L2Dir: Integrating L_2-Norm and Directional Alignment for Unsupervised Contrastive Representation Learning in Multimodal Retrieval

Tianyu Zong[1,2], Rui Dai[2,*], Hongzhu Yi[1], Yuanxiang Wang[1], Zhenghao Zhang[1], Zhenyu Guan[3], Yujia Yang[3], Bingkang Shi[4], Yueyang Ding[2], Xiangxiang Chu[2], Kaikui Liu[2], Jungang Xu[1,*]

[1]School of Computer Science and Technology, University of Chinese Academy of Sciences
[2]Amap, Alibaba Group[†] [3]Hangzhou Institute for Advanced Study, UCAS
[4]School of Cyber Security, University of Chinese Academy of Sciences
zongtianyu20@mails.ucas.ac.cn, xujg@ucas.ac.cn
{daima.dr, chuxiangxiang.cxx, damon}@alibaba-inc.com

Abstract

Multimodal representation learning primarily relies on contrastive objectives such as InfoNCE to align diverse modalities. However, these methods focus almost exclusively on directional alignment and often neglect the intrinsic role of embedding magnitudes (L_2-norm) in the contrastive process. To bridge this gap, we propose **L2Dir**, a plug-and-play framework designed to optimize $\boldsymbol{L_2}$-norm alignment and **Dir**ectional consistency jointly. As a highly efficient solution, L2Dir doesn't require extra data, distillation, or external supervision. It can be integrated seamlessly into existing pipelines by employing a lightweight MLP to reconstruct magnitudes from backbone features. Extensive evaluations across 95 tasks using UniIR and VLM2Vec-V2 frameworks demonstrate that L2Dir yields consistent and significant performance gains over established baselines across various backbones and scales, proving that explicit magnitude modeling is a versatile and potent method for refining unsupervised multimodal representations. The source code is available at https://github.com/tianyuzong/L2Dir_ACL2026.

1 Introduction

In multimodal retrieval tasks, existing approaches are primarily categorized into two paradigms. The first one comprises Twin-Tower architectures such as CLIP (Radford et al., 2021a) and UniIR (Wei et al., 2024), which encode modalities independently using dedicated encoders like ViT (Dosovitskiy et al., 2021) and RoBERTa (Liu et al., 2019). While known for their inference efficiency, these models often struggle with fine-grained cross-modal interaction and precise semantic alignment. The second one leverages Vision-Language Models (VLMs) such as LLaVA (Liu et al., 2023a) and the Qwen-VL series (Bai et al., 2023; Wang et al., 2025b), which adopt a ViT-Projector-LLM architecture to inject visual features into Large Language Models (LLMs). By harnessing the contextual reasoning and powerful representation space of LLMs, VLM-based methods, exemplified by recent works such as VLM2Vec-V2 (Meng et al., 2025), have achieved superior modal alignment and have become an emerging trend in complex retrieval tasks.

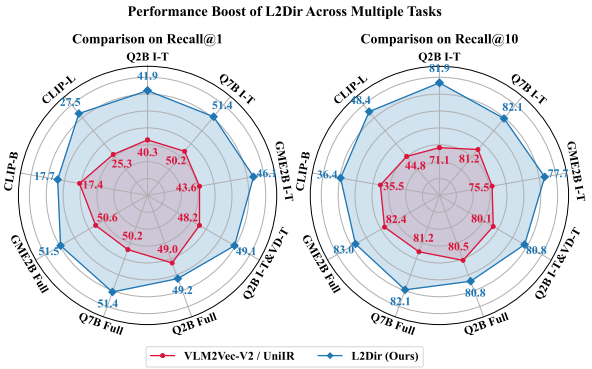


Figure 1: The radar plots report the Recall@1 (left) and Recall@10 (right) performance for L2Dir (Ours) versus the baselines. Our method achieves consistent improvements across all 9 scenarios, spanning different backbones (Qwen2-VL 2B/7B, GME 2B, CLIP-B/L), various retrieval frameworks (VLM2Vec-V2, UniIR), and multiple training data settings (Image-Text, VisDoc-T, and Full-set) for VLM2Vec-V2. Detailed in Sec. 4.

Regardless of the underlying architecture, these models primarily rely on unsupervised contrastive objectives, such as InfoNCE (van den Oord et al., 2018), to learn effective representations. This dominant paradigm focuses almost exclusively on aligning embedding directions within a unit hypersphere, while largely neglecting the magnitudes (L_2-norm) of the representation vectors. However, sole reliance on directional alignment can be limiting when semantically distinct pairs exhibit similar directions. The L_2-norm provides a complementary structural constraint that, when optimized alongside directional consistency, enforces a more rigor-

*Co-corresponding authors: Rui Dai and Jungang Xu.
†This work was conducted during the internship at Amap, Alibaba Group.

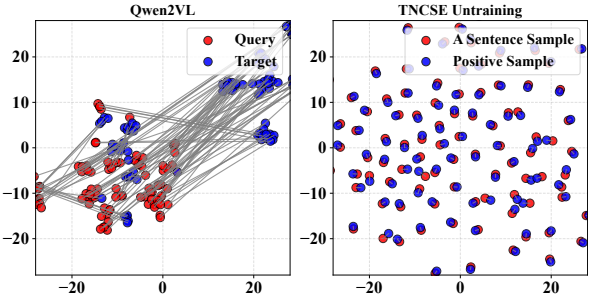


Figure 2: This figure visualizes the initial embedding distributions of Qwen2-VL-2B (MMEB-V2 (Meng et al., 2025)) and TNCSE (Wiki1M (Gao et al., 2021)) before training. We randomly select 100 samples from each dataset and project the embeddings into a 2D space employing t-SNE (Cieslak et al., 2020). Light gray lines connect corresponding positive pairs.

ous matching criterion for positive pairs. Integrating magnitude alignment as an auxiliary objective regularizes the feature space, leading to a more comprehensive alignment mechanism.

Despite its potential, semantic norm alignment in multimodal settings remains theoretically and empirically underexplored. Existing technique, such as TNCSE (Zong et al., 2025), is tailored explicitly for unimodal text scenarios. In contrast, multimodal tasks involve inherent structural disparities where visual and textual embeddings differ fundamentally in their distribution geometry. As visualized in Fig. 2, TNCSE embeddings spread evenly across the representation space. In contrast, queries and targets represented with Qwen2-VL are almost entirely segregated into two distinct groups that are far apart from each other. This structural disparity and clustered geometry render the original alignment objective less compatible, leading to persistent numerical instability and sub-optimal convergence (see Ablation Study 5.1). These findings underscore the necessity of a multimodality-aware redesign of the norm alignment mechanism to ensure stable and effective optimization.

To address these challenges, we propose L2Dir, a plug-and-play framework that jointly optimizes L_2-norm alignment and directional consistency during contrastive learning. This framework introduces a refined similarity metric that effectively couples L_2-norm constraints with the directional objectives of InfoNCE, enabling the collaborative optimization of embedding magnitude and direction within standard pipelines. By leveraging this metric, we formulate a stable contrastive loss that ensures robust multimodal training. We theoretically verify that the resulting gradient updates align with desired norm-consistency trends, effectively resolving the instability and convergence issues observed in prior methods. Doesn't require extra data or distillation, L2Dir achieves consistent retrieval gains across diverse architectures, ranging from CLIP-style towers to 7B-scale VLMs. In Fig. 1, we report the performance boost of L2Dir when integrated with mainstream backbones across various multimodal retrieval tasks. The radar chart clearly illustrates that L2Dir-enhanced models consistently outperform their base counterparts. We summarize our main contributions as follows:

- We propose L2Dir, a lightweight and plug-and-play framework that achieves synergistic optimization of L_2-norm alignment and directional consistency in unsupervised settings. This framework is efficient as it doesn't require auxiliary data, knowledge distillation, or external supervision during training.
- We introduce a novel contrastive loss for stable multimodal L_2-norm alignment. Analytically verified for its value range and gradient behavior, this loss provides a reliable objective, overcoming the convergence failures of prior methods in multimodal settings.
- Extensive evaluations across 95 diverse tasks from the VLM2Vec-V2 and UniIR frameworks demonstrate the broad efficacy of L2Dir. It consistently yields performance gains across multiple retrieval metrics and diverse architectures, including Qwen2-VL (2B/7B), GME 2B, and CLIP-B/L, proving its robustness as a general enhancement for multimodal representations.

2 Related Works

Multimodal retrieval has evolved from traditional Twin-Tower architectures to modern Vision-Language Model (VLM) paradigms. CLIP (Radford et al., 2021a) pioneers the Twin-Tower approach, utilizing InfoNCE loss to align heterogeneous modalities in a shared directional space. Subsequent frameworks, such as UniIR (Wei et al., 2024), generalize this paradigm to diverse retrieval scenarios via the M-BEIR benchmark, demonstrating the scalability of contrastive objectives. However, these methods focus exclusively on directional consistency. This approach effectively treats

embeddings as points on a unit hypersphere, leaving the L_2 norm unmodeled.Recent studies have begun to uncover the critical role of geometric properties in Transformer representations. Specifically, it has been demonstrated that the L_2 norm of embedding vectors serves as a reliable proxy for aleatoric certainty, where high-magnitude vectors correspond to clear semantic content and low magnitudes reflect inherent data ambiguity (Kirchhof et al., 2023). Despite its importance, the preservation of such geometric information faces architectural and objective-level challenges. For instance, while LayerNorm stabilizes Transformer training, it can suppress essential signals through forced numerical normalization, thereby limiting the dynamic range of features (Singhal and Kim, 2025). Furthermore, the traditional InfoNCE loss tends to force embeddings onto a unit hypersphere; this excessive pursuit of uniformity often leads to the loss of semantic density information, making it difficult for models to distinguish samples with varying confidence levels (Wang and Liu, 2021).The VLM paradigm addresses the alignment challenge by using efficient projection layers to inject visual features into LLMs. Within this landscape, GME (Zhang et al., 2024) and LamRA (Liu et al., 2025) explore unsupervised contrastive learning to refine these cross-modal bridges. More recently, VLM2Vec (Jiang et al., 2025) and VLM2Vec-V2 (Meng et al., 2025) have established new state-of-the-art benchmarks on MMEB (Jiang et al., 2025) and MMEB-V2 (Meng et al., 2025) by leveraging powerful backbones such as Phi-3.5-V (Abdin et al., 2024) and Qwen2-VL (Wang et al., 2025b). While techniques like GradCache optimize training efficiency, the underlying objective still relies solely on InfoNCE.This persistence highlights a critical gap across both paradigms, where the L_2 norm remains a vital but neglected dimension in multimodal alignment. In our work, we intervene at the architectural level to explicitly learn magnitude alignment while optimizing directional consistency. This design protects the geometric richness of the representation space and achieves more discriminative multimodal alignment, effectively complementing directional consistency to achieve a more comprehensive representation space.

3 Method

We first revisit L_2-norm constraints and analyze the monotonicity of the TNCSE-based L_{TN} loss. To address multimodal embedding gaps, we propose Tensor Norm Similarity (sim_{TN}) and a stable contrastive loss, InfoTN. Finally, we present the L2Dir framework, which enables the joint optimization of L_2-norm and embedding direction through the unified training of InfoTN and InfoNCE.

