Multi-modal Knowledge Preserving Adapter for Embedding Backward Compatibility

Jaeseok Byun[1,2*], Gukyeong Kwon[2†], Han-Kai Hsu[2], Meher Gitika Karumuri[2], Zhikang Zhang[2], Hao Yang[2], and Davide Modolo[2]

[1]Seoul National University, South Korea
[2]Amazon AGI, USA

Abstract. Upgrading embedding models typically requires expensive database re-indexing, as new query embeddings are incompatible with existing database embeddings. While Backward Compatible Training (BCT) mitigates this by enforcing compatibility during training, existing approaches often require updating the backbone model. This is impractical because of significant training cost, the risk of performance regression, and limited access to proprietary model weights. We introduce Multi-modal Knowledge Preserving Adapter (**MKP-Adapter**), the first adapter-only BCT approach for Multi-Modal Large Language Models (MLLMs) that requires no backbone updates. We identified that the primary challenge in adapter-only BCT is preserving the knowledge of the new embeddings while enforcing backward compatibility. Hence, we propose a multi-level preservation loss that maintains the geometric structure of the embedding spaces throughout BCT. Furthermore, a focal re-weighting strategy is integrated to prioritize learning from challenging samples. Experiments demonstrate that our method achieves strong backward compatibility across diverse multi-modal benchmarks (image, text, visual document, and video retrieval tasks) and model types. Notably, MKP-Adapter is trained solely on pre-extracted embeddings and requires only negligible additional latency relative to the original backbone forward pass, highlighting its efficiency.

Keywords: Embedding compatibility · Multi-modal retrieval · Adapter

1 Introduction

Embedding models have become fundamental components in modern information retrieval systems, enabling semantic search across diverse data modalities, including text, images, audio, and video [15, 18, 20, 28, 36]. However, when a new model is trained independently of an old model for performance improvement, the embeddings produced by the two models are not inherently compatible. This incompatibility forces systems to rebuild their entire embedding database,

*This work was done during an internship at Amazon.
†Corresponding author.

a process known as "backfilling", which typically requires heavy computational overhead, especially for large-scale retrieval systems.

To address this challenge, the Backward Compatible Training (BCT) paradigm [10, 14, 29, 41, 42, 52] has been introduced. To be specific, these approaches incorporate explicit backward compatibility constraints into the training objective of the new models and allow the new embeddings to be used for querying databases constructed with old embedding models. Because compatibility constraints must be incorporated during the training stage of the new embedding model, existing BCT methods cannot be applied to pre-trained models and thus require retraining from scratch to achieve compatibility. A recent approach, XBT [14], attempts to avoid such full retraining through a two-phase procedure: it first trains an adapter on text-only data to bridge old and new embeddings and then applies LoRA-tuning [9] to the new model backbone using contrastive learning. Although XBT reports promising compatibility results on CLIP [36] under standard cross-modal retrieval settings, XBT still requires backbone updates.

BCT approaches that require backbone updates, via LoRA-tuning or full retraining, suffer from several crucial limitations, particularly in the context of MLLMs [8, 16, 17, 23, 25, 26, 30, 43, 51]. First, high computational cost. Backbone updates require repeated forward and backward passes and become prohibitively expensive for large-scale modern foundation models. Second, the risk of degrading model capability. LoRA-tuning the new model may disturb pre-trained knowledge, while training the new model from scratch with compatibility constraints can compromise the original training objectives as observed by [41]. Mitigating these risks requires carefully designed BCT strategies and high-quality large-scale training data. Third, limited applicability due to weight access requirements. Backbone updates require direct access to model weights, making such approaches inapplicable to proprietary API based models that only support black box inference.

