, 24, 25].

Multimodal Wasserstein barycenter (MWB) loss. With this parameterization, we rewrite (4) as a max-min objective of the MWB loss $\mathcal{L}_{\mathrm{MWB}}$, which can be optimized to compute the barycenter map T_θ for approximating the WB:

$$\max_{\omega_{1:K}} \min_{\theta} \left\{ \mathcal{L}_{\mathrm{MWB}}(\omega_{1:K}, \theta) \triangleq \sum_{k=1}^{K} \lambda_k \mathop{\mathbb{E}}_{\boldsymbol{m}_k \sim \mathbb{P}_k} \left[\|\boldsymbol{m}_k - T_\theta(\boldsymbol{m}_0)\| - f_{\omega_k}(T_\theta(\boldsymbol{m}_0)) \right] \right\}. \tag{5}$$

To solve the problem (5), we train T_θ and $f_{\omega_{1:K}}$ by alternately maximizing over $\omega_{1:K}$ and minimizing over θ in the MWB loss, in which we estimate the expectation using mini-batch data at each training step. During the optimization, all modalities jointly update the barycenter, ensuring that no single modality dominates and that the modality bias is reduced. Then the WB embedding is computed as $\boldsymbol{b} = T_\theta(\boldsymbol{m}_0)$, serving as the anchor for subsequent alignment.

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

$$\mathcal{F}(f_{1:K}, T) := \mathcal{L}_{\mathrm{MWB}}(f_{1:K}, T), \quad \mathcal{L}(f_{1:K}) := \inf_{T:\mathcal{M}\to\mathcal{M}_B} \mathcal{F}(f_{1:K}, T), \quad \mathcal{L}^* := \mathcal{L}^*_{\mathrm{MWB}}. \tag{6}$$

Theorem 3.2 (Error analysis via duality gaps for the barycenter distribution)**.** *Let C_k be any transport costs. Assume that the maps $\boldsymbol{b} \mapsto C_k(\boldsymbol{m}_k, \boldsymbol{b}) - \widehat{f}_k(\boldsymbol{b})$ are β-strongly convex for $\boldsymbol{m}_k \in \mathcal{M}_k$, $k \in \{0, \dots, K\}$. Consider the duality gaps for an approximate solution $(\widehat{f}_{1:K}, \widehat{T})$:*

$$\mathcal{E}_1(\widehat{f}_{1:K}, \widehat{T}) \triangleq \mathcal{F}(\widehat{f}_{1:K}, \widehat{T}) - \mathcal{L}(\widehat{f}_{1:K}), \quad \mathcal{E}_2(\widehat{f}_{1:K}) \triangleq \mathcal{L}^* - \mathcal{L}(\widehat{f}_{1:K}), \tag{7}$$

Then the following inequality holds:

$$W^2\left(\widehat{T}_{\#}\mathbb{P}_0, \mathbb{Q}^*\right) \leq \frac{4}{\beta}(\mathcal{E}_1 + \mathcal{E}_2). \tag{8}$$

The proof is provided in the **Appendix A.2**. This theorem ensures that the Wasserstein distance between the estimated distribution $\widehat{T}_{\#}\mathbb{P}_0$ and the true barycenter $\mathbb{Q}^*$ is upper-bounded by the sum of these two errors. In our setting with Euclidean cost, this local condition can hold when the learned neural potentials are sufficiently smooth. In practice, regularization (e.g., weight decay) can improve the smoothness of neural potentials, and our experiments demonstrate the effectiveness of the proposed Wasserstein barycenter solver. Our experiments also demonstrate that the WB solver performs well with Euclidean cost.

3.3 Barycenter-anchored Volumetric Alignment

To bind modalities to the WB anchor while preserving the global structures of K-dimensional multimodal data, we introduce a geometric structure called the *barycenter simplex*. As illustrated in Fig. 3, the simplex takes the WB embedding $\boldsymbol{b}$ as the apex and is spanned by the vectors from the origin to $\boldsymbol{b}$ and the modality-to-barycenter gap vectors $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{m}_k$ for all modalities $\boldsymbol{m}_k$ ($k \geq 1$). Owing to its unique composition, the simplex volume quantifies two aspects of multimodal features: (1) the global alignment to the barycenter, and (2) the inter-modality discrepancy across modalities.

Given the WB embedding $\boldsymbol{b}$ (i.e., the vector from the origin to the barycenter) and gap vectors $\{\boldsymbol{r}_k\}_{k=1}^K$ (i.e., the vectors from each modality to the barycenter), we define the matrix $\mathbf{R} = (\boldsymbol{b}, \boldsymbol{r}_1, \cdots, \boldsymbol{r}_K)$ as the set of vectors spanning the barycenter simplex, whose square of volume can be computed according to Gantmacher [16] as:

$$\mathrm{V}^2(\boldsymbol{b}, \boldsymbol{r}_{1:K}) := \mathrm{V}^2(\boldsymbol{b}, \boldsymbol{r}_1, \cdots, \boldsymbol{r}_K) = \det(\mathbf{R}^T\mathbf{R}) = \begin{vmatrix} \langle \boldsymbol{b}, \boldsymbol{b} \rangle & \langle \boldsymbol{b}, \boldsymbol{r}_1 \rangle & \cdots & \langle \boldsymbol{b}, \boldsymbol{r}_K \rangle \\ \langle \boldsymbol{r}_1, \boldsymbol{b} \rangle & \langle \boldsymbol{r}_1, \boldsymbol{r}_1 \rangle & \cdots & \langle \boldsymbol{r}_1, \boldsymbol{r}_K \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{r}_K, \boldsymbol{b} \rangle & \langle \boldsymbol{r}_K, \boldsymbol{r}_1 \rangle & \cdots & \langle \boldsymbol{r}_K, \boldsymbol{r}_K \rangle \end{vmatrix}. \tag{9}$$

Barycenter-anchored volumetric contrastive (BVC) loss. We propose the barycenter-anchored volumetric contrastive (BVC) loss that contrasts the volume of the barycenter simplex. Given multimodal embeddings $\{\boldsymbol{m}_1^{(i)}, \dots, \boldsymbol{m}_K^{(i)}\}_{i=1}^B$, where B is the batch size and $\boldsymbol{m}_0^{(i)}$ is the anchor, the WB anchor is computed by $\boldsymbol{b}^{(i)} = T_\theta(\boldsymbol{m}_0^{(i)})$, and the modality-to-barycenter gap vectors are defined as $\boldsymbol{r}_{1:K}^{(i)} = \{\boldsymbol{r}_1^{(i)}, \dots, \boldsymbol{r}_K^{(i)}\}$. We treat each $(\boldsymbol{b}^{(i)}, \boldsymbol{r}_{1:K}^{(i)})$ as a *positive pair*, and construct two sets of

negative pairs: (i) by fixing $\boldsymbol{b}^{(i)}$ and pairing it with $\boldsymbol{r}_{1:K}^{(j)}$ $j \neq i$, and (ii) by fixing $\boldsymbol{r}_{1:K}^{(i)}$ and pairing it with $\boldsymbol{b}_j$ for $j \neq i$.

The BVC loss contrasts small volumes for positive pairs against large volumes for negative ones:

$$\mathcal{L}_{\text{BVC}} = -\mathop{\mathbb{E}}_{i}\left[\log \frac{\exp\left(-\text{V}(\boldsymbol{b}^{(i)}, \boldsymbol{r}_{1:K}^{(i)})/\tau\right)}{\sum_j \exp\left(-\text{V}(\boldsymbol{b}^{(i)}, \boldsymbol{r}_{1:K}^{(j)})/\tau\right)} + \log \frac{\exp\left(-\text{V}(\boldsymbol{b}^{(i)}, \boldsymbol{r}_{1:K}^{(i)})/\tau\right)}{\sum_j \exp\left(-\text{V}(\boldsymbol{b}^{(j)}, \boldsymbol{r}_{1:K}^{(i)})/\tau\right)}\right], \tag{10}$$

where $\text{V}(\boldsymbol{b}, \boldsymbol{r}_{1:K})$ denotes the volume of the barycenter simplex formed by the WB and modality-to-barycenter gap vectors, whose square is computed as (9). The BVC loss binds multiple modalities to the shared WB embedding, ensuring a more holistic and accurate alignment of n-modal data.

3.4 Training Loss

To further guide the training process following VAST [9], we employ a data-anchor matching (DAM) loss that distinguishes matched and mismatched data pairs:

$$\mathcal{L}_{\text{DAM}} = \mathbb{E}_{(\boldsymbol{b}, \boldsymbol{m}_{1:K}) \sim (\mathbb{Q}, \mathbb{P}_{1:K})}[y \log p_m(\boldsymbol{b}, \boldsymbol{m}_{1:K}) + (1-y)\log(1 - p_m(\boldsymbol{b}, \boldsymbol{m}_{1:K}))]. \tag{11}$$

To produce binary predictions p_m, the modality-specific anchor data are fed into the anchor encoder and integrated with the concatenated non-anchor features $\boldsymbol{m}_{1:K}$ by cross-attention layers. Then the output feature is passed through a two-layer MLP to generate the output probability.

The overall pre-training objective $\mathcal{L}$ combines the three losses, $\mathcal{L}_{\text{MWB}}$, $\mathcal{L}_{\text{BVC}}$, and $\mathcal{L}_{\text{DAM}}$, as follows:

$$\mathcal{L} := \mathcal{L}_{\text{MWB}} + \alpha_1 \mathcal{L}_{\text{BVC}} + \alpha_2 \mathcal{L}_{\text{DAM}}. \tag{12}$$

Particularly, the potentials $f_{\omega_{1:K}}$ in $\mathcal{L}_{\text{MWB}}$ are optimized via maximization, and are trained in an alternating manner against the minimization of the remaining networks. The training pipeline in summarized in Algorithm 1 in the Appendix B.1.