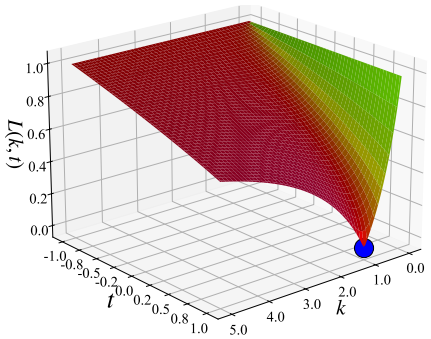


Figure 3: The gradient field of L_{TN} highlights regions dominated by the cosine similarity t (red) and the magnitude ratio k (green). The point $(1, 1)$ denotes perfect alignment, where both direction and magnitude are optimally matched.

3.1 Concise L_{TN} Definition

Existing unsupervised contrastive learning methods rely on the InfoNCE loss, which focuses solely on aligning the directions of positive samples. Conversely, TNCSE has demonstrated that explicitly focusing on the L2 Norm of the representation tensor can improve performance in semantic text similarity tasks. The tensor norm constraint objective is defined as Eq. 1:

$$L_{TN}\left(\mathbf{h}_q, \mathbf{h}_t\right) = \frac{\|\mathbf{h}_q - \mathbf{h}_t\|}{\|\mathbf{h}_q\| + \|\mathbf{h}_t\|}, \tag{1}$$

where $\mathbf{h}_q$ and $\mathbf{h}_t$ denote the representative of a query and its corresponding target in a positive pair, and γ is the angle between them. $\|\cdot\|$ represents the L_2 norm. By applying the Law of Cosines to the numerator $\|\mathbf{h}_q - \mathbf{h}_t\|$, Eq. 1 is expressed as Eq. 2:

$$L_{TN} = \frac{\sqrt{\|\mathbf{h}_q\|^2 + \|\mathbf{h}_t\|^2 - 2\,\|\mathbf{h}_q\|\,\|\mathbf{h}_t\|\cos\gamma}}{\|\mathbf{h}_q\| + \|\mathbf{h}_t\|}. \tag{2}$$

This loss is reformulated as a bivariate function $L_{TN}(k, t)$, where $k = \|\mathbf{h}_t\|/\|\mathbf{h}_q\| \in (0, \infty)$ denotes the norm ratio and $t = \cos\gamma \in [-1, 1]$ represents the cosine similarity. Since the cosine function is bounded, $t \in [-1, 1]$. Therefore, Eq. 1 can be expressed in the compact form of Eq. 3:

$$L_{TN}(k, t) = \frac{\sqrt{1 + k^2 - 2 \cdot kt}}{1 + k}. \tag{3}$$

The ideal minimum, $L_{TN}(k,t) = 0$, is achieved when both $k = 1$ perfect L2-Norm alignment and $t = 1$ perfect directional alignment. Our objective is to rigorously analyze and adapt this approach for multimodal retrieval, as its convergence properties and the existence of local optima have not been systematically investigated in prior work.

3.2 Proposed InfoTN Loss Function

We first visualize the initial t-SNE distribution in Fig. 2. The results reveal that Qwen2-VL exhibits severe modal separation, rendering L_{TN} unsuitable. Specifically, due to this substantial inter-modal disparity, directly minimizing the norm difference via L_{TN} is counterproductive. This straightforward objective fails to drive the L_2-norm ratio of positive pairs toward the expected unity ($k = 1$), instead causing persistent ratio instability during training[1]. Therefore, we shift the optimization focus from direct difference minimization to relative norm ranking within the L_2-norm space. We reconstruct L_{TN} into a similarity-based representation, as Eq. 4:

$$\text{sim}_{\text{TN}}(\mathbf{h}_q, \mathbf{h}_t) = 1 - L_{TN}. \tag{4}$$

To establish the mathematical foundation of this metric, we first determine its range. We prove that L_{TN} is strictly bounded within $[0, 1]$, as Eq. 5. The derivation is provided in Appendix B.

$$0 \leq L_{TN}(k,t) \leq 1. \tag{5}$$

This derivation directly ensures that sim_{TN} is bounded within $[0, 1]$, providing a stable numerical range for contrastive learning.

Beyond boundedness, we analyze the optimization landscape of L_{TN} to justify the effectiveness of sim_{TN}. L_{TN} is monotonically decreasing with respect to t and possesses a unique global minimum at $k = 1$, ensuring it reaches its optimum if and only if the positive pair achieves perfect alignment, and the rigorous mathematical proof and visualization are reported in Appendix A. The landscape is inherently consistent with our objective of achieving simultaneous alignment in both magnitude and direction. This well-behaved landscape identifies L_{TN} as a faithful reflection of dissimilarity; consequently, our proposed $\text{sim}_{\text{TN}}(\mathbf{h}_q, \mathbf{h}_t)$ inherits these desirable properties. It functions as a comprehensive similarity metric that is strictly and positively correlated with feature consistency, ensuring that an increase in the similarity score accurately reflects any optimization progress toward the target. This is further supported by the gradient field in Fig. 3, where the distinct regions of dominance for t and k provide complementary and consistent optimization signals across the functional space, guiding the model efficiently toward the target.

Leveraging these properties, sim_{TN} introduces a norm-based perspective to the standard contrastive framework, complementing the existing directional alignment. While conventional methods primarily rely on the InfoNCE loss to align directional semantics, we incorporate InfoTN to account for magnitude consistency explicitly. By integrating sim_{TN} into the training process, the model achieves a more comprehensive alignment of representations across both direction and magnitude. The standard InfoNCE and our proposed InfoTN loss are formulated as Eq. 6 and Eq. 7:

$$\mathcal{L}_{\text{NCE}}\left(\mathbf{h}_q, \mathbf{h}_t^+\right) = -\log \frac{e^{\text{sim}\left(\mathbf{h}_q, \mathbf{h}_t^+\right)/\tau}}{\sum_{\mathbb{D}} e^{\text{sim}\left(\mathbf{h}_q, \mathbf{h}_t^j\right)/\tau}}, \tag{6}$$

$$\mathcal{L}_{\text{InfoTN}}\left(\mathbf{h}_q, \mathbf{h}_t^+\right) = -\log \frac{e^{\text{sim}_{\text{TN}}\left(\mathbf{h}_q, \mathbf{h}_t^+\right)/\tau_{\text{TN}}}}{\sum_{\mathbb{D}} e^{\text{sim}_{\text{TN}}\left(\mathbf{h}_q, \mathbf{h}_t^j\right)/\tau_{\text{TN}}}}, \tag{7}$$

where $\mathbf{h}_q$ and $\mathbf{h}_t^+$ represent the query and positive target embeddings, while $\mathbb{D}$ denotes the set of all in-batch target samples. Temperature parameters τ and τ_{TN} are declared in Appendix C.

3.3 Proposed L2Dir Framework

To effectively implement the L2Dir framework, which jointly optimizes directional alignment $\mathcal{L}_{\text{InfoNCE}}$ and L2-Norm ranking $\mathcal{L}_{\text{InfoTN}}$, we introduce the **N**orm **A**lignment **P**rojector, **NAP**. The NAP is a simple randomly initialized FFN designed to reconstruct the L2-Norm features from the hidden states normalized by RMSNorm (Zhang and Sennrich, 2019) or LayerNorm, which are often suppressed by normalization layers in the backbone. The NA**P**'s **O**utput, **PO**, is used exclusively for the $\mathcal{L}_{\text{InfoTN}}$ loss. This design strategy ensures our method can be flexibly adopted to different encoder architectures. We detail the integration of NAP into two distinct classes of foundational models: UniIR-based CLIP models and VLM2Vec-V2-based Qwen2-VL models.

L2Dir for VLM Model. For models utilizing a ViT-Projector-LLM architecture, such as VLM2Vec-V2 based on Qwen2-VL and GME, we

[1]We report these findings in Appendix D.

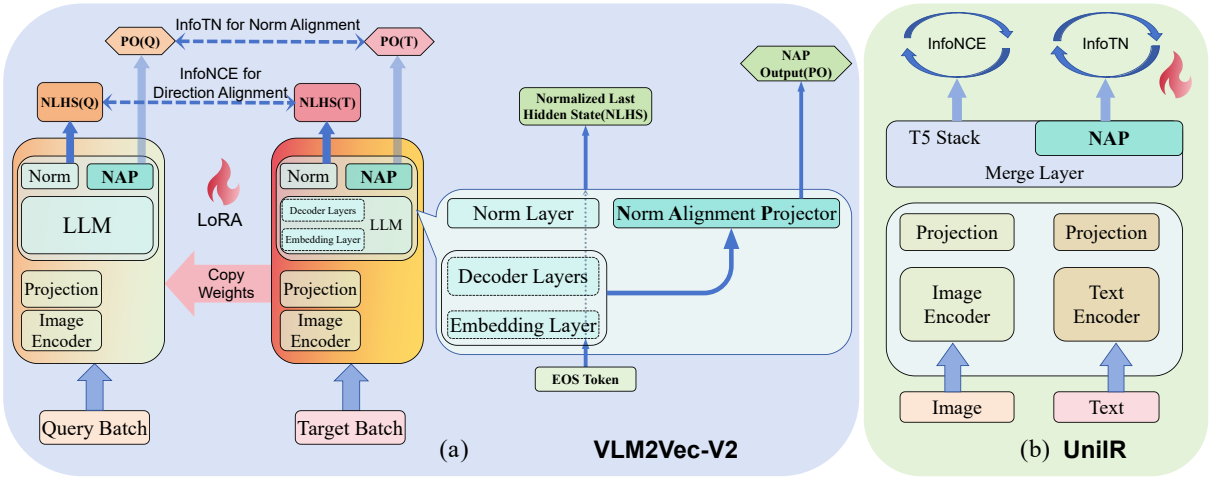


Figure 4: The overall architecture of the L2Dir framework which is instantiated on two multimodal backbones. (a) L2Dir is built upon VLM with the VLM2Vec-V2, with embeddings extracted via EOS token pooling. (b) L2Dir is built upon CLIP with the UniIR, with embeddings obtained via mean pooling.

use the LLM encoded Last Hidden States (LHS) for training. Since LHS is normalized by RMSNorm, its L2 Norm features are clipped, retaining only directional information for $\mathcal{L}_{\text{InfoNCE}}$. To reconstruct the norm features, we apply NAP. We do not use the Language Model head (LM head) for generative tasks because its excessively sparse logit distribution and large output dimension cause memory overflow even with LoRA (Hu et al., 2022) fine-tuning. NAP is an FFN placed immediately after the LLM output LHS, operating in parallel with the RMSNorm output flow. Its input and output dimensions align with the LHS dimension, preventing memory overflow. The NAP's output, PO, captures the reconstructed L2 Norm features. We employ PO in the $\mathcal{L}_{\text{InfoTN}}$ objective and LHS in the standard $\mathcal{L}_{\text{InfoNCE}}$, training them jointly

L2Dir for Twin-Tower Model. For Twin-Tower models, we adopt the UniIR framework built upon CLIP (Radford et al., 2021a). The UniIR framework adds a T5 stack (Raffel et al., 2020) as a Merge Layer after the CLIP encoders. This T5 stack output, which serves as the final merged embedding, still terminates with a LayerNorm. Consequently, the features are directional only, retaining no L2 Norm information for the $\mathcal{L}_{\text{InfoTN}}$ constraint. To capture essential norm features, NAP is applied in parallel with the final T5 LayerNorm. This strategy yields a representation with L2 Norm features. We utilize the LayerNorm output for the directional constraint $\mathcal{L}_{\text{InfoNCE}}$, while the PO is used for the $\mathcal{L}_{\text{InfoTN}}$ norm constraint. This architectural strategy enables the joint training of directional and L2 Norm features within the Twin-Tower structure.