To that end, we propose a practical alternative that achieves backward compatibility by training only an adapter (shown in Fig. 1(a)), without requiring any backbone updates. This adapter-only approach resolves aforementioned issues as follows. First, it operates solely on pre-extracted embeddings and requires no forward or backward passes through the backbone, making it efficient for training and readily applicable to MLLMs. Second, the pre-trained backbone remains intact, which allows to avoid performance regression and knowledge disturbance. Third, it requires no access to model weights, enabling direct deployment with API based models. While this adapter-only approach offers clear practical advantages over prior BCT methods, learning an adapter that achieves strong backward compatibility remains challenging. Unlike previous BCT approaches that jointly update or LoRA-tune the backbone to enforce compatibility while maintaining discriminative capability, our adapter-only setting requires the adapter alone to satisfy both objectives, with the backbone kept frozen.

A naive approach to obtaining such an adapter is to train it by directly projecting new embeddings onto old ones (*e.g.*, via $L1$ or $L2$ loss). However, forcing projected embeddings to match weaker or noisier old targets can lose

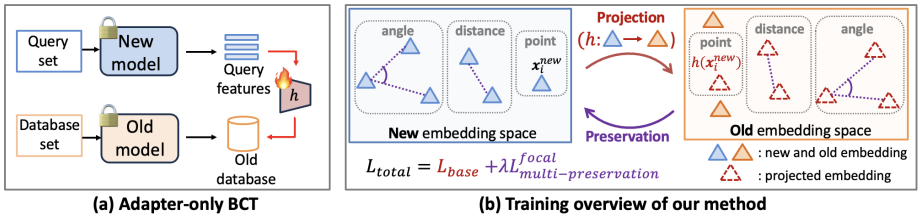


Fig. 1: Overview of MKP-Adapter.

the improved representational structure of the new model, which in turn can undermine backward compatibility performance. Our key idea is therefore to preserve the knowledge encoded in the new embeddings after projection. As illustrated in Fig. 1(b), we achieve this by combining the base projection loss (shown in red) with multi-level preservation regularization (shown in purple) at three granularities: (1) *point-wise*, a cycle-consistency loss that reconstructs each original embedding from its projected embedding; (2) *distance-wise*, a loss that preserves relative distances between embedding pairs; and (3) *angle-wise*, a loss that preserves angular relations among embedding triplets. Finally, to account for varying preservation difficulty across modalities and semantic contexts, we apply a focal re-weighting scheme that emphasizes harder samples.

We conduct a comprehensive evaluation under two practical model upgrade scenarios: (1) upgrading from dual-tower models (*e.g.*, CLIP [36]) to MLLMs (*e.g.*, GME [51]), and (2) upgrading between different versions of MLLMs. Our method achieves effective backward compatibility across diverse model upgrade scenarios on standard image-text retrieval benchmarks, including MS-COCO [22], Flickr30K [35], and Urban1K [49]. We further show that our approach generalizes across diverse tasks and modalities, including instruction-following queries in MMEB [17], visual document retrieval on ViDoRe [6], and video retrieval on MSR-VTT [46], VATEX [44], and MSVD [3]. Across all benchmarks, our method consistently and considerably outperforms baseline approaches in terms of backward compatibility. Moreover, the adapter adds only negligible overhead to the total system inference pipeline (0.01% of the original backbone forward pass cost) and can be trained within 1–2 hours on a moderate-scale (558k paired) dataset (see Sec. 5.4) on 8 A100 GPUs, demonstrating the practicality and efficiency of our approach. Overall, our contributions are summarized as follows:

1. We introduce the first adapter-only BCT framework across multiple modalities (image, text, visual documents, and video), particularly suited for using MLLMs as query encoders.
2. We propose a multi-level knowledge preservation loss to maintain the representation quality of the embeddings when training for backward compatibility. Our comprehensive evaluation validates the effectiveness of the proposed loss on diverse modalities and model upgrade scenarios.

3. MKP-Adapter can be trained within a few hours using pre-extracted embeddings and incurs only negligible overhead to the overall inference pipeline.