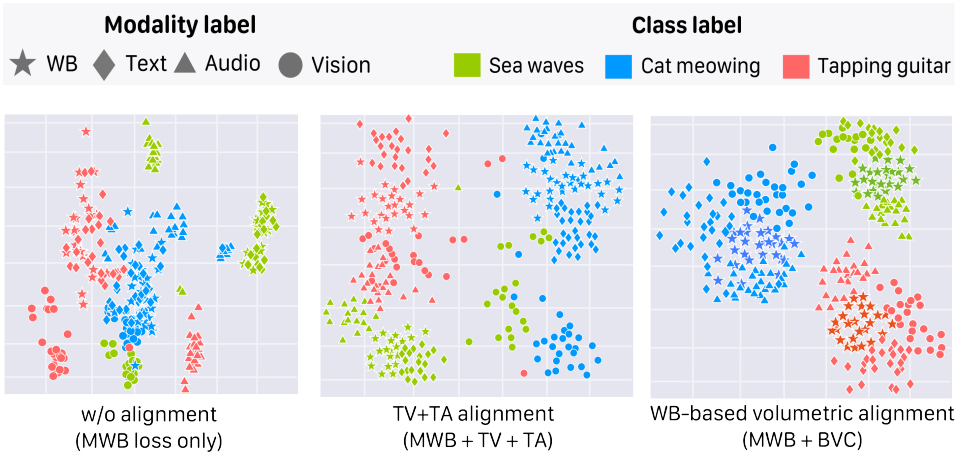


Figure 4: **Visualization of multimodal embeddings** before alignment (left) and after applying TV + TA cosine-based alignment (middle) and the proposed BVC alignment (right). The BVC loss promotes convergence of embeddings towards the WB anchors, forming compact clusters around the WB embeddings, which highlights improved balanced multimodal alignment.

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 [8], as shown in Fig. 4. Without alignment (left), embeddings of different modalities are scattered, exhibiting notable modality gaps within the same class. Incorporating pairwise alignment (middle) improves class clustering but still reveals modality separations. In contrast, with our BVC loss (right), embeddings of all modalities converge 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 balanced alignment around the WB embeddings.

4 Experiments

4.1 Experimental Setup

We adopt VAST [9] as the backbone, with BERT-B for text, BEATs for audio, and EVA-CLIP-ViT-G [52] for visual encoding. In total, the model comprises approximately 1B parameters (see **Appendix B.11** for computational overhead analysis). Unlike VAST, we discard its modality fusion layers and introduce lightweight MLPs for barycenter optimization. Following the setting of GRAM [12] and Triagnle [11], we continue pretraining based on VAST using our proposed loss functions on the VAST150k dataset [9]. VAST [9], GRAM [12], Triangle [11], and PMRL [35] adopt the same architectures and training strategy for a fair comparison. τ is 0.07 and the trade-off parameters in (12) are set as $\alpha_1 = 1, \alpha_2 = 0.1$ (see Appendix B.9 for sensitivity analysis). The barycenter weights λ_k ($k \in \bar{K}$) are uniformly set as $1/K$, where n is the number of modalities. The compared methods are implemented under identical experimental conditions.

Table 1: **Zero-shot video/audio classification** results on VGGSound5K. Results from our baseline VAST and the proposed BaryBind, and the performance gains are highlighted accordingly.

Method	Modality	Acc@1	Acc@5
ImageBind [17]	A	31.6	58.7
ImageBind [17]	V	37.9	65.9
LanguageBind [73]	A	34.1	62.8
LanguageBind [73]	V	39.6	64.5
GRAM [12]	V	41.6	74.3
GRAM [12]	A+V	42.3	76.4
Triangle [11]	A+V	44.8	80.0
OmniBind [64]	A	41.7	70.8
OmniBind [64]	V	45.4	73.2
OmniBind [64]	A+V	46.2	76.2
VAST [9]	A	40.3	71.7
VAST [9]	V	46.3	72.7
VAST [9]	A+V	48.1	79.6
BaryBind	A	45.7 (+5.4)	75.2 (+3.5)
BaryBind	V	48.3 (+2.0)	76.4 (+3.7)
BaryBind	A+V	**55.6** (+7.5)	**83.4** (+3.8)

We use benchmarks spanning diverse modalities to evaluate the BaryBind's capability in multimodal understanding across retrieval and classification tasks. These benchmarks include: (i) three-modality datasets DiDeMo [2] and ActivityNet [6], where video serves as the primary modality while audio and text provide auxiliary cues; (ii) four-modality datasets MSR-VTT [69] and VATEX [60], covering video (V), audio (A), text (T), subtitles (S) and we additionally introduce depth (D) modality, which is derived using ChronoDepth [49] and integrated via an additional lightweight head attached to the vision encoder; and (iii) audio-centered dataset VGGSound [8], where audio plays the dominant role and complementary information is also available in visual and textual forms. T-VAS denotes that the WB embedding is derived from the text, while the concatenated VAS embeddings serve as the condition. More details, evaluations on multimodal tasks, covering *video-QA* and *cross-modal generation*, and ablation studies are provided in **Appendix B.2**.

4.1.1 Zero-shot video/audio classification

We first evaluate BaryBind on VGGSound5K to assess its multimodal understanding ability in the zero-shot setting, particularly when jointly modeling video and audio signals. As shown in Table 1, BaryBind achieves the best top-1 and top-5 accuracy across all modality configurations, which verifies its strong generalization ability across modalities. It reaches (55.6%/83.4%) on the classification integrating video and audio, significantly outperforming the VAST baseline (48.1%/79.6%) and other competitors such as OmniBind (46.2%/76.2%), GRAM (42.3%/74.5%). The substantial gains with A+V highlight the enhanced holistic semantics understanding, enabled by the BVC loss that preserves the global geometry of n-modal data.

On the other hand, BaryBind improves the performance of the weaker audio-only modality classification (A) from 40.3% (VAST) to 45.7% for Acc@1, indicating its ability to alleviate under-optimization of weaker modalities. These results confirm that BaryBind learns a more balanced semantic space via our barycenter-based alignment strategy.

4.1.2 Cross-modal retrieval

We evaluate BaryBind on zero-shot text-to-video (T2V) and video-to-text (V2T) retrieval across four benchmarks. As shown in Table 2, BaryBind consistently achieves state-of-the-art results under all modality configurations. For instance, under the T-VA setting, it achieves 56.1 and 54.0 R@1 on DiDeMo (T2V/V2T), outperforming all prior methods. With T-VAS, BaryBind further sets new records on MSR-VTT and VATEX, demonstrating strong retrieval performance across both directions.

Table 2: **Zero-shot** text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall at 1 score (R@1). Results from our baseline VAST and BaryBind are highlighted accordingly.

Methods	Modality	MSR-VTT		DiDeMo		ActivityNet		VATEX	
		T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
X-CLIP [39]	T-V	46.1	46.8	45.2	42.3	44.3	42.6	-	-
ImageBind [17]	T-V	36.8	-	-	-	-	-	-	-
ViCLIP [62]	T-V	42.4	41.3	18.4	27.9	15.1	24.0	-	-
VideoPrism-b [72]	T-V	51.4	50.2	-	-	49.6	47.9	62.5	77.1
LanguageBind [73]	T-V	44.8	40.9	39.9	39.8	41.0	39.1	-	-
InternVL [10]	T-V	46.3	42.4	43.7	42.2	45.1	42.4	66.8	69.3
OmniBind [64]	T-V	47.4	45.2	43.5	42.6	44.3	40.8	-	-
NarVid [20]	T-V	51.8	50.3	52.4	50.5	51.8	46.6	73.8	76.3
Video-ColBERT [47]	T-V	51.9	48.8	51.7	50.1	52.7	47.8	72.4	73.7
VAST [9]	T-VA	49.3	43.7	49.5	48.2	51.4	46.8	80.0	77.3
VAST [9]	T-VAS	50.9	47.9	-	-	-	-	82.1	78.7
VAST [9]	T-VASD	51.2	48.3	-	-	-	-	82.4	79.2
GRAM [12]	T-VAS	54.2	51.6	-	-	-	-	83.2	81.9
GRAM [12]	T-VASD	54.9	51.3	-	-	-	-	83.8	82.1
Triangle [11]	T-VA	54.8	51.9	54.6	52.2	59.6	54.4	84.0	80.1
PMRL [35]	T-VAS	54.6	51.4	50.6	48.4	56.0	49.6	80.5	75.2
BaryBind	T-V	54.2	52.0	54.8	52.9	59.3	52.8	82.9	80.4
BaryBind	T-VA	56.1	53.6	**56.3**	**54.0**	**60.6**	**57.2**	84.7	81.8
BaryBind	T-VAS	57.1	53.8	-	-	-	-	85.5	84.3
BaryBind	T-VASD	**57.8**	**54.6**	-	-	-	-	**86.1**	**85.4**
InternVideo2-6B [63]	T-VA	54.9	49.1	55.7	51.6	61.2	52.8	82.7	76.4

Table 3: Text-to-audio (T2A) and audio-to-text (A2T) R@1 retrieval results on AudioCaps **w/ and w/o audio during training**.

Method	Training modality	T2A	A2T
VAST	T-VA	32.1	26.1
GRAM		33.2	27.4
Triangle		32.2	28.1
BaryBind		**35.7**	**32.5**
vs. baseline (VAST)		**+3.6**	**+6.4**
VAST	T-V	10.4	6.7
GRAM		12.8	7.3
Triangle		11.6	7.1
SMIL		13.3	9.5
ShaSpec		16.6	10.8
BaryBind		**21.2**	**14.5**
vs. baseline (VAST)		**+10.8**	**+ 7.8**

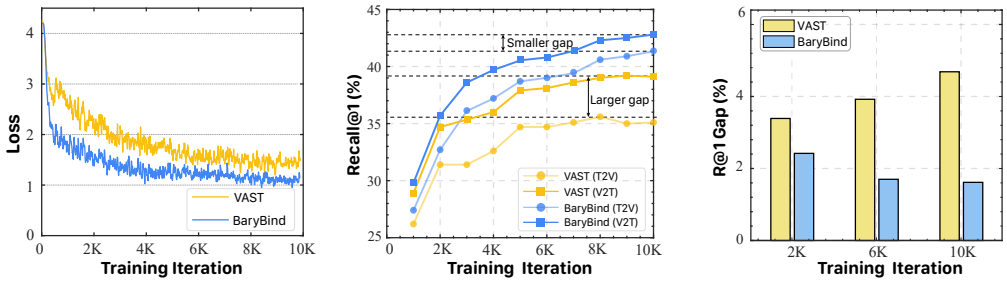
Table 4: Multimodal event classification on VGGSound5K **w/ and w/o video during inference** when training with audio and video.

Method	Inference modality	Acc@1	Acc@5
VAST	A+V	48.1	79.6
GRAM		42.3	76.4
Triangle		44.8	80.0
BaryBind		**55.6**	**83.4**
vs. baseline (VAST)		**+7.5**	**+3.8**
VAST	A	40.8	71.6
GRAM		38.5	70.1
Triangle		40.1	70.9
SMIL		43.5	70.2
ShaSpec		46.2	74.5
BaryBind		**49.4**	**78.3**
vs. baseline (VAST)		**+8.6**	**+6.7**

In addition to these overall gains, BaryBind effectively reduces the T2V/V2T performance gap, such as 2.8 on MSR-VTT under the T-VA setting vs 5.6 of the VAST baseline, which highlights improved balanced alignment. This suggests that BaryBind learns more balanced, unified representations, which can be attributed to the WB optimization and volumetric alignment process for establishing a balanced unified semantic space.