The loss function for L2Dir. In summary, for both the VLM models and the Twin-Tower models, the L2Dir framework achieves joint optimization of directional features $\mathcal{L}_{\text{InfoNCE}}$ and L2 norm features $\mathcal{L}_{\text{InfoTN}}$ by applying NAP in parallel after the normalization layer. Since both are negative log likelihood losses, they can be directly summed. We define the overall loss function $\mathcal{L}_{\text{L2Dir}}$ as Eq. 8:

$$\begin{aligned}\mathcal{L}_{\text{L2Dir}} = \lambda \cdot \mathcal{L}_{\text{NCE}}\left(\mathbf{LHS}_q, \mathbf{LHS}_t^+\right) \\ + (1-\lambda) \cdot \mathcal{L}_{\text{InfoTN}}\left(\mathbf{PO}_q, \mathbf{PO}_t^+\right),\end{aligned} \tag{8}$$

where $\lambda \in [0, 1]$, as detailed in Section 5.

4 Experiments

4.1 Experimental Setup

To widely assess the effectiveness of the L2Dir framework, we apply the L2Dir framework to two distinct architectures: the VLM structure, based on VLM2Vec-V2, and the Twin-Tower model, based on UniIR. We use these architectures to train and evaluate multimodal retrieval tasks.

L2Dir for VLM. All experiments are conducted within the L2Dir framework applied to representative VLM architectures, utilizing the MMEB V2-train dataset derived from the VLM2Vec-V2 pipeline. Our evaluation focuses on three pivotal multimodal retrieval categories: Image-Text (I-T, 36 tasks), VisDoc-Text (VD-T, 27 tasks), and Video-Text (V-T, 18 tasks). To rigorously validate the generalization of the proposed method, we implement three progressive training settings for the Qwen2-VL-2B backbone: (i) I-T Training Only, to assess zero-shot transferability on VD-T and V-T; (ii) Joint I-T + VD-T Training, to examine zero-shot performance on V-T; and (iii) Full

Types	T-I						T-IT				IT-T			
Tasks	VisualNews		MSCOCO		Fashion200k		EDIS		WebQA		OVEN		InfoSeek	
Metrics	@1	@10	@1	@10	@10	@50	@1	@10	@1	@10	@1	@10	@1	@10
UniIR-B	4.46	16.83	30.01	71.20	3.96	13.09	17.68	44.96	**34.61**	**70.89**	9.8	32.0	**6.6**	20.9
L2Dir-B	**4.77**	**19.12**	**31.11**	**71.73**	**4.54**	**13.44**	**18.02**	**47.30**	32.90	68.58	**10.1**	**33.5**	6.5	**21.0**
UniIR-L	8.07	26.90	44.26	83.28	9.66	23.79	21.01	53.22	**44.60**	**79.73**	17.05	42.10	8.28	24.68
L2Dir-L	**12.37**	**36.20**	**46.25**	**84.24**	**11.05**	**25.13**	**21.78**	**57.11**	38.15	76.18	**21.82**	**50.07**	**11.95**	**32.82**
Types	**I-T**						**IT-I**				**IT-IT**			
Tasks	**VisualNews**		**MSCOCO**		**FashionIQ**		**Fashion200k**		**CIRR**		**OVEN**		**InfoSeek**	
Metrics	@1	@10	@1	@10	@10	@50	@10	@50	@1	@10	@1	@10	@1	@10
UniIR-B	4.32	16.96	**43.20**	81.04	4.81	14.32	11.48	26.07	**5.95**	**37.15**	24.06	51.50	10.82	33.89
L2Dir-B	**4.78**	**18.83**	43.06	**81.60**	**5.11**	**15.14**	**11.83**	**27.52**	5.18	36.93	**26.55**	**53.37**	**11.31**	**36.65**
UniIR-L	7.84	26.46	59.50	92.08	9.88	24.69	18.94	36.93	**10.26**	**50.07**	38.13	65.04	19.04	45.03
L2Dir-L	**12.50**	**36.30**	**62.52**	**92.28**	**10.86**	**27.06**	**19.67**	**38.11**	7.91	45.56	**43.95**	**71.29**	**23.77**	**54.55**

Tasks		Image-Text				Video-Text				VisDoc-Text				All
		CLS	GD	QA	RET	CLS	MRET	QA	RET	OOD	VDRv1	VDRv2	VR	
Qwen2-VL 2B	**VLM2V-2**	63.36	75.40	56.43	66.21	34.27	30.42	**31.65**	23.85	**29.15**	**18.33**	13.76	37.09	42.63
	L2Dir	**64.00**	**76.63**	**57.52**	**67.81**	**34.68**	**36.43**	31.01	**25.45**	26.65	18.00	**14.78**	**45.32**	**43.98**
Qwen2-VL7B	**VLM2V-2**	**66.23**	78.70	56.78	**70.11**	42.71	**42.52**	30.84	**32.63**	34.39	20.58	12.43	**58.59**	47.24
	L2Dir	65.95	**78.73**	**57.61**	70.02	**43.31**	41.06	**32.50**	32.48	**37.43**	**39.15**	**20.29**	57.63	**50.42**
GME 2B	**VLM2V-2**	**62.14**	73.08	**56.05**	67.83	30.20	37.20	29.94	24.30	34.40	37.49	26.36	43.56	46.67
	L2Dir	61.49	**74.65**	55.78	**69.46**	**33.53**	**37.97**	**30.38**	**24.74**	**35.76**	**49.48**	**32.24**	**48.21**	**49.57**
Qwen2-VL 2B	**VLM2V-2**	**64.07**	76.55	56.47	67.44	**38.40**	36.18	32.08	**26.41**	36.95	51.05	**36.37**	63.32	51.93
	L2Dir	63.88	**79.83**	**57.30**	**68.18**	37.18	**38.44**	**32.56**	26.03	**38.05**	**53.88**	35.55	**64.16**	**52.70**
Qwen2-VL 2B	**VLM2V-2**	64.02	75.33	**56.89**	67.57	37.34	32.89	33.17	26.32	**38.00**	**54.51**	38.59	68.36	52.86
	L2Dir	**64.39**	**77.28**	56.10	**67.86**	**40.47**	**33.94**	**33.83**	**28.62**	35.75	51.31	**39.83**	**70.12**	**53.09**
Qwen2-VL7B	**VLM2V-2**	63.48	74.85	56.23	**69.38**	40.43	**39.90**	33.73	**30.78**	40.90	56.76	32.43	73.08	53.95
	L2Dir	**64.82**	**79.50**	**57.17**	69.18	**41.01**	37.16	**35.61**	30.49	**41.95**	**58.84**	**42.90**	**74.71**	**55.80**
GME 2B	**VLM2V-2**	61.71	76.23	**57.14**	69.20	36.43	34.15	32.79	25.09	41.14	**67.49**	43.29	**66.34**	54.79
	L2Dir	**63.20**	**78.33**	57.09	**70.73**	**39.31**	**37.00**	**33.99**	**27.27**	**41.39**	67.28	**46.01**	65.49	**55.95**

Model	Q2B	Q7B	G2B	Q2B	Q2B	Q7B	G2B	Metrics	@1	@10
VLM2V2	42.6/42.2	47.2/46.7	46.7/46.2	51.9/51.4	52.7/52.3	54.0/53.4	54.8/54.2	**UR-B/L**	17.4/25.3	35.5/44.8
L2Dir	**44.0/43.6**	**50.4/49.9**	**49.6/49.0**	**52.7/52.1**	**53.1/52.6**	**55.8/55.2**	**56.0/55.4**	**LD-B/L**	**17.7/27.5**	**36.4/48.4**

Model	Q2B	Q7B	G2B	Q2B	Q2B	Q7B	G2B
VLM2V2	40.3/63.5/71.1	50.2/74.4/81.2	43.6/68.3/75.5	48.2/73.0/80.1	49.0/**73.7**/80.5	50.2/74.4/81.2	50.6/75.5/82.4
L2Dir	**41.9/64.7/81.9**	**51.4/75.2/82.1**	**46.1/70.5/77.7**	**49.1/73.4/80.8**	**49.2/73.7/80.8**	**51.4/75.2/82.1**	**51.5/76.3/83.0**

Table 1: The table consists of four sections. The first reports detailed Recall metrics for CLIP models under the UniIR and L2Dir frameworks. The second presents Hit@1 results comparing VLM models across the VLM2Vec-V2 and L2Dir frameworks. The third summarizes average performance by reporting MRR@1 and NDCG@1 across 81 tasks for the VLM2Vec-V2 comparison, as well as Recall@1 and Recall@10 for the UniIR comparison. The fourth provides the mean Recall@1/5/10 across 81 tasks specifically for the VLM2Vec-V2/L2Dir comparison. , , and denote training on I-T data, I-T + VD-T data, and the full dataset, respectively. Note that models and frameworks in the third and fourth sections are presented using abbreviations. Better performance is **Bolded**.

Dataset Training, for a comprehensive evaluation across all modalities. To further demonstrate the broad applicability of L2Dir, we extend our baselines to include larger-scale and alternative architectures, specifically Qwen2-VL-7B and GME 2B. To evaluate retrieval performance, we primarily report Hit@1, MRR@1, NDCG@1, and Recall@1, 5, 10. These metrics measure the model's accuracy in ranking relevant items within the top results.

L2Dir for Twin-Tower Model. For the Twin-Tower architecture, we apply the L2Dir framework to the UniIR structure, utilizing the M-BEIR dataset, which is explicitly associated with the UniIR framework. Our evaluation focuses on six distinct cross-modal retrieval tasks: Text-Image (T-I, 3 tasks), Image-Text (I-T, 3 tasks), Image, Text-Text (IT-T, 2 tasks), Text-Image, Text (T-IT, 2 tasks), Image, Text-Image (IT-I, 2 tasks), and Image, Text-Image, Text (IT-IT, 2 tasks). Retrieval performance is measured via Recall@5/10, with Recall@10/50 reported for specific tasks to maintain consistency with the UniIR framework's original settings. Under the feature fusion mode of the UniIR framework, we employ the ViT-B/32 and ViT-L/14 versions of CLIP as our baselines.