2 Related Work

2.1 Multi-modal embedding models

Early multi-modal embedding models [1,2,15,19,36,48] adopt dual-tower architectures trained on large-scale noisy web image-text data. Following LLaVA [25], which connects a large language model to a frozen visual backbone, various MLLM variants [23,24,34,43,53] have shown strong performance on multi-modal generation tasks. Several works have since extended MLLMs to embedding-based retrieval: E5-V [16] fine-tunes MLLMs with contrastive learning on text-only data, while GME [51] leverages large-scale image, text, and image-text composition data. VLM2Vec [17] is trained on a carefully curated dataset called the Massive Multimodal Embedding Benchmark (MMEB) training set and achieves strong performance on the corresponding MMEB evaluation benchmark. UniME [8] further improves by using knowledge distillation and hard negative mining. In this work, we consider compatibility scenarios spanning these diverse MLLM-based query encoders.

2.2 Embedding compatibility

Two representative paradigms have been proposed for embedding compatibility: forward compatible training (FCT) and backward compatible training (BCT). The pioneering FCT work [37] converts old database embeddings to be compatible with a new model via lightweight adapters, but relies on auxiliary features from a separate model, which are often unavailable in practice. FastFill [13] considers online backfilling by estimating the appropriate ordering of the backfilling sample. RBT [42] updates both query and database embeddings using a unified adapter. Recently, Embedding Converter [47] achieves strong forward compatibility for text embeddings by leveraging relative distance information. While FCT only requires training a lightweight adapter, it necessitates converting all database embeddings through the adapter, which can incur substantial computational overhead at massive scale database. Lightweight adapters have also been explored for task and domain adaptation [5, 11, 39], but these methods focus on transforming embeddings within a single model rather than maintaining compatibility across model upgrades.

In contrast, BCT modifies the training procedure of the new model to ensure compatibility with a fixed gallery of old embeddings, avoiding database updates. However, incorporating backward compatibility constraints can degrade the discriminative capability of the new model [41]. Prior works mitigate this by aligning class centers [29], expanding embedding dimensionality [52], or introducing alignment losses [10]. These methods require compatibility to be enforced during training, making them unsuitable for frozen pre-trained models and largely

limited to unimodal settings. XBT [14] extends BCT to cross-modal settings via LoRA tuning [9], enabling application to frozen models, but it still requires updating the backbone parameters. In contrast, we propose an adapter-only backward compatibility approach that requires neither backbone modification nor weight access, making it well-suited for using MLLMs as query encoders across diverse retrieval scenarios. We note our method is not tied to any specific model architecture or modality, as it operates directly on pre-extracted embeddings.

3 Problem Settings

3.1 Problem formulation

Our goal is to learn a mapping function from the new embedding space to the old embedding space for backward compatibility, defined as $h : \mathbb{R}^{d^{new}} \rightarrow \mathbb{R}^{d^{old}}$. Let f_{old} and f_{new} denote the frozen pre-trained old and new embedding models, respectively. The corresponding embeddings are given by $\mathbf{x}_i^{\text{old}} = f_{\text{old}}(s_i)$ and $\mathbf{x}_i^{\text{new}} = f_{\text{new}}(s_i)$, where s_i denotes a data sample. The dataset spans multiple modalities (images, text, visual documents, and videos). To reflect realistic deployment settings, we assume all embeddings are pre-extracted, requiring no additional forward passes through either embedding model during training. All embeddings are $L2$-normalized before any adapter operations or indexing.