4.2 Robustness to Missing Modalities

Barybind is designed to capture balanced representations from arbitrary subsets of multimodal inputs, enabling the aligned representations of available modalities to serve as proxy features when others are missing. We evaluate BaryBind under two settings: (1) text-to-audio (T2A) and audio-to-text (A2T) retrieval on AudioCaps [23] with missing audio during training, and (2) multimodal event classification with missing video during inference on VGGSound, where the model is trained using both video and audio. In both cases, the barycenter is optimized using only the accessible modalities. Results in Table 3 and 4 show that BaryBind preserves robust and graceful degradation under missing-modality conditions as compared to traditional anchor-based methods VAST [9], GRAM [12], and missing modality-tailored methods SIML [38] and ShaSpec [57], which demonstrate the effectiveness of our barycenter modeling for the shared semantics.



(a) Total training loss comparison (b) Recall@1 w.r.t training steps (c) V2T/T2V retrieval gap

Figure 5: **Training dynamics comparison with cosine-based VAST:** (a) Training loss comparison between BaryBind and VAST. (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 T2V/V2T R@1 gap). (c) V2T/T2V retrieval gaps throughout training.

4.3 Ablation Studies

Table 5: **Ablation study on loss functions.** We report the zero-shot top-1 classification accuracy (Acc@1) on VGGSound5K and top-1 retrieval recall at 1 score (R@1) on MSR-VTT. The performance gains over the baseline are highlighted accordingly.

Loss function components				Classification			Retrieval	
				VGGSound			MSR-VTT	
TV+TA CL	MWB	BVC	DAM	A	V	V+A	T2V	V2T
✓	✗	✗	✗	38.1	42.8	44.5	46.8	40.1
✓	✗	✗	✓	40.3	43.6	46.3	49.3	43.7
✓	✓	✗	✗	43.6 (+5.5)	45.6	48.6	49.1	47.3 (+7.2)
✓	✓	✗	✓	44.3 (+4.0)	46.8	49.8	50.6	48.8 (+5.1)
✗	✗	✓	✗	42.2	44.8	47.2	50.6	46.4
✗	✗	✓	✓	42.9	47.1	50.1	51.5	46.8
✗	✓	✓	✗	45.2 (+7.1)	47.8 (+5.0)	54.7 (+10.2)	54.8 (+8.0)	51.1 (+11.0)
✗	✓	✓	✓	45.7 (+5.4)	48.3 (+4.7)	55.6 (+9.3)	56.1 (+6.8)	53.6 (+9.9)

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. 5, 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 embeddings toward a unified barycenter. Such training behavior reflects our barycentric optimization converges stably, and also demonstrates the effectiveness of the barycenter-based alignment in constructing a balanced representation space.

Impact of loss functions. To evaluate the contribution of each loss component, we conduct an ablation study on VGGSound [8] for audio/video classification and on MSR-VTT [69] for text–video retrieval under a tri-modal setting (text, video, audio). The baseline adopts pairwise cosine contrastive losses (TV+TA CL). As summarized in Table 5, 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 48.8%. In this sense, MWB constructs a barycentric semantic anchor to filter out modality-specific biases from the original anchor (e.g., text). The barycenter volume contrastive (BVC) loss further enhances global consistency among modalities, enabling full-modality gains such as 55.6% in audio classification and 56.1/53.6 R@1 for T2V/V2T retrieval. Overall, the combination of MWB and BVC promotes coordinated alignment to the WB, yielding a more balanced representation space.

Effect of batch size. Since BaryBind performs distribution-level alignment through Wasserstein barycentric optimization, we investigate the effect of batch size on the learned barycentric space. The training is conducted on two A100 GPUs without gradient accumulation. We evaluate under the T-VA setting on MSR-VTT retrieval and VGGSound classification. As shown in Table 6, the performance remains stable across different batch sizes. Although a very small batch size (16) introduces a slight performance decrease due to noisier empirical distribution estimation, the overall gap is marginal. When the batch size is larger than 32, the performance becomes highly consistent, demonstrating the robustness of the learned barycentric representation under moderate batch-size variations.

Table 6: Ablation on batch size.

Batch Size	T2V/V2T R@1	A+C Acc@1
16	55.2/52.6	54.5
32	55.8/53.4	55.1
64	55.9/53.2	55.4
128	56.1/53.5	55.8
256	56.1/53.6	55.6

5 Conclusion

In this work, we tackle the multimodal representation learning problem and propose BaryBind, a novel approach that aligns multiple modalities to a multimodal Wasserstein barycenter through a volumetric alignment objective. We demonstrate that this barycenter-based alignment strategy is scalable and effectively alleviates modality imbalance in complex scenarios involving more than three modalities. Moreover, BaryBind exhibits robustness in missing-modality settings. Overall, this work advances multimodal learning by introducing a barycenter-aligned joint representation that enables more balanced and effective multimodal understanding. In the future, we plan to further incorporate reconstruction into this framework to enable joint understanding and generation.

Acknowledgments and Disclosure of Funding

This work was supported in part by National Key R&D Program under Grant 2021YFA1003000, in part by NSFC under Grant 125B2028, 12125104, 12426313, 12501709, and 12401671, in part by the National Postdoctoral Program for Innovative Talents under Grant BX20240276, in part by China Postdoctoral Science Foundation under Grant 2025M773058, and in part by the Fundamental Research Funds for the Central Universities, China under Grant xzy022025047.

References

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

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2018.

[4] Florian Beier, Robert Beinert, and Gabriele Steidl. Multi-marginal gromov–wasserstein transport and barycentres. *Information and Inference: A Journal of the IMA*, 12(4):2753–2781, 2023.

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

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[7] Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *JMLR*, 26(141):1–47, 2025.

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020.

[9] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *NeurIPS*, 2023.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024.

[11] Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. *arXiv preprint arXiv:2509.24734*, 2025.

[12] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *ICLR*, 2025.

[13] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, pages 685–693. PMLR, 2014.

[14] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, pages 1–5, 2023.

[15] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

[16] Felix R Gantmacher. Matrix theory. *Chelsea, New York*, 21:6, 1959.

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind one embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33: 21271–21284, 2020.

[19] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, pages 976–980. IEEE, 2022.

[20] Chan Hur, Jeong hun Hong, Dong hun Lee, Dabin Kang, Semin Myeong, Sang hyo Park, and Hyeyoung Park. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *CVPR*, 2025.

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

[22] Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[23] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *ACL*, pages 119–132, 2019.

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

[25] Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Anastasis Kratsios, Evgeny Burnaev, and Aleksandr Korotin. Energy-guided continuous entropic barycenter estimation for general costs. *NeurIPS*, 37:107513–107546, 2024.

[26] Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *ICLR*, 2021.

[27] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[28] Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M Solomon. Continuous regularized wasserstein barycenters. In *NeurIPS*, volume 33, pages 17755–17765, 2020.

[29] Changshuo Liu, Jiaqi Zhu, Wenqiao Zhang, Xiaokui Xiao, and Beng Chin Ooi. Coda: Cohort- and drift-aware foundation model for multimodal clinical reasoning. In *NeurIPS*, 2026.

[30] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE TPAMI*, 2024.

[31] Junfeng Liu and Xingquan Li. ChiP-STAR: Spatial-topological attention for pre-trained generative chip routing. In *Advances in Neural Information Processing Systems*, volume 39, 2026.

[32] Junfeng Liu, Min Zhou, Shuai Ma, and Lujia Pan. MATA*: Combining learnable node matching with A* algorithm for approximate graph edit distance computation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1503–1512, 2023.

[33] Mengmeng Liu, Jiuming Liu, Michael Ying Yang, Chaokang Jiang, Jiangtao Li, Yunpeng Zhang, Hesheng Wang, Francesco Nex, and Hao Cheng. Streamvlo: Streaming visual–lidar odometry with cumulative drift compensation. In *CVPR*, pages 39086–39097, 2026.

[34] Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Francesco Nex, Hao Cheng, and Michael Ying Yang. UNIVERSE: Unified video action models for autonomous driving with flexible mask-modulated modality generation. *arXiv preprint arXiv:2607.05133*, 2026.

[35] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *IEEE TPAMI*, 2026.

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

[37] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *CVPR*, pages 26752–26762, 2024.

[38] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2302–2310, 2021.

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

[40] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[41] Zijie Meng. Decoupling semantics from distortions: Multi-scale two-stream vision-language alignment for AI-generated image quality assessment. *arXiv preprint arXiv:2606.16799*, 2026.

[42] Zijie Meng, Jiwen Liu, Yufei Liu, Chengzhuo Tong, Xiaoqiang Liu, Yuanxing Zhang, Yulong Xu, and Pengfei Wan. ARGUS: Stacked multi-view identity mosaic injection for subject-preserving video generation. *arXiv preprint arXiv:2606.11670*, 2026.

[43] Zijie Meng, Yufei Liu, Chengqian Ma, Zhiyu Li, Jiyuan Liu, Wenhua Nie, Bingcai Wei, Shuqin Chen, Weichen Xu, Jiquan Yuan, et al. Omnidrive: An LLM-choreographed multi-agent world model with unified latent co-compression for multi-view driving video generation. *arXiv preprint arXiv:2606.17536*, 2026.

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

[45] Moritz Piening and Robert Beinert. A novel sliced fused gromov-wasserstein distance. In *AAAI*, volume 40, pages 24828–24836, 2026.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[47] Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *CVPR*, 2025.

[48] Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *NeurIPS*, 2024.

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

[50] Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. In *ICLR*, 2025.

[51] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 21(1):171–176, 1958.

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

[53] Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE TPAMI*, 47(8):1–16, 2025.

[54] Xiaole Tang, Xiaoyi He, Xiang Gu, and Jian Sun. Baryir: Learning multi-source unified representation in continuous barycenter space for generalizable all-in-one image restoration. *arXiv preprint arXiv:2505.21637*, 2025.

[55] Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. Learning continuous wasserstein barycenter space for generalized all-in-one image restoration. *IEEE TPAMI*, 2026.

[56] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[57] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, pages 15878–15887, 2023.

[58] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail A. Dianat, Majid Rabbani, Raghuveer M. Rao, and Zhiqiang Tao. Text Is MASS: Modeling as stochastic embedding for text-video retrieval. *CVPR*, pages 16551–16560, 2024.

[59] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advancing long description understanding for video clip models. In *EMNLP*, pages 16061–16075, 2024.

[60] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019.

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

[62] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024.

[63] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling video foundation models for multimodal video understanding. *ArXiv preprint: 2403.15377*, 2024.