Baseline Implementation and Fairness. To ensure a fair comparison, all baselines are locally reproduced, with their hyperparameters strictly maintained according to the official open-source settings. This rigorous replication ensures that our proposed

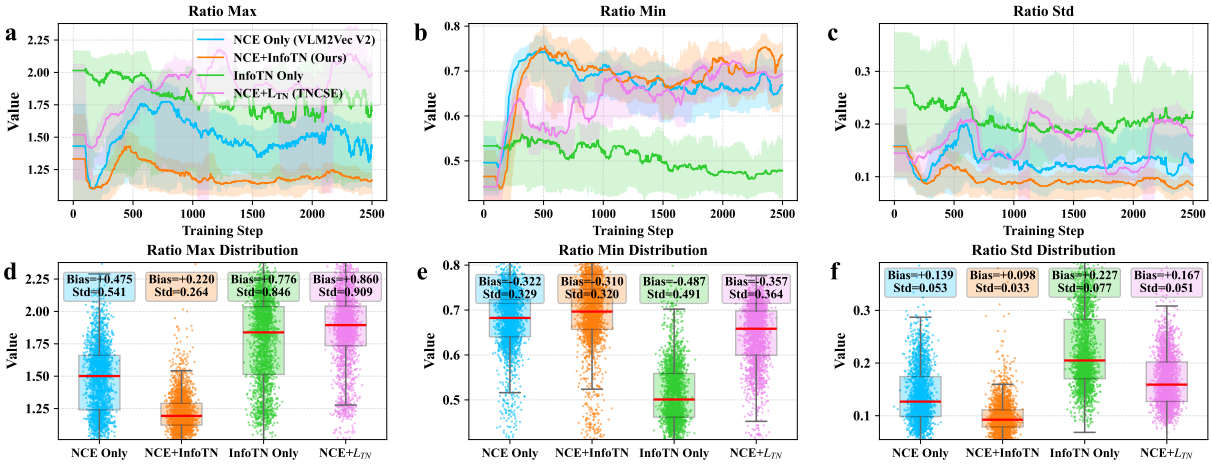


Figure 5: The ratio is defined as $\|\mathbf{h}_q\| \,/\, \left\|\mathbf{h}_t^+\right\|$ for positive pairs within a batch. (a–c) Temporal evolution of Ratio Max, Ratio Min, and Ratio Std; (d–f) Statistical distributions across experimental groups. In panels (d) and (e), Bias and Std are calculated relative to the target value of 1.0; in panel (f), they represent the standard arithmetic mean and deviation of the data. Boxplots show the interquartile range (IQR; 25th to 75th percentiles) as boxes, with the median marked by a red line. Whiskers extend to the most extreme data points within 1.5 × IQR from the quartiles.

framework is evaluated against the baselines under identical experimental environments. Our hyperparameter settings are detailed in Appendix C.

4.2 Experimental Results and Analysis

We report the comprehensive evaluation results in Table 1, which consistently validate the superiority of our proposed method across multiple benchmarks. On the M-BEIR, our approach, applied to CLIP-B and CLIP-L backbones, outperforms the baselines across the majority of the 14 multimodal retrieval tasks, demonstrating its effectiveness in traditional Twin-Encoder architectures. Our method consistently outperforms the baselines across all backbones and evaluated metrics within the extensive MMEB-V2 benchmark. Specifically, we observe superior performance in both Hits@1, MRR@1, NDCG@1, and Recall@1, 5, and 10 compared to existing methods. We consider average performance across 81 tasks spanning three meta-categories: Image-Text (36 tasks: CLS, QA, RET, GD), Video-Text (18 tasks: CLS, QA, RET, MRET), and VisDoc-Text (27 tasks: VDRv1/v2, VR, OOD). The results indicate that our approach maintains a significant lead in retrieval performance across these diverse categories, regardless of the training configuration. This consistent gain, even under data-restricted settings such as training on I-T data only, underscores the robustness of the norm alignment mechanism and its exceptional cross-modal generalization capabilities. The details of Table 1 are supplemented in Appendix E. In Appendix G, we report the number of trainable parameters introduced by the NAP network and the corresponding performance improvement.

Metric	@1	@5	@10
InfoNCE Only	63.7	84.1	88.7
InfoNCE + L_{TN}	63.8	83.5	85.6
InfoTN Only	3.3	5.8	7.2
InfoNCE+InfoTN (Ours)	**64.9**	**85.0**	**89.5**

Table 2: This table reports the average Recall@1/5/10 performance across 36 image-text retrieval tasks under four loss function settings.

5 Ablation Study

All ablation studies are conducted on the VLM2Vec-V2 framework.

5.1 Stability Analysis of Our Setting

To evaluate training dynamics, we monitor the magnitude ratio ($\|\mathbf{h}_q\| \,/\, \left\|\mathbf{h}_t^+\right\|$) for positive pairs. Ideally, this ratio should converge toward unity with minimal variance. As shown in Fig. 5(a) and 5(b), while all configurations trend toward the 1.0 equilibrium, except InfoTN Only, our InfoNCE + InfoTN setting maintains narrower $5^{th} \sim 95^{th}$ percentile intervals, demonstrating superior stability. Fig. 5(c) and 5(f) further quantify this stability, showing that InfoNCE + InfoTN yields a standard deviation near zero and the smallest bias relative to zero, confirming its superior convergence among all loss settings. Moreover, statistical analysis of the initial 2,500 steps (Fig. 5(d) and 5(e)) reveals that our setting consistently achieves the minimum Bias and Std relative to 1.0. These results confirm that InfoTN facilitates a more rapid and precise

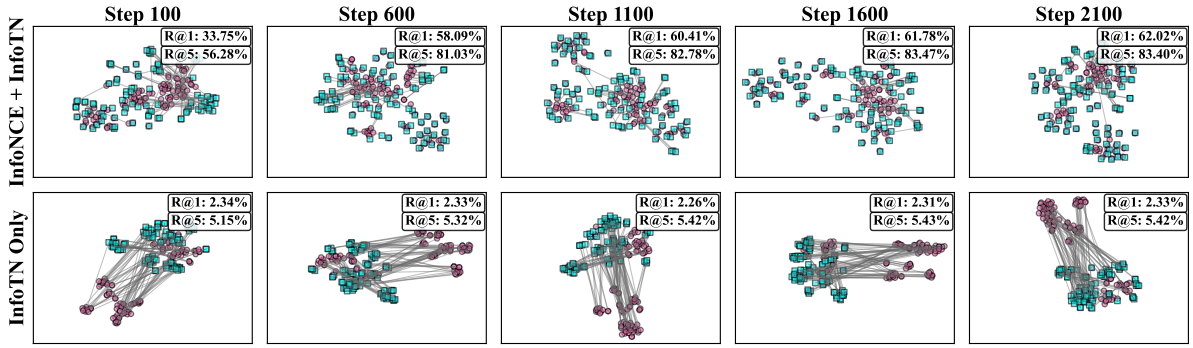


Figure 6: This figure presents t-SNE visualizations of 100 randomly selected positive sample pairs under two loss function settings. The embeddings are sampled every 500 steps, from step 100 to 2100.

alignment of representation norms, ensuring a more stable optimization process than InfoNCE only. Table 2 reports the performance on 36 I-T tasks for the checkpoints derived from the four aforementioned loss function configurations.

5.2 The Necessity of InfoNCE

While Fig. 3 visualizes InfoTN's contribution to directional alignment, we investigate if it can function independently. Theoretically, InfoNCE ensures uniform directional distribution by repelling negative samples, whereas InfoTN focuses solely on norm consistency. Without InfoNCE's repulsive force, the inherent representation collapse in Qwen2-VL is significantly exacerbated. To validate this, we compare our complete setting against an InfoTN-only baseline. t-SNE visualizations of positive pairs (Fig. 6) reveal that under InfoTN-only, embeddings remain collapsed with a persistent modal gap. In contrast, the integration of InfoNCE effectively mitigates collapse and bridges the modal separation as training progresses. Quantitatively, InfoTN-only yields negligible improvements in average Recall@1/5 across 36 I-T retrieval tasks, confirming that InfoNCE is indispensable for driving effective optimization.

5.3 Decoupling Loss and Architecture

To verify that the performance gains stem from the design of InfoTN rather than the increased parameter count from the NAP module, we compared InfoNCE + InfoTN (NAP) against InfoNCE + InfoNCE (NAP). Under identical experimental conditions, the former outperformed the latter in Hit@1/5/10 metrics across 36 I-T retrieval tasks, as reported in Table 3. These results demonstrate that InfoTN can more effectively leverages the features extracted by NAP to enhance representation quality through directional alignment optimization.

Metric	@1	@5	@10
InfoNCE Only (Baseline)	63.7	84.3	88.7
InfoNCE+InfoNCE (NAP)	63.8	84.5	88.9
InfoNCE+InfoTN (NAP)	**64.9**	**85.0**	**89.5**

Table 3: This table reports the ablation results for decoupling the additional network NAP and InfoTN.

Settings	Dir+Norm	Dir Only	Norm Only
VLM2V2 (Q)	63.7/84.1	62.6/83.8	58.9/80.1
L2Dir (Q)	**64.9/85.0**	**64.7/85.0**	**60.3/81.1**
VLM2V2 (G)	63.6/84.6	63.0/84.1	59.3/80.5
L2Dir (G)	**64.4/84.8**	**63.8/84.2**	**61.8/87.4**

Table 4: This table reports the Hit@1/5 performance of models utilizing Qwen2-VL-2B and GME2B as backbones across 36 image-text retrieval tasks, evaluated by various metrics.

5.4 Impact of NAP in Retrieval Tasks

We investigate the role of the embedding norm generated by the NAP network in retrieval, while the LHS determines the semantic direction. By comparing three similarity strategies, including direction-only, combined direction and norm, and norm-only matching, we observe that direction-only retrieval achieves the best performance, whereas norm-only matching yields the lowest scores, as reported in Table 4. This observation has a clear geometric interpretation. Under norm-only matching, embeddings are distributed onto concentric hyperspheres centered at the origin, which causes a query to be matched with any target sharing a similar radius regardless of its semantic direction. Consequently, semantically unrelated samples with close norms are erroneously retrieved, and performance degrades. Crucially, the norm produced by the NAP network is not intended to serve as a direct retrieval cue but instead functions as a vital auxiliary training signal. It regularizes the representation space and guides the model toward a more structured geometry, thereby facilitating su-

λ	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
36 I-T Avg.	63.9	64.4	**64.9**	64.5	64.2

Table 5: Performance comparison across different values of λ on the Qwen2VL-2B model. Results are reported as the average score across 36 Image-Text (I-T) tasks. The best performance is highlighted in bold.

perior directional alignment in the LHS. Instead of providing explicit matching information during inference, the NAP network acts as a structural stabilizer, enhancing the overall discriminative power of the multimodal embedding space. In Appendix F, we report the impact of training models without NAP, using hidden states without RMSNorm for magnitude alignment instead.

5.5 Hyperparameter Analysis of λ in $\mathcal{L}_{\text{L2Dir}}$

We investigate the impact of the hyperparameter λ in Eq. 8 by evaluating values in the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. These experiments are conducted using the Qwen2VL-2B model, with training restricted exclusively to the Image-Text dataset. By observing the performance on representative Image-Text tasks, we find that the optimal results are achieved at $\lambda = 0.5$. Consequently, we extend this configuration to all other task settings and frameworks within this study. Throughout this work, λ is uniformly set to 0.5. The comprehensive experimental results are detailed in Table 5.

5.6 InfoTN vs. Simple Regularization

To verify that L2Dir's gains stem from magnitude consistency rather than generic regularization, we replace InfoTN with a unit-norm geometric regularizer $\mathcal{L}_{\text{geo}} = \lambda \sum (\|\mathbf{v}_i\| - 1)^2 \left(\lambda = 10^{-3}\right)$. Evaluated on 12 representative image-text retrieval tasks (1,000 steps), geometric regularization degrades performance versus the baseline (Hit@1: 55.60 vs. 56.65), whereas L2Dir yields significant improvements (57.32). This confirms that rigid norm constraints hinder discriminative feature learning. Instead, L2Dir complements InfoNCE's angular alignment by modeling magnitude consistency, enabling richer geometric reconstruction in the joint embedding space.

5.7 Semantic Significance of Magnitudes

To investigate whether learned embedding norms encode interpretable semantic signals, we analyze statistical correlations between magnitudes and multiple linguistic metrics on 50K Wikipedia corpora, including Word Entropy, Char Entropy, Digit Count, and Stopword Ratio. Observations reveal a geometric silencing effect in standard normalized architectures, where magnitude signals exhibit weak or decoupled correlations with semantic density indicators. After introducing InfoTN, this semantic coupling is substantially restored, enabling the magnitude dimension to actively encode informational saliency and structural complexity while reducing sensitivity to redundant noise, experimental results are presented in Table 6. The mechanism achieves multiplicative recovery of information density coupling, strengthens perception of structural complexity, improves sensitivity to factual details, and significantly enhances robustness to semantic noise. Based on the above, L2Dir demonstrates that magnitude is a vital carrier of semantic saliency rather than a mere regularization term. By harmonizing magnitude and Direction during training, we ensure that directional vectors inherit a precise and semantically grounded structure. This restoration effectively bridges the gap between the rich geometric properties of early embeddings and the training stability of modern Transformers.

Metric	**Baseline**	**Ours**
Word Entropy	r = +0.069	r = **+0.216**
Char Entropy	r = +0.191	r = **+0.328**
Digit Count	r = +0.094	r = **+0.162**
Stopword Ratio	r = -0.149	r = **-0.042**

Table 6: Pearson correlation analysis between embedding magnitudes and linguistic metrics under InfoNCE baseline and InfoTN enhanced settings.

6 Conclusion

In this work, we propose InfoTN, a training objective designed to constrain the L_2-norm of positive representations in unsupervised contrastive learning, and introduce L2Dir, a training framework optimized for multi-modal retrieval tasks. Leveraging the VLM2Vec-V2 framework, we evaluate our approach using Qwen2-VL (2B/7B) and GME 2B models across 81 multi-modal retrieval tasks. The results demonstrate that L2Dir consistently outperforms baselines. Furthermore, experiments conducted on CLIP (Base/Large) models within the UniIR framework show that L2Dir achieves superior performance across 14 image-text retrieval tasks. Finally, we provide extensive ablation studies to analyze the stability and necessity of joint training with InfoTN and InfoNCE, revealing that the complementarity between these two objectives is crucial for achieving superior alignment.

Limitation

Our current work focuses on fine-tuning generative models for retrieval, which typically compromises their original generative capabilities. We do not investigate the proposed method within an autoregressive training framework in this study, primarily because substantial computational resources are required for joint optimization (Yu et al., 2025a). Consequently, exploring the synergy between contrastive and autoregressive objectives under limited hardware constraints remains a key direction for future research.

Ethical Considerations

This study adheres to rigorous ethical standards and poses no significant risks to ethics. Our research does not involve the use of personal data, sensitive information, or high-risk application scenarios. All experiments were conducted using publicly available multimodal benchmark datasets in strict accordance with their respective usage guidelines. No ethically controversial datasets or models are utilized. Furthermore, our methodology involves no processing of original data that could introduce privacy concerns or algorithmic bias.

AI Assistants Usage

In this paper, the AI Assistants are used solely for text polishing.

Acknowledgement

We gratefully acknowledge Prof. Qian Sheng[2] from the Department of Mathematics, School of Science, North China University of Technology, for his insightful guidance on the theoretical derivations. We also thank Baorong Liu[3] and Likun Zhang[4] for their assistance in analyzing the source code of the VLM2Vec-V2 framework and reproducing the baseline experiments.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4895–4901. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9448–9458.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483. IEEE.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 190–200. The Association for Computer Linguistics.

Matthew C. Cieslak, Ann M. Castelfranco, Vittoria Roncalli, Petra H. Lenz, and Daniel K. Hartline. 2020. t-distributed stochastic neighbor embedding (t-sne): A tool for eco-physiological transcriptomic analysis. *Marine Genomics*, 51:100723.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Alon Diament, Maria Gorodetski, Adam Jankelow, Ayya Keshet, Tal Shor, Daphna Weissglas-Volkov, Hagai Rossman, and Eran Segal. 2023. A multimodal dataset of 21, 412 recorded nights for sleep and respiratory research. *CoRR*, abs/2311.08979.

[2] qiansheng@ncut.edu.cn
[3] liu.baorong007@gmail.com
[4] zhanglikun20@mails.ucas.ac.cn

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5804–5813. IEEE Computer Society.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2019. Natural adversarial examples. *CoRR*, abs/1907.07174.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12031–12041. IEEE.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information*

Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. 2023. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 17085–17104. PMLR.

Hilde Kuehne, Ali Bilgin Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 780–787. IEEE Computer Society.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2556–2563. IEEE Computer Society.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *CoRR*, abs/2107.09609.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021a. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6761–6771. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. 2023b. EDIS: entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4877–4894. Association for Computational Linguistics.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 4015–4025. Computer Vision Foundation / IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021b. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2105–2114. IEEE.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6492–6505. Association for Computational Linguistics.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024b. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark V2: raising the bar for visual retrieval. *CoRR*, abs/2505.17166.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *CoRR*, abs/2507.04590.

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. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Rishi Singhal and Jung-Eun Kim. 2025. Impact of layer norm on memorization and generalization in transformers. *CoRR*, abs/2511.10566.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2495–2504. Computer Vision Foundation / IEEE.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025a. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *CoRR*, abs/2502.18017.

Weizhi Wang, Yu Tian, Linjie Yang, Heng Wang, and Xifeng Yan. 2025b. Open-qwen2vl: Compute-efficient pre-training of fully-open multimodal llms on academic resources. *CoRR*, abs/2504.00595.

Xin Wang, Jiawei Wu, Jun-Kun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24news: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6768–6775. European Language Resources Association.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11307–11317. Computer Vision Foundation / IEEE.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. 2025a. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *CoRR*, abs/2503.19900.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025b. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.

Huaying Yuan, Jian Ni, Yueze Wang, Junjie Zhou, Zhengyang Liang, Zheng Liu, Zhao Cao, Zhicheng Dou, and Ji-Rong Wen. 2025. Momentseeker: A comprehensive benchmark and A strong baseline for moment retrieval within long videos. *CoRR*, abs/2502.12558.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: improving universal multimodal retrieval by multimodal llms. *CoRR*, abs/2412.16855.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018a. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464.

Luowei Zhou, Nathan Louis, and Jason J. Corso. 2018b. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 50. BMVA Press.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

Tianyu Zong, Bingkang Shi, Hongzhu Yi, and Jungang Xu. 2025. TNCSE: tensor norm constraints for unsupervised contrastive learning of sentence embeddings. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 26192–26201. AAAI Press.

A Monotonicity Analysis of $\mathcal{L}_{TN}(k, t)$

The objective function is $\mathcal{L}_{\text{TN}}(k, t)$. The domain is defined as $k \in [0, +\infty)$ and $t \in [-1, 1]$.

A.1 Monotonicity with respect to t

The partial derivative of $\mathcal{L}_{\text{TN}}(k, t)$ with respect to t is:

$$\frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial t} = -\frac{k}{1+k} \cdot \frac{1}{\sqrt{1+k^2-2kt}}. \quad \text{(A-1)}$$

Define $f(k,t) = 1 + k^2 - 2kt$. The singular points are found by solving $f(k,t) = 0$.The discriminant of $f(k,t)$ with respect to k is $\Delta = 4(t^2-1)$.

- If $|t| < 1$, $\Delta < 0$, $\implies$ $f(k,t) > 0$ for all $k \geq 0$.
- If $t = 1$, $f(k,1) = (k-1)^2$, which implies $f(k,t) = 0$ at $(k,t) = (1,1)$.
- If $t = -1$, $f(k,-1) = (k+1)^2$, which implies $f(k,t) = 0$ at $k = -1$ (not in domain).

The only singularity in the feasible domain is at $(k,t) = (1,1)$.

For $k > 0$, the prefactor $-\frac{k}{1+k} < 0$, and $\sqrt{f(k,t)} > 0$ for $(k,t) \neq (1,1)$.

$$\frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial t} < 0, \quad \forall (k,t) \in [0,+\infty) \times [-1,1] \setminus \{(1,1)\} \quad \text{(A-2)}$$

Conclusion: $\mathcal{L}_{\text{TN}}(k,t)$ is strictly decreasing in t almost everywhere. Given $\mathcal{L}_{\text{TN}}(k,t)$ is continuous on the closed domain, its global minimum over $t \in [-1,1]$ occurs at $t = 1$ for all $k > 0$.

A.2 Monotonicity with respect to k

The partial derivative of $\mathcal{L}_{\text{TN}}(k,t)$ with respect to k is:

$$\frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial k} = \frac{(k-1)\cdot(1+t)}{(1+k)^2 \cdot \sqrt{k^2-2kt+1}} \quad \text{(A-3)}$$

The sign of the derivative is determined by $(k-1)$.

- If $k \in [0,1)$:$(k-1) < 0 \implies \frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial k} < 0$.
- If $k > 1$: $(k-1) > 0 \implies \frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial k} > 0$.
- If $k = 1$: $(k-1) = 0 \implies \frac{\partial \mathcal{L}_{\text{TN}}(k,t)}{\partial k} = 0$.

Conclusion: For any fixed $t \in [-1,1)$, $\mathcal{L}_{\text{TN}}(k,t)$ attains a local minimum at $k = 1$ along the k-direction.

Final Summary: The minimum of the L2 Norm loss function $\mathcal{L}_{\text{TN}}(k,t)$ is uniquely attained when both t and k reach their optimal values derived from the partial derivatives:

$$\min \mathcal{L}_{\text{TN}}(k,t) \text{ is achieved at } (k^*,t^*) = (1,1) \quad \text{(A-4)}$$

This implies that the $\mathcal{L}_{\text{TN}}$ objective pushes the model to align positive pairs perfectly in both direction ($t \to 1$) and L2 norm ($k \to 1$).

A.3 Visualization of L_{TN} Partial Derivatives

In this subsection, we visualize the partial derivatives of L_{TN} with respect to t and k. The corresponding top-down views of these derivatives provide an intuitive perspective on the optimization landscape, clearly reflecting the monotonicity of L_{TN} within the specific domains of t and k as illustrated in Fig. A-1.

B Strict Mathematical Derivation of $0 \leq \mathcal{L}_{\text{TN}}(\mathbf{k},\mathbf{t}) \leq 1$

B.1 Definition and Constraints

The $\mathcal{L}_{\text{TN}}$ function is defined as:

$$\mathcal{L}_{\text{TN}}(k,t) = \frac{\sqrt{1+k^2-2kt}}{1+k} \quad \text{(A-5)}$$

Constraints:

- Norm ratio: $k \geq 0$
- Cosine similarity: $t \in [-1,1]$

B.2 Proving the Lower Bound ($\mathcal{L}_{\text{TN}}(k,t) \geq 0$)

We analyze the term inside the square root, $f(k,t) = 1 + k^2 - 2kt$, using the upper constraint $t \leq 1$.

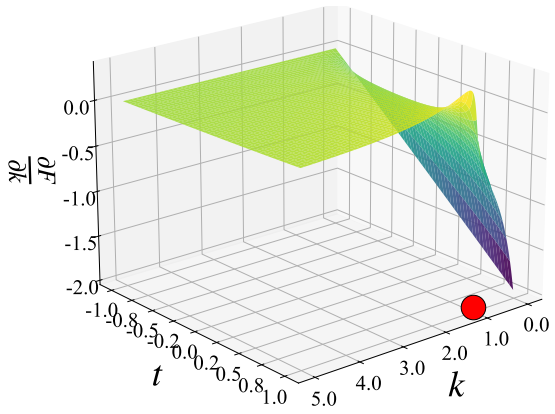
$$\begin{aligned} f(k,t) &= 1 + k^2 - 2kt \\ &\geq 1 + k^2 - 2k(1) \quad (-2k \leq 0 \text{ and } t \leq 1) \\ &= 1 - 2k + k^2 \\ &= (k-1)^2 \end{aligned} \quad \text{(A-6)}$$

Since $(k-1)^2 \geq 0$ for all real k, we have $f(k,t) \geq 0$.