[64] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. In *ICLR*, 2025.

[65] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

[66] Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *NeurIPS*, 2025.

[67] Chenghao Xu, Jiexi Yan, and Cheng Deng. Keep and extent: Unified knowledge embedding for few-shot image generation. *IEEE TIP*, 34:2315–2324, 2025. doi: 10.1109/TIP.2025.3557578.

[68] Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qiuchen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren Zhou. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In *ICML*, 2023.

[69] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[70] Tianrun Xu, Guanyu Chen, Ye Li, Yuxin Xi, Zeyu Mu, Ruichen Wang, Tianren Zhang, Haichuan Gao, and Feng Chen. Ouro: A self-bootstrapped framework for enhancing multimodal scene understanding. In *ICCV*, pages 18240–18251, 2025.

[71] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[72] Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding. In *ICML*, 2024.

[73] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024.

A Theoretical Results

A.1 Proof of theorem 3.1

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

$$\mathcal{L}^* = \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{M}_B)} \sup_{f_0,\dots,f_k\in\mathcal{C}(\mathcal{M}_B)} \underbrace{\sum_{k=1}^{K} \lambda_k \left\{ \int_{\mathcal{M}_k} f_k^{C_k}(\boldsymbol{m}_k) \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k) + \int_{\mathcal{M}_B} f_k(\boldsymbol{b}) \mathrm{d}\mathbb{Q}(\boldsymbol{b}) \right\}}_{\triangleq \widetilde{\mathcal{F}}(\mathbb{Q}, f_{1:K})}. \quad (13)$$

where

$$f_k^{C_k}(\boldsymbol{m}_k) = \inf_{b\in\mathcal{M}_B} [C_k(\boldsymbol{m}_k, \boldsymbol{b}) - f_k(\boldsymbol{b})]. \quad (14)$$

We denote the expression under the inf and inf sup in (13) as functionals $\mathcal{L} : \mathcal{C}(\mathcal{M}_B)^K$ and $\widetilde{\mathcal{F}} : \mathcal{P}(\mathcal{M}_B) \times \mathcal{C}(\mathcal{M}_B)^K$, respectively. For simplicity, we also introduce the following notations

$$\bar{f} \triangleq \sum_{k=1}^{K} \lambda_k f_k \quad \text{and} \quad M \triangleq \inf_{\boldsymbol{b}\in\mathcal{M}_B} \bar{f}(\boldsymbol{b}) = \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{M}_B)} \int_{\mathcal{M}_B} \bar{f}(\boldsymbol{b}) d\mathbb{Q}(\boldsymbol{b}), \quad (15)$$

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

Firstly, due to the compactness of $\mathcal{M}_B$, the space $\mathcal{P}(\mathcal{M}_B)$ is compact with respect to the weak topology. For fixed potentials $f_{1:K} \in \mathcal{P}(\mathcal{M}_B)^{K+1}$ we have that $\widetilde{\mathcal{F}}(\cdot, f_{1:K})$ is linear, convex and continuous. Secondly, for a fixed $\mathbb{Q}$, the functional $\widetilde{\mathcal{F}}(\mathbb{Q}, \cdot)$ is a concave due to the concavity of C-transform. These properties enable the application of Sion's minimax theorem ([51], Theorem 3.4), which allows the interchange of the sup and inf in (13). Thus with (15) we obtain

$$\begin{aligned}
\mathcal{L}^* &= \sup_{f_0,\dots,f_K\in\mathcal{C}(\mathcal{M}_B)} \underbrace{\inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{M}_B)} \sum_{k=1}^{K} \lambda_k \left\{ \int_{\mathcal{M}_k} f_k^{C_k}(\boldsymbol{m}_k) \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k) + \int_{\mathcal{M}_B} f_k(\boldsymbol{b}) \mathrm{d}\mathbb{Q}(\boldsymbol{b}) \right\}}_{\triangleq \mathcal{L}(f_{1:K})} \\
&= \sup_{f_0,\dots,f_K\in\mathcal{C}(\mathcal{M}_B)} \left\{ \sum_{k=1}^{K} \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\boldsymbol{m}_k) \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k) + \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{M}_B)} \int_{\mathcal{M}_B} \bar{f}(\boldsymbol{b}) \mathrm{d}\mathbb{Q}(\boldsymbol{b}) \right\} \\
&= \sup_{f_0,\dots,f_K\in\mathcal{C}(\mathcal{M}_B)} \underbrace{\left\{ \sum_{k=1}^{K} \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\boldsymbol{m}_k) \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k) + \inf_{\boldsymbol{b}\in\mathcal{M}_B} \bar{f}(\boldsymbol{b}) \right\}}_{\triangleq \widetilde{\mathcal{L}}(f_{1:K})}. \quad (16)
\end{aligned}$$

Now we show that the sup in (16) can be restricted to potentials $\tilde{f}_{1:K}$ which satisfy the congruence condition $\sum_{k=1}^{K} \lambda_k \tilde{f}_k = 1$. It is enough to show that for every tuple $f_{1:K}$ there exists a congruent tuple $\tilde{f}_{1:K} \in \mathcal{C}(\mathcal{M}_B)^K$ such that $\widetilde{\mathcal{L}}(\tilde{f}_{1:K}) \geq \widetilde{\mathcal{L}}(f_{1:K})$.

For this, we consider the congruent potentials given any tuple $f_{1:K}$

$$(\tilde{f}_0, \cdots, \tilde{f}_K) = \left(f_0, \cdots, f_{K-1}, f_K - \frac{\bar{f}}{\lambda_K} \right). \quad (17)$$

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

$$\begin{aligned}
\widetilde{\mathcal{L}}(\tilde{f}_{1:K}) - \widetilde{\mathcal{L}}(f_{1:K}) &= \lambda_K \int_{\mathcal{M}_k} \left(\tilde{f}_K^{C_K}(\boldsymbol{m}_K) - f_K^{C_K}(\boldsymbol{m}_K)\right) d\mathbb{P}_K(\boldsymbol{m}_K) - M \\
&= \lambda_K \int_{\mathcal{M}_k} \left[\left(f_K - \frac{\bar{f}}{\lambda_K}\right)^{C_K}(\boldsymbol{m}_K) - f_K^{C_K}(\boldsymbol{m}_K)\right] d\mathbb{P}_K(\boldsymbol{m}_K) - M \\
&\geq \lambda_K \int_{\mathcal{M}_k} \left[\left(f_K - \frac{M}{\lambda_K}\right)^{C_K}(\boldsymbol{m}_K) - f_K^{C_K}(\boldsymbol{m}_K)\right] d\mathbb{P}_K(\boldsymbol{m}_K) - M \\
&= \lambda_K \int_{\mathcal{M}_k} \frac{M}{\lambda_K} d\mathbb{P}_K(\boldsymbol{m}_K) - M = 0, \qquad (18)
\end{aligned}$$

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

Finally, since $\widetilde{\mathcal{L}}(f_{1:K}) = \mathcal{L}(f_{1:K})$ for congruent potentials $f_{1:K}$, with (14) we obtain

$$\begin{aligned}
\mathcal{L}^* = \sup_{\sum_k \lambda_k f_k = 0} \mathcal{L}(f_{1:K}) &= \sup_{\sum_k \lambda_k f_k = 0} \sum_{k=1}^{K} \lambda_k \int_{\mathcal{M}_k} f_k^{C_k}(\boldsymbol{m}_k) \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k), \\
&= \sup_{\sum_k \lambda_k f_k = 0} \sum_{k=1}^{K} \lambda_k \inf_{\boldsymbol{b}\in\mathcal{M}_B} \int_{\mathcal{M}_k} \left[C_k(\boldsymbol{m}_k - \boldsymbol{b}) - f_k(\boldsymbol{b})\right] \mathrm{d}\mathbb{P}_k(\boldsymbol{m}_k). \qquad (19)
\end{aligned}$$

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

$$\mathcal{L}^* = \sup_{\sum_k \lambda_k f_k = 0} \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{M}_B)} \sum_{k=1}^{K} \lambda_k \mathop{\mathbb{E}}_{\substack{\boldsymbol{m}_k\sim\mathbb{P}_k \\ \boldsymbol{b}\sim\mathbb{Q}}} \left[\|\boldsymbol{m}_k - \boldsymbol{b}\| - f_k(\boldsymbol{b})\right], \qquad (20)$$

which completes the proof. □

A.2 Proof of Theorem 3.2

Proof. Let $\mathbb{P}_{1:K}$ denote the joint distribution over the multimodal tuples $(\boldsymbol{m}_0, \ldots, \boldsymbol{m}_K) =: \boldsymbol{m}_{1:K}$, where $\boldsymbol{m}_k \in \mathcal{M}_k$. The barycenter map is a function $T : \mathcal{M}_0 \to \mathcal{M}_B$ that acts on the anchor modality. For a set of potential functions $f_{1:K}$ and a map T, we rewrite the total cost functional $\mathcal{F}$ and its corresponding minimal cost functional $\mathcal{L}$ as follows:

$$\mathcal{F}(f_{1:K}, T) \triangleq \mathbb{E}_{\boldsymbol{m}_{1:K}\sim\mathbb{P}_{1:K}} \left[\sum_{k=1}^{K} \lambda_k \left(C_k(\boldsymbol{m}_k, T(\boldsymbol{m}_0)) - f_k(T(\boldsymbol{m}_0))\right)\right], \qquad (21)$$

$$\mathcal{L}(f_{1:K}) \triangleq \inf_{T:\mathcal{M}_0\to\mathcal{M}_B} \mathcal{F}(f_{1:K}, T). \qquad (22)$$

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

$$T^f \in \mathop{\arg\inf}_{T:\mathcal{M}\to\mathcal{M}_B} \mathcal{F}(\widehat{f}_{1:K}, T). \qquad (23)$$

Given $T^f : \mathcal{M}_0 \to \mathcal{M}_B$, the functional $\mathcal{L}$ in (22) can be written as:

$$\mathcal{L}(\widehat{f}_{1:K}) = \mathcal{F}(\widehat{f}_{0:k}, T^f). \qquad (24)$$

We can observe that the first gap $\mathcal{E}_1$ is the difference between (21) and (24):

$$\mathcal{E}_1(\widehat{f}_{1:K}, \widehat{T}) = \mathcal{F}(\widehat{f}_{1:K}, \widehat{T}) - \mathcal{L}(\widehat{f}_{1:K}) = \mathcal{F}(\widehat{f}_{1:K}, \widehat{T}) - \mathcal{F}(\widehat{f}_{1:K}, T^f). \qquad (25)$$

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

$$\mathcal{L}^* \triangleq \sum_{k=1}^{K} \lambda_k \mathrm{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}^*). \tag{26}$$

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

$$\mathcal{L}^* = \sum_{k=1}^{K} \lambda_k \int_{\mathcal{M}_k} C_k(\boldsymbol{m}_k, T^*(\boldsymbol{m}_0)) d\mathbb{P}_k(\boldsymbol{m}_k). \tag{27}$$

Due to the congruence condition on the potentials $\widehat{f}_{1:K}$ and the property $T^*_{\#}\mathbb{P}_0 = \mathbb{Q}^*$ for all k, we have:

$$\sum_{k=1}^{K} \lambda_k \mathbb{E}_{\boldsymbol{m}_k \sim \mathbb{P}_k}[\widehat{f}_k(T^*(\boldsymbol{m}_0))] = \mathbb{E}_{\boldsymbol{b} \sim \mathbb{Q}^*}\left[\sum_{k=1}^{K} \lambda_k \widehat{f}_k(\boldsymbol{b})\right] = 0. \tag{28}$$

This allows us to reformulate the optimal value $\mathcal{L}^*$. Using the definition of $\mathcal{F}$ in (21), we find:

$$\begin{aligned}
\mathcal{L}^* &= \sum_{k=1}^{K} \lambda_k \mathbb{E}_{\boldsymbol{m}_k \sim \mathbb{P}_k}[C_k(\boldsymbol{m}_0, T^*(\boldsymbol{m}_k))] - \underbrace{\mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k \widehat{f}_k(T^*(\boldsymbol{m}_0))\right]}_{=0 \text{ from } (28)} \\
&= \mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k \left(C_k(\boldsymbol{m}_k, T^*(\boldsymbol{m}_0)) - \widehat{f}_k(T^*(\boldsymbol{m}_0))\right)\right] = \mathcal{F}(\widehat{f}_{1:K}, T^*).
\end{aligned}$$

With (24) we derive the second gap $\mathcal{E}_2$ can be written as

$$\mathcal{E}_2 = \mathcal{L}^* - \mathcal{L}(\widehat{f}_{1:K}) = \mathcal{F}(\widehat{f}_{1:K}, T^*) - \mathcal{F}(\widehat{f}_{1:K}, T^f). \tag{29}$$

We introduce the function $g_k(\boldsymbol{m}_k, \boldsymbol{b}) \triangleq C_k(\boldsymbol{m}_k, \boldsymbol{b}) - \widehat{f}_k(\boldsymbol{b})$, which is assumed to be β-strongly convex with respect to $\boldsymbol{b}$. Using this, we can rewrite our total cost functional $\mathcal{F}$ from (21) as:

$$\mathcal{F}(f_{1:K}, T) = \mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k g_k(\boldsymbol{m}_k, T(\boldsymbol{m}_0))\right]. \tag{30}$$

As a result of convexity, it follows that a necessary condition for T^f to minimize $\mathcal{F}(f_{1:K}, T)$ is the vanishing of its first variation, yielding

$$\mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k \nabla_{\boldsymbol{b}} g_k(\boldsymbol{m}_k, T^f(\boldsymbol{m}_0))\right] = 0. \tag{31}$$

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

$$\begin{aligned}
\mathcal{E}_1 &= \mathcal{F}(\widehat{f}_{1:K}, \widehat{T}) - \mathcal{F}(\widehat{f}_{1:K}, T^f) \\
&= \mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k \left(g_k(\boldsymbol{m}_k, \widehat{T}(\boldsymbol{m}_0)) - g_k(\boldsymbol{m}_k, T^f(\boldsymbol{m}_0))\right)\right] \\
&\geq \mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\sum_{k=1}^{K} \lambda_k \left(\langle \nabla_{\boldsymbol{b}} g_k(\boldsymbol{m}_k, T^f(\boldsymbol{m}_0)), \widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\rangle + \frac{\beta}{2}\|\widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\|^2\right)\right] \\
&= \mathbb{E}_{\boldsymbol{m}_{1:K} \sim \mathbb{P}_{1:K}}\left[\left\langle \sum_{k=1}^{K} \lambda_k \nabla_{\boldsymbol{b}} g_k(\boldsymbol{m}_k, T^f(\boldsymbol{m}_0)), \widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\right\rangle\right] + \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0 \sim \mathbb{P}_0}\left[\|\widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\|^2\right] \\
&\overset{(31)}{=} 0 + \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0 \sim \mathbb{P}_0}\left[\|\widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\|^2\right].
\end{aligned} \tag{32}$$

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

$$\mathcal{E}_2 \geq \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0\sim\mathbb{P}_0}\left[\|T^f(\boldsymbol{m}_0) - T^*(\boldsymbol{m}_0)\|^2\right]. \tag{33}$$

Now we sum the inequalities for $\mathcal{E}_1$ (32) and $\mathcal{E}_2$ (33):

$$\begin{aligned}
\mathcal{E}_1 + \mathcal{E}_2 &\geq \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0\sim\mathbb{P}_0}\left[\|\widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\|^2\right] + \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0\sim\mathbb{P}_0}\left[\|T^f(\boldsymbol{m}_0) - T^*(\boldsymbol{m}_0)\|^2\right] \\
&= \frac{\beta}{2}\mathbb{E}_{\boldsymbol{m}_0\sim\mathbb{P}_0}\left[\|\widehat{T}(\boldsymbol{m}_0) - T^f(\boldsymbol{m}_0)\|^2 + \|T^f(\boldsymbol{m}_0) - T^*(\boldsymbol{m}_0)\|^2\right] \\
&\geq \frac{\beta}{4}\mathbb{E}_{\boldsymbol{m}_0\sim\mathbb{P}_0}\left[\|\widehat{T}(\boldsymbol{m}_0) - T^*(\boldsymbol{m}_0)\|^2\right] \geq \frac{\beta}{4}W^2(\widehat{T}_\#\mathbb{P}_0, T^*_\#\mathbb{P}_0) = \frac{\beta}{4}W^2(\widehat{T}_\#\mathbb{P}_0, \mathbb{Q}^*).
\end{aligned} \tag{34}$$

□

A.3 Analysis on barycenter simplex volume

Given the matrix containing vectors spanning the barycenter simplex

$$\mathbf{R} = (\boldsymbol{b}, \boldsymbol{r}_1, \cdots, \boldsymbol{r}_K),$$

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

$$\langle \boldsymbol{b}, \boldsymbol{b}\rangle = \langle \boldsymbol{r}_1, \boldsymbol{r}_1\rangle = \langle \boldsymbol{r}_2, \boldsymbol{r}_2\rangle = 1. \tag{35}$$

We define the cosine of the angles between these vectors as

$$\cos\theta := \langle \boldsymbol{b}, \boldsymbol{r}_1\rangle, \quad \cos\beta := \langle \boldsymbol{b}, \boldsymbol{r}_2\rangle, \quad \cos\gamma := \langle \boldsymbol{r}_1, \boldsymbol{r}_2\rangle. \tag{36}$$

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

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

Therefore, the simplex volume in the bimodal case reduces to:

$$\mathrm{Vol}_{n=2} = \sqrt{1 - \cos^2\theta} = \sqrt{\sin^2\theta} = \sin\theta,$$

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

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

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

Substituting (35) and (36) into the determinant, the expression simplifies to

$$\begin{aligned}
\mathrm{Vol}^2_{n=3} &= 1 - \cos^2\theta - \cos^2\beta - \cos^2\gamma + 2\cos\theta\cos\beta\cos\gamma, \\
&= \sin^2\theta + \sin^2\beta + \sin^2\gamma + 2\cos\theta\cos\beta\cos\gamma - 2.
\end{aligned}$$

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

The volume becomes small when each modality vector closely follows the barycenter direction (small angles α and β). In addition, if the gap vectors between modalities, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, are nearly aligned (small inter-modality angle γ), 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.

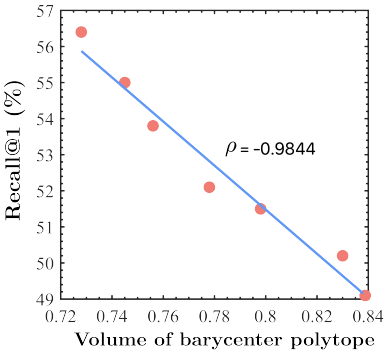


Figure 6: The simplex's volume is correlated ($\rho = -0.9844$) with the downstream retrieval performance.

Unlike pairwise cosine similarity, which only captures alignment between two modalities, the barycenter simplex 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.

Geometric Intuition. Geometrically, traditional methods rely on asymmetric alignment, forcing non-anchor modalities (e.g., video) to conform to a modality-specific anchor (e.g., text). In contrast, BaryBind establishes a shared "geometric consensus" by optimizing a Wasserstein barycenter across all modalities. Instead of leaning towards a fixed specific modality, the multimodal embeddings of BaryBind converge toward the central manifold supported by the refined WB anchor. This balanced constraint shares the optimization burden and mitigates modality-specific bias, logically explaining the reduced T2V/V2T performance gap observed in our experiments.

B Experimental Details and More Results

B.1 Training algorithm

Algorithm 1 Training Algorithm

Input: Per-modality encoders $E_{\phi_{1:K}}$; Multimodal data set $\mathcal{S}_{1:K}$; NN-based WB map T_θ; NN-based potentials $f_{\omega_{1:K}}$; Learning rate γ

Output: Learned T_θ,$E_{\phi_{1:K}}$,$f_{\omega_{1:K}}$

while not converged **do**
  Sample multimodal batch data $\boldsymbol{z}_{1:K} \sim \mathcal{S}_{1:K}$
  Extract embeddings $\{\boldsymbol{m}_k = E_{\phi_k}(\boldsymbol{z}_k) \sim \mathbb{P}_k\}_{k=1}^K$
  Compute WB anchor $\boldsymbol{b} = T_\theta(\boldsymbol{m}_0)$
  Compute $\mathcal{L}_{\text{MWB}}$ by Eqn. (5)
  Compute $\mathcal{L}_{\text{BVC}}, \mathcal{L}_{\text{DAM}}, \mathcal{L}$ by Eqns. (10), (11), (12)
  # Optimize network blocks
  $\omega_k = \omega_k + 2\gamma \frac{\partial \mathcal{L}_{\text{MWB}}}{\partial \omega_k}, k = 1, \cdots, K$
  $\phi_k = \phi_k - \gamma \frac{\partial \mathcal{L}}{\partial \phi_k}, k = 1, \cdots, K$
  $\theta = \theta - \gamma \frac{\partial \mathcal{L}}{\partial \theta}$
end while

B.2 Experimental Details

We perform both continued pretraining on top of VAST models and training from scratch of the BaryBind model using 2×A100 GPUs. For continued pretraining, We sampled 150,000 examples from the VAST-27M dataset [9] due to the unavailability of the full dataset. Each example consists of a video (comprising frames and an audio track) paired with a corresponding caption. We further evaluate the training-from-scratch models on MSR-VTT and ActivityNet in §B.10. The model backbone follows the VAST multimodal encoder, utilizing BERT-B, modality-specific audio and

vision encoders, and other modality encoders as applicable, resulting in a total of approximately 1.3 billion parameters.