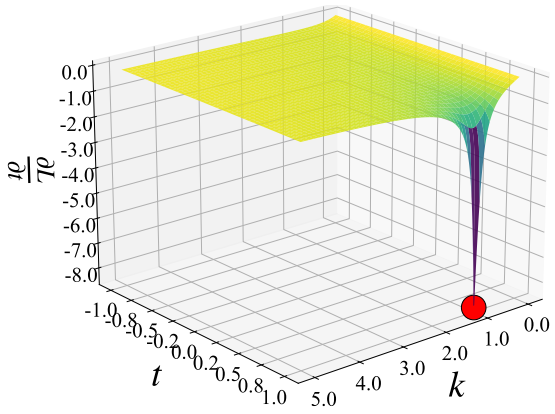
Since the numerator is the square root of a non-negative number ($\sqrt{f(k,t)} \geq 0$) and the denominator $(1+k)$ is strictly positive for $k \geq 0$:

$$\mathcal{L}_{\text{TN}}(k,t) = \frac{\sqrt{f(k,t)}}{1+k} \geq 0 \quad \text{(A-7)}$$

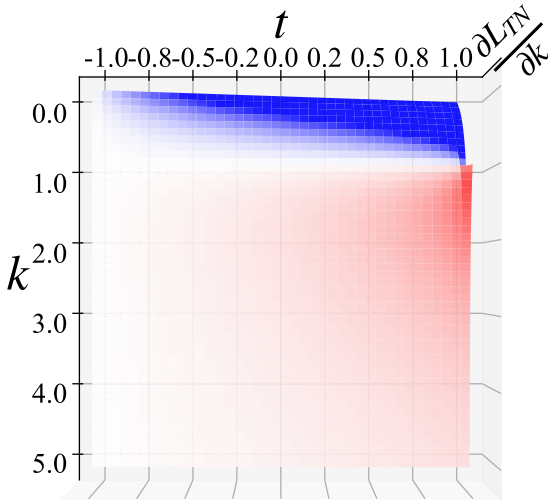
The equality $\mathcal{L}_{\text{TN}}(k,t) = 0$ holds if and only if $f(k,t) = 0$, which requires $(k-1)^2 = 0$ and $t = 1$, meaning $k = 1$ and $t = 1$.



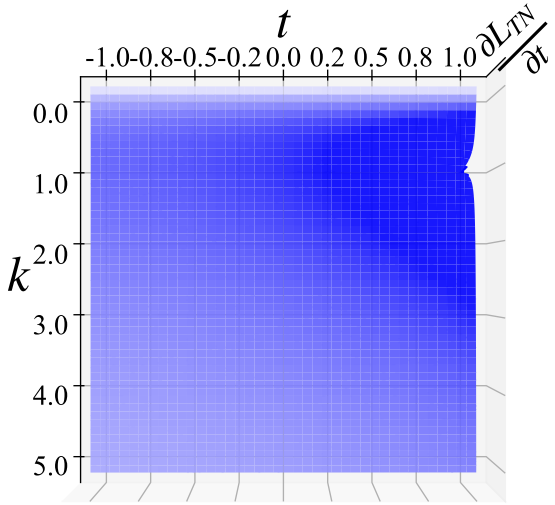
(a) Trend Chart of $\frac{\partial L_{TN}(k,t)}{\partial k}$.



(b) Trend Chart of $\frac{\partial L_{TN}(k,t)}{\partial t}$.



(c) View from above to observe the sign and magnitude of $\frac{\partial L_{TN}(k,t)}{\partial k}$.



(d) View from above to observe the sign and magnitude of $\frac{\partial L_{TN}(k,t)}{\partial t}$.

Figure A-1: Quantitative analysis of the L_{TN} loss with respect to cosine similarity (t) and norm ratio (k). (a)–(b) Visualization of the partial derivatives $\frac{\partial \mathcal{L}_{LTN}}{\partial t}$ and $\frac{\partial \mathcal{L}_{LTN}}{\partial k}$, respectively. (c)–(d) Top-down sign distributions of these gradients, where blue and red regions indicate negative and positive values, with color intensity reflecting magnitude. The point $(1, 1)$ is marked in (a) and (b) to denote perfect alignment.

B.3 Proving the Upper Bound ($\mathcal{L}_{\text{TN}}(k,t) \leq 1$)

We analyze $f(k,t)$ using the lower constraint $t \geq -1$.

$$\begin{aligned} f(k,t) &= 1 + k^2 - 2kt \\ &\leq 1 + k^2 - 2k(-1) \quad (as\ above) \\ &= 1 + k^2 + 2k \\ &= (k+1)^2 \end{aligned} \tag{A-8}$$

Taking the square root of both sides (since both are non-negative):

$$\sqrt{f(k,t)} \leq \sqrt{(k+1)^2} \tag{A-9}$$

Since $k \geq 0$, $\sqrt{(k+1)^2} = k + 1$.

$$\sqrt{1 + k^2 - 2kt} \leq k + 1 \tag{A-10}$$

Dividing both sides by the positive term $(1 + k)$:

$$\frac{\sqrt{1 + k^2 - 2kt}}{1 + k} \leq \frac{k+1}{k+1} \tag{A-11}$$

$$\mathcal{L}_{\text{TN}}(k,t) \leq 1 \tag{A-12}$$

The equality $\mathcal{L}_{\text{TN}}(k,t) = 1$ holds if and only if $t = -1$, irrespective of the value of k.

Conclusion: Combining the derivations for the lower and upper bounds, we rigorously confirm the range of the $\mathcal{L}_{\text{TN}}$ loss function

$$\mathbf{0} \leq \mathcal{L}_{\text{TN}}(\mathbf{k}, \mathbf{t}) \leq \mathbf{1}. \tag{A-13}$$

C Hyperparameter Settings

This section details the hyperparameter configurations used in our experiments. For both **UniIR**[5]

[5]https://github.com/TIGER-AI-Lab/UniIR

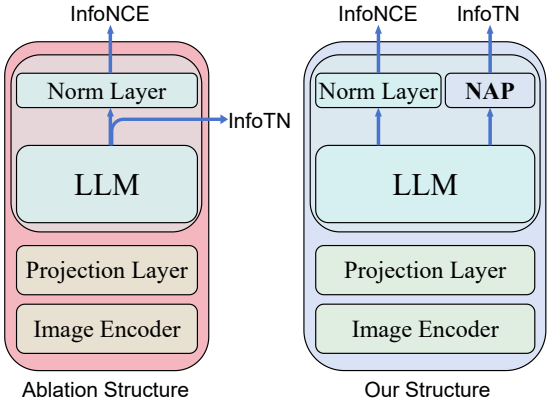


Figure F-1: This figure illustrates the difference in training architecture with and without the NAP.

and **VLM2Vec-V2**[6] frameworks, we strictly adhere to the default configurations provided in their respective open-source repositories to reproduce the baseline results faithfully.

C.1 UniIR Framework

Within the UniIR framework, all models are evaluated under the *Feature Fusion* mode. The temperature parameter for the directional InfoNCE loss is set to the default $\tau = 0.01$, our proposed norm-alignment loss utilizes $\tau_{TN} = 0.02$. Fellow default parameters, epoch set to 20.

- **CLIP-Base**: We apply a learning rate of 5.0×10^{-5} for the T5 module and 2.5×10^{-6} for the CLIP backbone.
- **CLIP-Large**: We apply a learning rate of 5.0×10^{-4} for the T5 module and 1×10^{-5} for the CLIP backbone.

C.2 VLM2Vec-V2 Framework

For models within the VLM2Vec-V2 framework, we employ Low-Rank Adaptation (LoRA) for fine-tuning ($r = 64, \alpha = 16$). We use a linear learning rate scheduler with a peak learning rate of 5×10^{-5} and 100 warmup steps. Following the default setting, the InfoNCE temperature is $\tau = 0.02$, whereas the InfoTN loss employs $\tau_{TN} = 0.01$.

- **2B-scale Models (Qwen2-VL-2B, GME 2B)**: We utilize GradCache to achieve an effective batch size of 1024. The total training steps are set to 5000, conducted on a cluster of 8× H100 GPUs.
- **7B-scale Models (Qwen2-VL-7B)**: Due to memory constraints, we utilized 8× H20 (8× 96GB) GPUs with a physical batch size of 4 per GPU. To account for gradient accumulation differences and computational capacity (not explicitly detailed in the original VLM2Vec-V2 report), we set the training duration to $20,000$ steps.

[6] https://github.com/TIGER-AI-Lab/VLM2Vec

D Supplementary Experiments on L_{TN} Instability

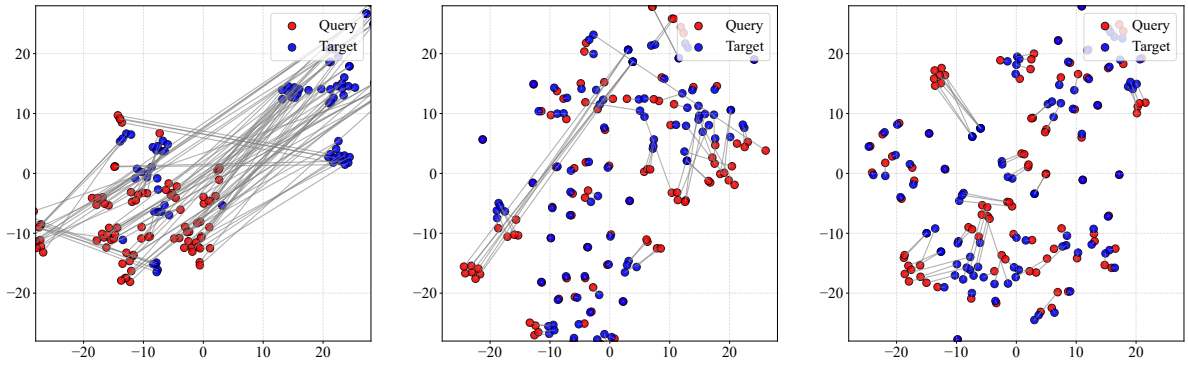
To verify that the unsuitability of L_{TN} stems from inconsistent initial norm distributions, we designed a controlled study using three distinct backbones: the initial Qwen2-VL-2B, our L2Dir-2B (pre-trained on the I-T dataset), and the publicly available VLM2Vec-V2-2B.

First, we visualized the t-SNE distributions of 100 random query-target positive pairs for each backbone, as Fig. D-1. Compared to the severe modal separation of the initial Qwen2-VL, both L2Dir-2B and VLM2Vec-V2 exhibit a more uniform representation space. We then trained all three backbones using a unified TNCSE-based loss function (InfoNCE + L_{TN}) under identical configurations. As shown in Fig. D-2, the L_2-norm ratios for L2Dir and VLM2Vec-V2 consistently converge toward the expected unity ($k = 1$), correlating with their better-conditioned initial distributions. In contrast, for the original Qwen2-VL, the convergence of the norm ratio toward $k = 1$ is far from ideal throughout the training process.