During the continued pretraining phase, we use a single frame and a single 10-second audio clip per sample; in the training-from-scratch setting, we use eight randomly sampled video frames and a single 10-second audio clip per video.

For the continued pretraining, BaryBind is trained for 10k steps on the 150k-sample subset of VAST27M. Retrieval performance is evaluated every 100 steps on the MSR-VTT test set, and the checkpoint with the best performance is selected. We employ an initial learning rate of 1e-4 with a linear decay schedule and a batch size of 256. In the training-from-scratch experiments, all encoders are trained from scratch on the MSR-VTT training dataset for 4 epochs, using an initial learning rate of 1e-4 with a linear decay schedule and a batch size of 64.

Table 7: 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

Table 7 summarizes the modality configuration, data statistics, and frame settings across benchmarks.

MSR-VTT [69] 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 [2] 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 [6] 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 [60] 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.

VGGSound5K [8] 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.

B.3 Broader Multimodal Understanding and Generation

Multimodal learning has recently expanded beyond conventional image–text understanding to complex temporal, embodied, and generative scenarios [29, 70, 43, 15, 71, 27, 31, 32]. In autonomous systems, StreamVLO integrates visual and LiDAR observations with spatio-temporal correlation modeling for robust streaming odometry [33]. UNIVERSE further unifies video and trajectory generation within a mask-modulated architecture [34]. These developments highlight the growing need to organize heterogeneous sensory and semantic information within a coherent representation space. Reliable multimodal understanding additionally requires preserving and faithfully exploiting visual evidence. Related challenges also arise in visual generation and evaluation: ARGUS aggregates multi-view identity evidence for subject-consistent video generation [42], MST-CLIPIQA decouples semantic coherence from perceptual distortions through multi-scale vision–language alignment [41], and KAE develops unified knowledge embedding for few-shot image generation [67]. Collectively, these studies demonstrate that modern multimodal systems must reconcile heterogeneous modalities, granularities, and objectives while retaining both global semantics and modality-specific information.

Table 8: **Finetuning** text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall at 1 score (R@1). Results from our baseline VAST and BaryBind are highlighted accordingly.

Methods	Modality	MSR-VTT		DiDeMo		ActivityNet		VATEX	
		T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
CLIP4Clip [36]	T-V	45.6	45.9	43.0	43.6	40.3	41.6	63.0	78.3
InternVideo-L [61]	T-V	53.1	54.4	57.9	59.1	62.2	62.8	69.8	80.6
mPLUG-2 [68]	T-V	53.1	-	56.4	-	-	-	-	-
ViCLIP [62]	T-V	52.5	51.8	49.4	50.2	49.8	48.1	-	-
T-MASS [58]	T-VA	52.7	-	53.3	-	-	-	65.6	-
VALOR-L [30]	T-VAS	54.4	-	57.6	-	63.4	-	76.9	-
VideoCLIP-XL [59]	T-V	54.6	54.0	62.3	62.7	58.4	59.2	-	-
TempMe [50]	T-V	49.0	47.6	48.0	48.4	44.9	45.3	69.6	71.8
VAST [9]	T-VA	55.8	57.6	65.6	62.0	68.8	66.7	86.9	84.1
VAST [9]	T-VAS	56.6	57.6	-	-	-	-	87.5	84.0
BaryBind	T-V	57.4	57.8	67.2	64.6	67.3	65.2	85.0	82.4
BaryBind	T-VA	61.0	61.4	**69.6**	**66.3**	**73.5**	**70.2**	88.2	85.2
BaryBind	T-VAS	**65.7**	**65.4**	-	-	-	-	**89.5**	**86.4**

BaryBind addresses this general representation challenge from an optimal-transport perspective, using a multimodal Wasserstein barycenter as a shared anchor to bind multiple modalities without collapsing their complementary structures.

B.4 Finetuning Cross-modal Retrieval Results

We present the fine-tuning results in Table 8, which shows BaryBind maintains its advantage, outperforming prior models such as InternVideo, VideoCLIP-XL, and VAST. For instance, our T-VA setting achieves 60.3 R@1 on MSR-VTT and 72.1 on ActivityNet for T2V, surpassing VAST by +4.5 and +2.3 points, respectively. The performance further improves under T-VAS, where BaryBind reaches 65.7 R@1 on MSR-VTT, indicating that our binding strategy generalizes well with additional modalities and scales effectively with fine-tuning.

B.5 Clarification on Distinction from Prior Anchor-based Approaches

Our core contribution is the introduction of Wasserstein Barycenter (WB) optimization for establishing a balanced representation space. This differs fundamentally from pure anchor-based alignment approaches (e.g., GRAM and Triangle), by introducing an additional optimization process prior to alignment. This methodology is presented in Section 3, especially Section 3.2 with theoretical analysis. Specifically, BaryBind introduces a learnable WB anchor via WB optimization. The WB optimization transports modality-specific distributions toward a barycenter, defined as the distribution closest to all modality-specific distributions under the Wasserstein distance, ensuring multimodal semantic consensus for subsequent volumetric alignment. Both comparisons and ablations validate that the WB optimization is the main contributor to the superior performance.

On the other hand, along the line of alignment, GRAM and Triangle define a parallelotope and a triangle, respectively, both centered at the origin. In contrast, BaryBind constructs a **simplex** spanned by the WB anchor and gap vectors. Contrasting this volume enforces paired multimodal data toward a shared representation. This mechanism provides an optimized geometric reference (the WB anchor), followed by contrastive alignment guided by the WB anchor, which is substantially different from pure anchor-based methods.

The advantage of the WB-based representation is further supported by results across multiple tasks, including multimodal retrieval (consistently smaller gaps across benchmarks, and improved robustness under missing-modality settings), multimodal classification, and video-QA. Ablations further show that removing WB leads to consistent performance drops, indicating that it is the primary contributor.

B.6 Scalability of BaryBind with Increasing Modality Number

We evaluate the computation efficiency of different similarity metrics with varying modality number n. For each metric, we sample $B = 64$ sets of n vectors in $\mathbb{R}^D$ with $D = 512$, simulating multimodal

Table 9: 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 $D = 512$.

Number of modality n	**2**	**3**	**4**	**5**	**10**	**20**
Pairwise cosine similarity	3.0×10^{-7}	7.0×10^{-7}	1.0×10^{-6}	1.3×10^{-6}	2.9×10^{-6}	4.9×10^{-6}
Barycenter simplex volume	4.9×10^{-6}	6.8×10^{-6}	5.9×10^{-6}	9.8×10^{-6}	3.6×10^{-5}	8.8×10^{-5}

embeddings. As shown in Table 9, the simplex volume remains computationally efficient and scales reasonably as the modality number increases. The negligible overhead incurred when extending to more modalities highlights the 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 [49] 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.

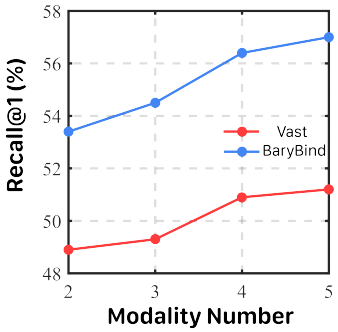


Figure 7: Zero-shot text-to-video retrieval R@1 results as scaling from 2 (T-V) modalities to 5 (T-VASD) modalities.

B.7 Robustness on different WB initializers

BaryBind learns the **Wasserstein Barycenter (WB)** anchor $\boldsymbol{b}$ that serves as a balanced representation aiming to approximate the ideal geometric semantic center. We utilize the specific modality as a semantic warm-start to initialize the barycenter mapping, and the $\boldsymbol{b}$ is derived from the initial specific modality $\boldsymbol{m_0}$, i.e.,: $\boldsymbol{b} = T_\theta(\boldsymbol{m}_0)$. This allows us to leverage the rich semantic priors to accelerate convergence, while the MWB objective ensures the final representation shifts away from the modality bias towards the geometric center. Crucially, although T_θ takes m_0 as input, its optimization gradients are derived from the Wasserstein distance to all K modalities simultaneously. Therefore, the direction of the update actively pulls the embedding out of the modality-specific manifold.

To verify the robustness of the WB initializer choice, we perform ablation on using vision, text, and T+V+A (arithmetic fusion) as the initializer across multimodal tasks. From Tables 10 and 11, we observe two findings. (1) Simply choosing a specific modality or fused embedding as the anchor without our barycenter optimization leads to inferior performance and modality imbalance. (2) With WB-driven optimization, all initializations (including the symmetric T+V+A fused initialization) gain clear performance boosts, indicating that the WB formulation itself, rather than the initial anchor choice, is the key contributor to balanced performance. Notably, the text-driven WB performs comparably to the fused-driven and video-driven ones (e.g., 55.6% vs. 55.8% and 55.2% on V+A setting of VGGSound), reflecting the high robustness of BaryBind to different initializations.