Furthermore, our analysis of the norm ratio distributions and raw data scatter plots reinforces this conclusion: backbones with a well-aligned initial state respond as expected to L_{TN}, whereas the significant modal disparity in raw VLMs hinders effective norm alignment. These results provide robust evidence that directly incorporating L_{TN} into unsupervised contrastive learning is suboptimal for a multimodal backbone without a sufficiently aligned initial distribution.

E Supplementary Results of the Main Experiment

We evaluate the performance of different VLM backbones within the VLM2Vec-V2 framework, reporting the Overall score and detailed subcategory results across 81 tasks. Specifically, we report Recall@K (K = 1, 5, 10) to quantify the retrieval effectiveness of each backbone, as shown in Table E-1. Furthermore, we report the specific performance of GME2B trained on the Image-Text (I-T)



(a) Qwen2-VL (b) VLM2Vec-V2-Qwen2-VL (c) L2Dir-Qwen2-VL

Figure D-1: This series of subfigures visualizes the embedding distributions of Qwen2-VL, VLM2Vec-V2-Qwen2-VL, and L2Dir-Qwen2-VL across 100 random sample pairs in a 2-dimensional space.

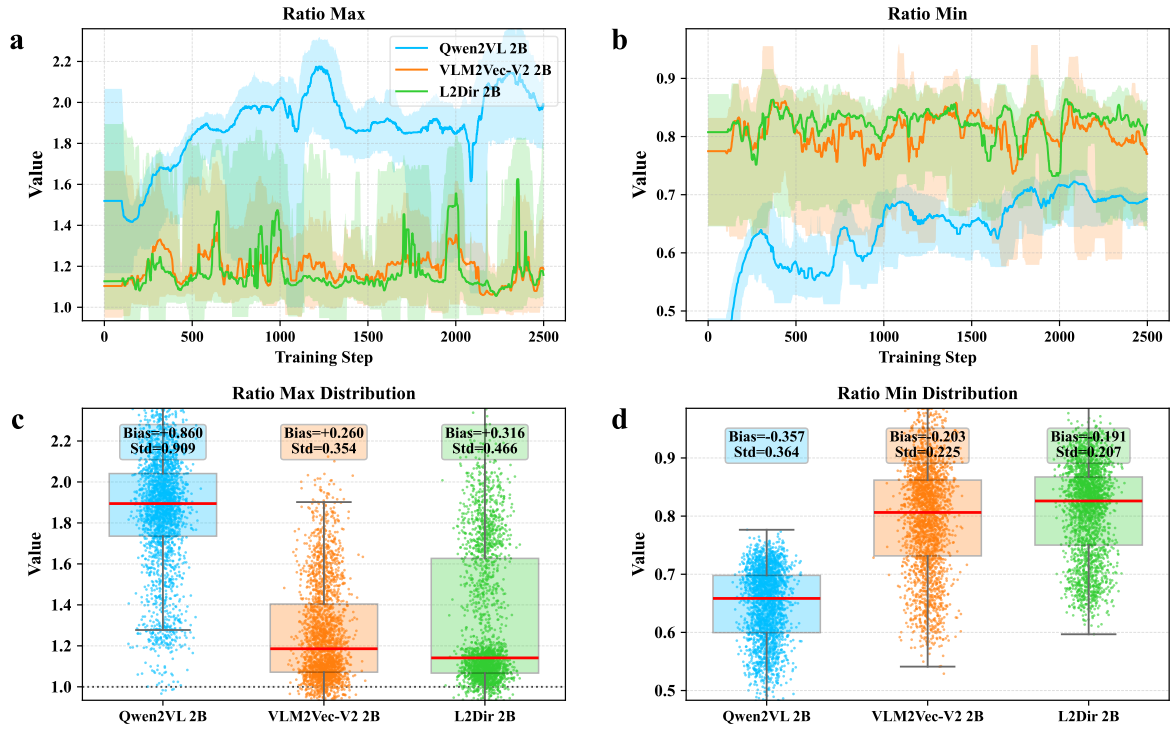


Figure D-2: A comparative study of embedding norm ratios across three distinct initialization models: Qwen2-VL 2B, VLM2Vec-V2 2B, and L2Dir 2B. Subfigures (a) and (b) illustrate the evolution of ratio-max and ratio-min during the training process, where the solid lines represent the rolling median and the shaded areas indicate the 5th to 80th percentile range. Subfigures (c) and (d) provide the statistical distributions of these ratios. The box plots and jittered scatter plots show the convergence behavior relative to the ideal target value (1.0, indicated by the horizontal dashed line). The calculated Bias (mean deviation from 1.0) and Std (root mean square deviation from 1.0) are annotated for each model.

dataset across all 81 tasks, as shown in Tables E-2, E-3, and E-4.

F Necessity of NAP

The purpose of adding a feedforward neural network **NAP** in L2Dir is to reconstruct the norm feature of the normalized last hidden state for joint training. Since the LLM output representation in

Method	36 Avg Hit@1
VLM2Vec-V2	63.7
w/o. NAP	62.4
Our Setting	**64.9**

Table F-1: This table reports ablation results about whether to add FFN.

Qwen2-VL is not normalized and inherently possesses norm features, this representation passes through a final RMSNorm for normalization. The resulting LHS is then utilized for unsupervised contrastive learning training. To obtain the norm feature, an intuitive approach is to use the hidden state input to InfoTN without RMSNorm for norm alignment, while employing LHS for InfoNCE to achieve directional alignment, as shown in Fig. F-1. However, this may be influenced by generative pretraining, introducing noise unrelated to norm alignment. Therefore, our approach involves initializing a decoupled FNN that maps raw hidden states to a new representation space, making their norm features more suitable for the InfoTN loss. We evaluate on 36 image-text retrieval tasks with consistent training hyperparameters, reporting results in Table F-1.

G NAP Overhead vs. Performance Gains

This section reports the relative increase in trainable parameters introduced by the NAP module (with respect to the original model) and compares it with the corresponding improvement in model performance, as Table G-1. The gains for VLM models are reported based on training on the image-text dataset.

Model	**CLIP-B**	**CLIP-L**	**Qwen2-VL 2B**	**Qwen2-VL 7B**	**GME 2B**
Original Params (trainable).	158.36M	438.04M	9.20M	20.51M	9.20M
Trainable Params. (NAP)	0.26M	0.59M	0.05M	0.12M	0.05M
Parameter Ratio (%)	0.17%	0.13%	0.55%	0.58%	0.55%
Recall@10 Gains (14 or 81 Avg.)	+0.90%	+3.60%	+10.80%	+0.90%	+2.20%

Table G-1: Comparison of parameter overhead and performance gains.

Recall@1														
Tasks		Image				Video				VisDoc				All
		CLS	GD	QA	RET	CLS	MRET	QA	RET	OOD	VDRv1	VDRv2	VR	
Qwen2-VL 2B	VLM2V2	63.4	75.4	56.4	66.2	34.3	29.6	**31.6**	23.9	**2.8**	**18.3**	4.1	35.6	40.4
	L2Dir	**64.0**	**76.6**	**57.5**	**67.8**	**34.7**	**35.8**	31.0	**25.5**	2.6	18.0	**5.8**	**43.9**	**41.9**
Qwen2-VL7B	VLM2V2	**66.2**	78.7	56.8	**70.1**	42.7	**41.5**	30.8	**32.6**	3.3	20.6	5.0	**56.7**	44.9
	L2Dir	66.0	**78.7**	**57.6**	70.0	**43.3**	40.2	**32.5**	32.5	**3.6**	**39.1**	**8.0**	55.8	**47.5**
GME 2B	VLM2V2	**62.1**	73.1	**56.1**	67.8	30.2	36.5	29.9	24.3	**3.7**	37.5	10.0	41.8	43.6
	L2Dir	61.5	**74.7**	55.8	**69.5**	**33.5**	**37.3**	**30.4**	**24.7**	3.4	**49.5**	**11.8**	**46.6**	**46.1**
Qwen2-VL 2B	VLM2V2	**64.1**	76.6	56.5	67.4	**38.4**	35.5	32.1	**26.4**	3.3	51.1	14.8	61.4	48.2
	L2Dir	63.9	**79.8**	**57.3**	**68.2**	37.2	**37.7**	**32.6**	26.0	**3.4**	**53.9**	**15.5**	**62.2**	**49.1**
Qwen2-VL 2B	VLM2V2	64.0	75.3	**56.9**	67.6	37.3	32.0	33.2	26.3	**3.4**	**54.5**	**16.3**	66.4	49.0
	L2Dir	**64.4**	**77.3**	56.1	**67.9**	**40.5**	**33.1**	**33.8**	**28.6**	3.2	51.3	15.3	**68.1**	**49.2**
Qwen2-VL7B	VLM2V2	63.5	74.9	56.2	**69.4**	40.4	**38.9**	33.7	**30.8**	3.8	56.7	12.2	71.0	50.2
	L2Dir	**64.8**	**79.5**	**57.2**	69.2	**41.0**	36.2	**35.6**	30.5	**4.0**	**58.8**	**15.6**	**72.7**	**51.4**
GME 2B	VLM2V2	61.7	76.2	**57.1**	69.2	36.4	33.3	32.8	25.1	3.5	**67.4**	18.1	**64.4**	50.6
	L2Dir	**63.2**	**78.3**	57.1	**70.7**	**39.3**	**36.1**	**34.0**	**27.3**	**3.5**	67.2	**18.2**	63.5	**51.5**
Recall@5														
Tasks		Image				Video				VisDoc				All
		CLS	GD	QA	RET	CLS	MRET	QA	RET	OOD	VDRv1	VDRv2	VR	
Qwen2-VL 2B	VLM2V2	83.7	**91.5**	78.5	86.8	65.0	71.5	**100.0**	44.8	**10.7**	**34.6**	**17.1**	57.1	63.5
	L2Dir	**83.9**	**91.5**	**79.7**	**88.2**	**66.2**	**75.5**	**100.0**	**47.5**	10.3	34.1	16.7	**65.0**	**64.7**
Qwen2-VL7B	VLM2V2	**84.6**	91.8	78.8	**89.5**	70.3	**77.8**	**100.0**	**56.5**	15.1	38.4	13.0	77.2	67.2
	L2Dir	84.5	**91.9**	**79.9**	88.9	**72.7**	77.7	**100.0**	55.4	**17.4**	**61.5**	**26.6**	**77.5**	**71.5**
GME 2B	VLM2V2	**83.8**	**90.8**	**78.5**	88.6	61.9	**77.8**	**100.0**	46.4	**16.0**	58.2	26.1	62.8	68.3
	L2Dir	83.3	90.4	77.7	**89.0**	**64.1**	76.9	**100.0**	**47.4**	15.9	**69.7**	**29.8**	**69.5**	**70.5**
Qwen2-VL 2B	VLM2V2	**84.3**	91.1	78.5	87.5	**68.4**	76.5	**100.0**	**49.0**	17.9	73.7	35.6	**82.4**	73.0
	L2Dir	84.0	**91.3**	**79.1**	**88.4**	66.8	**77.6**	**100.0**	48.5	**18.6**	**75.2**	**37.4**	81.7	**73.4**
Qwen2-VL 2B	VLM2V2	84.7	90.7	**80.1**	**87.8**	68.4	**75.1**	**100.0**	48.4	**18.1**	**74.0**	**37.9**	**86.4**	73.7
	L2Dir	**84.8**	**91.2**	79.0	87.6	**71.5**	73.6	**100.0**	**51.1**	17.2	73.2	37.8	86.3	**73.7**
Qwen2-VL7B	VLM2V2	**83.0**	88.5	78.0	**89.0**	**69.2**	**79.8**	**100.0**	53.2	20.5	78.7	34.1	88.5	74.4
	L2Dir	82.9	**91.1**	**79.3**	88.4	67.5	77.7	**100.0**	**54.3**	**22.0**	**79.0**	**40.8**	**89.7**	**75.2**
GME 2B	VLM2V2	82.9	91.2	**79.8**	89.0	67.5	75.7	**100.0**	46.6	**22.1**	**84.7**	43.9	**85.4**	75.5
	L2Dir	**84.2**	**92.5**	79.7	**90.4**	**70.0**	**76.6**	**100.0**	**50.8**	21.9	83.4	**46.4**	84.8	**76.3**
Recall@10														
Tasks		Image				Video				VisDoc				All
		CLS	GD	QA	RET	CLS	MRET	QA	RET	OOD	VDRv1	VDRv2	VR	
Qwen2-VL 2B	VLM2V2	88.1	**93.8**	84.5	91.0	80.5	**99.9**	**100.0**	54.8	**17.2**	**43.9**	**28.0**	63.7	71.1
	L2Dir	**88.3**	93.6	**85.8**	**92.4**	**81.6**	99.9	**100.0**	**57.7**	15.9	42.5	25.1	**73.3**	**71.9**
Qwen2-VL7B	VLM2V2	**89.3**	94.2	85.2	**93.1**	84.9	99.9	**100.0**	**65.8**	24.2	47.0	20.2	82.9	74.1
	L2Dir	89.2	**94.4**	**86.4**	92.7	**86.0**	**100.0**	**100.0**	64.8	**28.0**	**69.1**	**37.7**	**83.1**	**78.6**
GME 2B	VLM2V2	**88.1**	**93.3**	**84.6**	92.8	76.8	99.9	**100.0**	56.4	**25.1**	66.9	35.1	69.8	75.5
	L2Dir	88.0	93.2	84.0	**92.9**	**79.4**	**99.9**	**100.0**	**57.1**	24.8	**76.3**	**40.7**	**75.8**	**77.7**
Qwen2-VL 2B	VLM2V2	**89.1**	93.3	85.0	91.6	**82.4**	99.9	**100.0**	**58.4**	28.7	79.7	47.8	**87.8**	80.1
	L2Dir	88.4	**93.6**	**85.4**	**92.6**	81.2	**99.9**	**100.0**	58.4	**30.2**	**82.5**	**50.1**	87.4	**80.8**
Qwen2-VL 2B	VLM2V2	88.9	93.1	**86.0**	**91.8**	80.8	99.9	**100.0**	57.8	**27.6**	**80.3**	49.6	**91.1**	80.5
	L2Dir	**89.2**	**93.5**	85.5	91.7	**83.5**	**99.9**	**100.0**	**60.2**	26.2	80.0	**51.2**	90.2	**80.8**
Qwen2-VL7B	VLM2V2	87.5	91.5	84.9	**92.8**	**83.5**	**99.9**	**100.0**	62.4	32.1	**83.8**	45.7	92.7	81.2
	L2Dir	**87.8**	**93.4**	**85.8**	92.5	82.5	99.9	**100.0**	**63.5**	**34.4**	83.3	**52.4**	**93.6**	**82.1**
GME 2B	VLM2V2	87.4	93.4	**85.9**	93.1	81.1	99.9	**100.0**	56.5	34.1	**89.4**	55.7	**90.7**	82.4
	L2Dir	**88.7**	**94.9**	85.7	**93.6**	**82.6**	**100.0**	**100.0**	**59.9**	**34.5**	87.8	**58.7**	89.7	**83.0**