Table 10: **Ablation on the anchor selection under the T-VA setting.**

Anchor Modality	VGGSound			MSR-VTT			
	V	A	V + A	T2V		V2T	
	Acc@1	Acc@1	Acc@1	R@1	R@5	R@1	R@5
Video (w/o WB opt.)	45.8	39.6	46.5	50.9	72.6	47.4	74.8
Text (w/o WB opt.)	46.3	40.3	48.1	51.5	72.2	46.8	73.1
T+V+A (w/o WB opt.)	46.1	41.8	50.5	51.2	73.5	46.9	74.5
Video-initialized WB	47.9 (+2.8)	44.8 (+5.2)	55.2 (+8.7)	54.9 (+4.0)	74.6 (+2.0)	52.9 (+5.5)	78.5 (+3.7)
Text-initialized WB	**48.3** (+2.0)	**45.6** (+5.3)	55.6 (+7.5)	56.0 (+4.5)	**75.4** (+3.2)	53.2 (+6.4)	78.1 (+5.0)
T+V+A initialized WB	**48.3** (+2.2)	45.2 (+3.4)	**55.8** (+5.3)	**56.1** (+4.9)	75.2	**53.6** (+6.7)	**78.8** (+4.3)

Table 11: Comparison of arithmetic fusion and WB-based models on MSR-VTT and VideoQA under different modalities

Anchor	Modality	Retrieval (R@1)		VideoQA (Acc)	
		T2V	V2T	MSRVTT-QA	MSVD-QA
T+V+A (w/o WB opt.)	T-VA	51.2	46.9	47.6	55.1
T+V+A-initialized WB	T-VA	**56.1**	**53.6**	**52.0**	**60.3**
Text-initialized WB	T-VA	**56.0**	**53.2**	**52.2**	**60.3**
T+V+A+S (w/o WB opt.)	T-VAS	51.7	47.3	50.3	57.9
T+V+A+S-initialized WB	T-VAS	**57.1**	**53.8**	**55.7**	**63.8**
Text-initialized WB	T-VAS	**57.2**	**53.9**	**55.4**	**63.6**
T+V+A+S+D (w/o WB opt.)	T-VASD	51.8	47.2	–	–
T+V+A+S+D-initialized WB	T-VASD	**57.9**	**55.0**	–	–
Text-initialized WB	T-VASD	**57.8**	**55.1**	–	–

B.8 Normalization Strategies for Polytope Volume Metric

While the raw barycenter simplex volume V serves as a global alignment metric, it depends on the number of modalities and embedding dimensionality, which limits interpretability and cross-setting comparability. To address this, we consider normalization strategies such as $(V/n!)^{1/n}$ (viewing the WB simplex as an n-simplex), which scale the volume to better reflect per-modality contributions and make the metric more comparable across different modality configurations.

Table 12: Zero-shot generalization on MSR-VTT under different normalization strategies for the simplex volume.

Normalization	Setting	T2V	V2T
$(V/2!)^{1/2}$	T-V	53.9	52.2
V	T-V	54.1	52.0
$(V/3!)^{1/3}$	T-VA	54.8	52.9
V	T-VA	55.1	52.7
$(V/4!)^{1/4}$	T-VAS	57.3	54.2
V	T-VAS	57.2	54.5

Table 12 reports zero-shot retrieval results on MSR-VTT under different normalization strategies. We observe that applying normalization slightly reduces the absolute retrieval scores compared to the raw volume V, likely because normalization reduces the magnitude of the gradient signal from the volume metric, slightly weakening the alignment supervision. Nonetheless, the overall performance remains stable, indicating that $(V/n!)^{1/n}$ provides a reliable and interpretable metric without significantly sacrificing retrieval accuracy.

B.9 Hyperparameter Sensitivity Analysis

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

$$\mathcal{L} = \mathcal{L}_{\text{MWB}} + \alpha_1 \mathcal{L}_{\text{BVC}} + \alpha_2 \mathcal{L}_{\text{DAM}}. \tag{37}$$

As reported in Table 13, adjusting α_1 on MSR-VTT shows that increasing the weight of the BVC loss gradually improves retrieval performance, and the best results are achieved when $\alpha_1 = 1$. A further increase leads to a performance decline, indicating that over-restricting barycentric geometry may weaken instance-level discrimination.

Table 13: Sensitivity analysis of α_1 on MSR-VTT validation set.

α_1	0.3	0.6	1	1.5	1.8
T2V (R@1)	56.2	55.7	**56.5**	56.0	55.3
V2T (R@1)	58.0	57.5	**58.3**	57.9	56.5

Table 14: Sensitivity analysis of α_2 on MSR-VTT validation set.

α_2	0.02	0.06	0.1	0.14	0.18
T2V (R@1)	55.6	55.8	**56.5**	56.2	56.0
V2T (R@1)	57.6	57.2	**58.3**	57.9	57.7

Table 14 reveals that α_2 also has an optimal range, with $\alpha_2 = 0.1$ yielding the strongest retrieval performance. Excessively small or large weights cause suboptimal alignment due to either under-constrained or modality-biased instance matching.

Overall, $\alpha_1 = 1$ and $\alpha_2 = 0.1$ deliver the highest multimodal alignment on the MSR-VTT validation set, demonstrating the complementary strengths of BVC and DAM.

B.10 Training from Scratch Evaluation

Due to the unavailability of full VAST27M dataset [9], we adopt continuing training based on the VAST model in §4.1. Here we further evaluate the models trained from scratch on MSR-VTT and ActivityNet. Four representative models: VAST [9], which performs pairwise alignment with cosine-similarity heads and lightweight fusion layers; GRAM [12], Triangle [11], and the proposed BaryBind. Table 15 reports the R@1 results for both text-to-video (T2V) and video-to-text (V2T) retrieval. BaryBind achieves the highest performance across all settings, surpassing VAST, GRAM, and Triangle on both datasets. On MSR-VTT, BaryBind improves over VAST by +6.4% (T2V) and +3.8% (V2T), while also outperforming GRAM and Triangle by comparable margins. A similar pattern appears on ActivityNet, where BaryBind reaches 30.8% (T2V) and 32.4% (V2T). The training curves in Fig. 5 show that BaryBind continues to refine the latent space throughout training, whereas the other methods plateau earlier. These observations indicate that a barycenter-based objective provides stronger structural guidance when no pretrained semantic prior is available, enabling the model to form a more coherent multimodal representation.

Table 15: Training-from-scratch T2V and V2T results on MSR-VTT and ActivityNet. The training follows the T-VA setting.

Methods	MSR-VTT		ActivityNet	
	T2V	V2T	T2V	V2T
VAST [9]	35.0	39.1	26.7	25.2
LanguageBind [73]	31.8	36.4	29.3	26.1
GRAM [12]	35.7	39.4	27.2	29.1
Triangle [11]	39.4	40.1	27.8	29.3
BaryBind	**44.8**	**43.2**	**33.6**	**34.2**

B.10.1 Effect of Relaxing the Congruence Constraint

The proposed formulation enforces the congruence constraint $\sum_k \lambda_k f_k = 0$ through a hard parameterization. To investigate the necessity of strictly enforcing the congruence constraint, we replace the hard parameterization with a soft penalty formulation and optimize the following objective:

$$\mathcal{L}_{\text{soft}} = \mathcal{L}_{\text{MWB}} + \eta \left\| \sum_k \lambda_k f_k \right\|_2^2 . \tag{38}$$

During alternating optimization, we record the objective function for updating the potentials:

$$\mathcal{L}_f = -\sum_k \lambda_k \mathbb{E}_{m_k \sim \mathbb{P}_k} \left[f_k(T_\theta(m_k)) \right] + \eta \left\| \sum_k \lambda_k f_k \right\|_2^2 , \tag{39}$$

where $\eta = 10$ is used in the experiment. The optimization dynamics are summarized in Table 16. Although the early optimization stage exhibits noticeable fluctuations, the potential optimization gradually stabilizes after approximately 3,000 iterations without persistent divergence.

Table 16: Stability analysis of potential optimization during alternating updates.

Iterations	Mean $\mathcal{L}_f$	Range	Δ
0–500	-0.376	[-1.000,-0.154]	0.846
500–1000	-0.237	[-0.309,-0.171]	0.138
1000–2000	-0.181	[-0.242,-0.134]	0.108
2000–3000	-0.098	[-0.138,-0.045]	0.093
3000–10000	-0.030	[-0.049,-0.011]	0.038

We further compare the hard and soft formulations in Table 17. The soft penalty achieves comparable performance with only marginal degradation, demonstrating that the congruence constraint mainly enables a stable optimization.

Table 17: Comparison between hard and soft congruence constraints.

Constraint	T2V/V2T R@1	A+C Acc@1
Hard constraint	56.1/53.6	55.6
Soft constraint	55.9/53.2	55.1

B.10.2 Stability of WB Map Adaptation under Encoder Distribution Shifts

During encoder fine-tuning, the feature distributions continuously evolve. We therefore analyze the adaptation dynamics of the WB map T_θ. Specifically, we track the WB mapping loss:

$$\mathcal{L}_T \triangleq \sum_{k=1}^{K} \lambda_k \mathbb{E}_{m_k \sim \mathbb{P}_k} \left[\|m_k - T_\theta(m_0)\|_2^2 - f_{\omega_k}(T_\theta(m_0)) \right] . \tag{40}$$

As shown in Table 18, the WB map gradually adapts to the evolving encoder distributions. During the initial stage, the mapping loss changes as the WB map responds to encoder updates. A rapid adaptation phase is observed between 500 and 2,000 iterations, where the loss decreases substantially. After approximately 3,000 iterations, the variation becomes significantly smaller, indicating that the WB map reaches a stable adaptation regime with the evolving feature distributions.

Table 18: Loss analysis of WB map T_θ adaptation during encoder fine-tuning.

Iterations	Mean $\mathcal{L}_T$	Range	Δ
0–500	0.971	[0.913,1.000]	0.087
500–1000	0.602	[0.317,0.913]	0.597
1000–2000	0.136	[0.073,0.330]	0.257
2000–3000	0.052	[0.020,0.095]	0.075
3000–10000	0.029	[0.000,0.022]	0.022

B.11 Computational Overhead Comparison During Training

We provide a comparison of computational overhead among VAST, GRAM, and BaryBind. All experiments are conducted on 2×NVIDIA A100 80GB GPUs with mixed-precision (FP16/AMP). The additional computation in BaryBind mainly comes from the Wasserstein barycenter optimization and auxiliary alignment losses. The increase in per-step time remains moderate while providing improved multimodal alignment performance.

Table 19: Computational overhead 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

B.12 Semantic Consensus in the Learned WB Space

Empirical illustration. To illustrate the semantic consensus property of the learned WB space, we perform cross-modal retrieval through the WB space. Given a query from one modality, it is first mapped into the WB space and then used to retrieve samples from other modalities. As shown in Table 20 and Table 21, the retrieved samples consistently share the same high-level semantic concepts (e.g., cooking and basketball), while preserving modality-specific variations. These results empirically demonstrate that the learned WB space captures shared semantic structures across modalities.

B.13 More Quantitative Results on Scaling to More Modalities

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

As shown in Table 22, both VAST and BaryBind benefit from the inclusion of additional modalities, demonstrating improved performance on MSR-VTT and VATEX in terms of both T2V and V2T retrieval. Notably, BaryBind consistently outperforms VAST across all modality configurations and datasets, highlighting its stronger scalability and generalization capacity.

B.14 Additional Evaluations on VideoQA and Cross-modal Generation

To further demonstrate the versatility and broader impact of BaryBind, we conduct additional experiments on Video Question Answering (VideoQA) and cross-modal generation tasks.

VideoQA. Following the data and captioner setting of VAST [9], we evaluate BaryBind on the VideoQA benchmarks MSRVTT-QA and MSVD-QA. As shown in Table 23, BaryBind achieves competitive accuracy on both benchmarks and outperforms recent global alignment methods. Notably, the superior performance on QA tasks highlights the effectiveness of the learned representations in capturing fine-grained semantics.

Table 20: Empirical illustration of semantic consensus through Text→WB→X retrieval.

Query	Retrieval Path	Modality (Retrievaled sample)	Caption
A man is cooking in the kitchen	Text→WB→Video	Video8834	A man is cooking in the kitchen; he states he will return later to add to the dish
		Video9687	A man chopping lobster and taking off the shell
		Video9038	A person is preparing some food
A man is cooking in the kitchen	Text→WB→Audio	Audio (Video7376)	Broth is being added to a soup pot and stirred with a rubber spatula
		Audio (Video8834)	A man is cooking in the kitchen; he states he will return later to add to the dish
		Audio (Video7799)	A guy chops up garlic and pours it over chicken frying in a pan
A man is cooking in the kitchen	Text→WB→Subtitle	Subtitle (Video7362)	A person is cooking on stage
		Subtitle (Video9687)	A man chopping lobster and taking off the shell
		Subtitle (Video8772)	The man is making a sauce in the kitchen

Table 21: Empirical illustration of semantic consensus through Video→WB→X retrieval.

Query	Retrieval Path	Modality (Retrieved sample)	Caption
video of a basketball event	Video→WB→Text	Text (Video7763)	People are playing basketball
		Text (Video9029)	A player is putting a basketball into the basket from distance
		Text (Video8124)	Basketball players making a shot in the last seven seconds
		Text (Video9013)	Basketball highlights of players scoring
video of a basketball event	Video→WB→Audio	Audio (Video8124)	Basketball players making a shot in the last seven seconds
		Audio (Video7763)	People are playing basketball
		Audio (Video9013)	Basketball highlights of players scoring
video of a basketball event	Video→WB→Subtitle	Subtitle (Video8124)	Basketball players making a shot in the last seven seconds
		Subtitle (Video9013)	Basketball highlights of players scoring
		Subtitle (Video8652)	A man runs into the crowd when trying to catch a basketball

Table 22: Downstream performance with increasing number of modalities.

Modalities					MSR-VTT				VATEX			
					VAST		BaryBind		VAST		BaryBind	
Text	Video	Audio	Sub.	Depth	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
✓	✓	✗	✗	✗	48.7	43.2	54.2	52.0	78.8	77.0	82.3	79.8
✓	✓	✓	✗	✗	49.3	43.7	56.0	53.2	80.0	77.3	84.2	81.3
✓	✓	✓	✓	✗	50.9	47.9	57.2	53.6	82.1	78.7	84.6	83.5
✓	✓	✓	✓	✓	51.2	49.3	57.9	54.4	82.4	79.2	84.9	83.8

Cross-modal generation. We use the VGGSound dataset to evaluate the generation performance, which consists of approximately 200K video clips annotated with 309 sound classes. We utilize the captioner provided by VAST to generate video captions, using 1024 videos for testing and the remainder for training. We adopt a single encoder–decoder configuration compatible with UnCLIP-style generators, specifically a CLIP ViT-H/14 image–text encoder paired with the ImageBind audio encoder, compatible with the Stable UnCLIP decoder. We evaluate text-to-image (T2I), audio-to-image (A2I), and text-audio-to-image (TA2I) generation performance.

Qualitative examples of the generation results are provided in the right part of Figure 1, illustrating the semantic consistency achieved by BaryBind.

B.15 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.16 Limitations and Future Work

Limitation: We note that the introduction of the iterative barycenter optimization involves a marginal increase in training computational overhead compared to simple pairwise alignment. However, as

Table 23: Evaluation on VideoQA benchmarks. Methods utilizing audio or subtitle modalities are **bolded**.

Method	MSRVTT-QA	MSVD-QA
mPLUG-2	48.0	58.1
VALOR-L	49.2	60.0
VAST	50.1	60.2
GRAM	52.3	61.5
Triangle	52.0	61.9
BaryBind	**55.4**	**63.6**

Table 24: Cross-modal generation results on VGGSound (FID ↓). T2I, A2I, and TA2I denote text-to-image, audio-to-image, and joint text-audio-to-image generation, respectively.

Method	T2I	A2I	TA2I
ImageBind	50.1	53.6	46.8
VAST	48.1	57.2	50.6
GRAM	46.7	56.5	45.2
Triangle	46.4	54.5	44.3
BaryBind	**43.6**	**50.3**	**38.6**

shown in Table 19, this cost is negligible and is well-justified by the notable gains in cross-modal retrieval accuracy.

We consider the integration of BaryBind into large multimodal language models (LLMs) as a direction for future work. Specifically, we are interested in exploring how BaryBind can enhance the multimodal understanding capabilities of these models by providing a unified, balanced token space that is not centered around language. This approach could help alleviate the issue of "hallucinations" in large multimodal models. Furthermore, by focusing on a more balanced representation across modalities, we aim to promote a deeper understanding of the physical world, moving beyond linguistic reasoning towards true multimodal perception that can more accurately capture and reason about real-world phenomena. In addition, we plan to incorporate reconstruction into the BaryBind framework to enable multimodal understanding and generation, broadening its applicability in both discriminative and generative multimodal tasks.

B.17 Impact Statement

This paper advances multimodal learning by leveraging optimal-transport–based barycenter theory to achieve more balanced multimodal representation learning. The potential societal impact includes enabling more efficient and robust approaches for general multimodal understanding. To the best of our knowledge, this work does not pose notable ethical or societal risks. Our goal is to support the responsible development of scalable and balanced multimodal learning methods and theories.

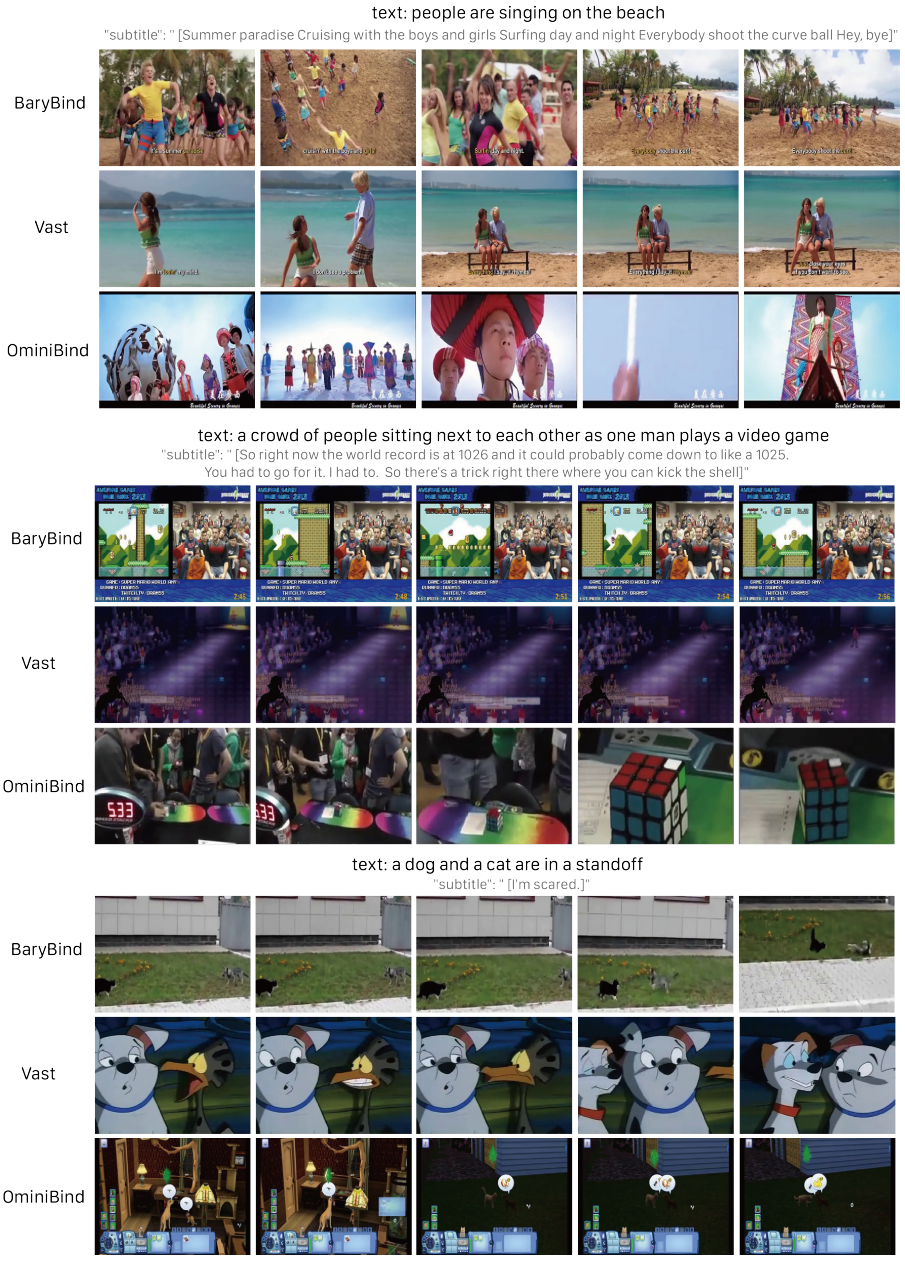


Figure 8: Visual results of text-to-video retrieval. We display 5 frames from the top-1 video.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main contributions are summarized in the introduction Section. These contributions are consistent with the results we presented in Section 3 and Section 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations and future work in Appendix B.16.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: All assumptions are stated before the theorems. Due to space constraints, we provide the complete proofs in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide all necessary details for reproduction in Section 4 and Appendix B.2.

   Guidelines:

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

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

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

Answer: [No]

Justification: While the data are publicly available, the code is not included in the supplemental material at this stage. However, we provide comprehensive implementation details in Section 4 and Appendix B.2 and algorithm in Appendix B.1 to ensure reproducibility.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We specify all training and test details in Section 4 and Appendix B.2.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: We use fixed random seeds and deterministic configurations for all experiments to ensure reproducibility. Therefore, error bars are not reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We specify the hardware (2×A100 GPUs) and the training duration in Appendix B.2. Further details on computational overhead are provided in Appendix B.11.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research, including data usage and methodology, adheres to all its guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the potential positive and negative societal impacts of our work in Appendix B.17.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We use only existing, publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All existing assets (datasets and backbone models) are properly cited in Section 4. And we use them only for academic research in compliance with their original licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: This paper does not release new datasets or code assets as part of the current submission.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method in this research does not invovle LLMs as an important, original, or non-standard component.

    Guidelines:

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