Table E-1: This table compares the proposed L2Dir method against the VLM2Vec-V2 (VLM2V2) baseline using various backbones (Qwen2-VL-2B/7B, GME 2B). Evaluation metrics include Recall@1, @5, and @10 across Image, Video, and VisDoc categories. Bold text indicates the best performance for each model-backbone pair. L2Dir demonstrates a robust performance boost across diverse architectures and data configurations. , , and denote training on I-T data, I-T + VD-T data, and the full dataset, respectively.

Modality	Group	Task	Hit@1	
			Baseline	Ours
Image-Text	CLS	**Country211** (Radford et al., 2021b)	21.40	**21.70**
		HatefulMemes (Kiela et al., 2020)	61.20	**61.30**
		ImageNet-1K (Russakovsky et al., 2015)	**82.40**	82.10
		ImageNet-A (Hendrycks et al., 2019)	**46.00**	44.80
		ImageNet-R (Hendrycks et al., 2021)	86.10	**88.00**
		N24News (Wang et al., 2022)	77.80	**78.20**
		ObjectNet (Barbu et al., 2019)	**54.40**	45.20
		Place365 (Zhou et al., 2018a)	35.30	**37.80**
		SUN397 (Xiao et al., 2010)	**70.80**	69.90
		VOC2007 (Everingham et al., 2010)	**86.00**	85.90
		10 Avg.	**62.14**	61.49
	GD	**MSCOCO** (Lin et al., 2014)	**63.90**	**63.90**
		RefCOCO (Kazemzadeh et al., 2014)	80.40	**82.40**
		RefCOCO-Matching	76.90	**82.70**
		Visual7W-Pointing (Zhu et al., 2016)	**71.10**	69.60
		4 Avg.	73.08	**74.65**
	QA	**A-OKVQA** (Schwenk et al., 2022)	**48.50**	47.00
		ChartQA (Masry et al., 2022)	**49.30**	49.10
		DocVQA (Mathew et al., 2021)	90.30	**90.90**
		GQA (Ainslie et al., 2023)	**51.50**	48.60
		InfographicsVQA (Mathew et al., 2022)	59.20	**59.70**
		OK-VQA (Marino et al., 2019)	**55.40**	53.40
		ScienceQA (Lu et al., 2022)	**39.10**	38.90
		TextVQA (Singh et al., 2019)	71.20	**72.90**
		Visual7W (Zhu et al., 2016)	53.00	**54.10**
		VizWiz (Gurari et al., 2018)	43.00	**43.20**
		10 Avg.	**56.05**	55.78
	RET	**CIRR** (Liu et al., 2021b)	53.40	**54.80**
		EDIS (Liu et al., 2023b)	80.30	**87.00**
		FashionIQ (Wu et al., 2021)	25.50	**27.70**
		MSCOCO_i2t	71.60	**73.50**
		MSCOCO_t2i	73.30	**73.70**
		NIGHTS (Diament et al., 2023)	**67.40**	66.30
		OVEN (Hu et al., 2023)	69.30	**71.20**
		VisDial (Das et al., 2017)	80.00	**81.40**
		VisualNews_i2t (Liu et al., 2021a)	76.80	**77.70**
		VisualNews_t2i (Liu et al., 2021a)	74.90	**75.00**
		WebQA (Chang et al., 2022)	88.40	**90.00**
		Wiki-SS-NQ (Ma et al., 2024a)	53.10	**55.20**
		12 Avg.	67.83	**69.46**

Table E-2: This table reports the experimental results of image-text retrieval for GME2B, trained under the VLM2Vec-V2 and L2Dir frameworks, respectively, using only the Image-Text dataset.

Modality	Group	Task	Hit@1	
			Baseline	Ours
Video-Text	CLS	**Breakfast** (Kuehne et al., 2014)	**11.78**	11.55
		HMDB51 (Kuehne et al., 2011)	34.40	**38.20**
		K700 (Carreira et al., 2019)	23.80	**34.50**
		SmthSmthV2 (Goyal et al., 2017)	28.30	**37.70**
		UCF101 (Soomro et al., 2012)	**52.70**	45.70
		4 Avg.	30.20	**33.53**
	MRET	**Charades-STA** (Gao et al., 2017)	**19.53**	19.39
		MomentSeeker (Yuan et al., 2025)	36.39	**38.01**
		QVHighlight (Lei et al., 2021)	55.68	**56.51**
		3 Avg.	37.20	**37.97**
	QA	**ActivityNetQA** (Yu et al., 2019)	**51.50**	50.90
		EgoSchema (Mangalam et al., 2023)	21.80	**24.00**
		MVBench (Li et al., 2024)	**30.33**	29.90
		NExTQA (Xiao et al., 2021)	**20.10**	19.90
		Video-MME (Fu et al., 2025)	26.00	**27.19**
		4 Avg.	29.94	**30.38**
	RET	**DiDeMo** (Hendricks et al., 2017)	**25.70**	25.00
		MSR-VTT (Xu et al., 2016)	26.40	**27.50**
		MSVD (Chen and Dolan, 2011)	**42.69**	41.49
		VATEX (Wang et al., 2019)	20.46	**22.69**
		YouCook2 (Zhou et al., 2018b)	6.26	**7.01**
		5 Avg.	24.30	**24.74**

Table E-3: This table reports the experimental results of video-text retrieval for GME2B, trained under the VLM2Vec-V2 and L2Dir frameworks, respectively, using only the Image-Text dataset.

Modality	Group	Task	Hit@1	
			Baseline	Ours
VisDoc-Text	OOD	**MMLongBench-doc** (Ma et al., 2024b)	36.04	**39.98**
		MMLongBench-page (Ma et al., 2024b)	**4.30**	**4.30**
		ViDoSeek-doc (Wang et al., 2025a)	91.68	**94.75**
		ViDoSeek-page (Wang et al., 2025a)	**5.60**	4.03
		3 Avg.	34.40	**35.76**
	VDRv1 (Faysse et al., 2025)	**ViDoRe_arxivqa**	37.40	**52.60**
		ViDoRe_docvqa	19.73	**21.06**
		ViDoRe_infovqa	49.80	**60.53**
		ViDoRe_shiftproject	23.00	**39.00**
		ViDoRe_artificial_intelligence	50.00	**65.00**
		ViDoRe_energy	52.00	**62.00**
		ViDoRe_government_reports	45.00	**56.00**
		ViDoRe_healthcare_industry	44.00	**62.00**
		ViDoRe_tabfquad	44.64	**63.21**
		ViDoRe_tatdqa	9.36	**13.43**
		10 Avg.	37.49	**49.48**
	VDRv2 (Macé et al., 2025)	**ViDoRe_biomedical_lectures_v2**	25.63	**28.75**
		ViDoRe_biomedical_lectures_v2_ML	23.59	**26.25**
		ViDoRe_economics_reports_v2	31.03	**41.38**
		ViDoRe_economics_reports_v2_ML	31.03	**34.05**
		ViDoRe_esg_reports_human_labeled_v2	25.00	**30.77**
		ViDoRe_esg_reports_v2	26.32	**35.09**
		ViDoRe_esg_reports_v2_ML	21.93	**29.39**
		7 Avg.	26.36	**32.24**
	VR (Yu et al., 2025b)	**VisRAG_ArxivQA**	39.46	**49.75**
		VisRAG_ChartQA	39.68	**53.97**
		VisRAG_InfoVQA	**65.04**	60.72
		VisRAG_MP-DocVQA	25.38	**31.64**
		VisRAG_PlotQA	23.64	**26.65**
		VisRAG_SlideVQA	**68.17**	66.55
		6 Avg.	43.56	**48.21**

Table E-4: This table reports the experimental results of visdoc-text retrieval for GME2B, trained under the VLM2Vec-V2 and L2Dir frameworks, respectively, using only the Image-Text dataset.