3.2 Base adapter

A naive approach to training such an adapter is to learn a mapping $h : \mathbb{R}^{d^{new}} \rightarrow \mathbb{R}^{d^{old}}$ by minimizing the L2 loss (mean squared error) between the projected embedding and the old embedding, such that $h(\mathbf{x}_i^{new}) \approx \mathbf{x}_i^{old}$. Here, d^{new} and d^{old} denote the dimensionalities of the new and old embedding models, respectively. Formally, given N samples, we minimize the distance between embeddings:

$$\mathcal{L}_{base} = \sum_{i=1}^{N} \|h(\mathbf{x}_i^{new}) - \mathbf{x}_i^{old}\|_2^2. \tag{1}$$

While we use mean squared error in this formulation, alternative distance metrics such as mean absolute error may also be employed. We note that while FCT [13, 47] applies such a mapping in the opposite direction (old $\rightarrow$ new), our backward compatibility setting reverses this direction by projecting new embeddings into the old space (new $\rightarrow$ old). For multi-modal retrieval scenarios such as image–text retrieval, we form the training set by aggregating pre-extracted embeddings from different modalities into a unified dataset. Consequently, each mini-batch contains embeddings from multiple modalities, and the adapter h is trained in a modality-agnostic manner using a single shared mapping across modalities. More details and alternative adapter design choices are provided in the supplementary materials (S.M).

4 Method

4.1 Motivation

Unlike knowledge distillation [7, 33] or forward-compatible training [13, 37, 47], which use newer or stronger embeddings as the supervision target, an adapter-only BCT is inherently more challenging. Specifically, the base loss $\mathcal{L}_{\text{base}}$ in Eq. (1) encourages new embeddings to fit the old embedding space as the target, which typically exhibits a weaker and noisier semantic structure than the new one. Therefore, using old embeddings as the target naively forces the new embeddings to follow this noisy structure, which can sacrifice the discriminative capacity of the new model and potentially lead to poor backward compatibility performance. Our key insight is that, rather than simply fitting into the old space, the projection should preserve the semantic structure and knowledge encoded in the new embeddings. Building upon the base loss, we add regularization losses that explicitly preserve the representation quality of the embeddings before and after the projection. In particular, we introduce multi-level preservation losses that maintain the refined similarity structure of the new embeddings at increasing geometric complexity, along with a focal re-weighting strategy that prioritizes challenging samples for improving robustness and generalization.

4.2 Multi-level preservation loss

[Point-wise preservation loss] We first introduce a point-wise preservation loss inspired by cycle consistency [54] to ensure that individual embedding information is preserved after projection. The core idea is that if the projected embedding can be reconstructed back to the original, the essential information of the original embedding is preserved. To achieve this, we introduce an auxiliary adapter $g : \mathbb{R}^{d^{old}} \rightarrow \mathbb{R}^{d^{new}}$ that maps embeddings from the old space back to the new space. For each embedding, we project the new embedding to the old space via the main adapter $h : \mathbb{R}^{d^{new}} \rightarrow \mathbb{R}^{d^{old}}$, then reconstruct it back to the new space via g. By minimizing the reconstruction error, we enforce that h maintains sufficient semantic information needed to recover the original embedding. Formally, we train two adapters jointly with the following point-level loss:

$$\mathcal{L}_{point} = \sum_{i=1}^{N} \|g(h(\mathbf{x}_i^{new})) - \mathbf{x}_i^{new}\|_2^2. \tag{2}$$

g is used only for training and h is the learned adapter used during inference.

[Distance-wise preservation loss] While the point-wise loss preserves information at the level of individual embeddings, it does not explicitly enforce the preservation of relationships between embeddings, which are crucial for retrieval tasks where relative similarities determine ranking. Inspired by [33,47], we introduce an additional distance-wise preservation loss that maintains relative distances between embeddings, thereby encouraging the model to preserve the relative similarity structure of the original embeddings in the projected space. To

capture comprehensive pairwise mutual relationships, we follow [47] and consider both global distances computed over randomly sampled pairs within each batch, and local distances computed based on the precomputed m-nearest neighbors of each sample selected in the *new* embedding space ($m = 100$). Further details are provided in S.M. While this formulation resembles techniques used in knowledge distillation [7, 33], a key distinction is that, instead of learning from an external teacher model that typically provides a stronger supervisory signal, we use the original new embeddings themselves as self-targets to preserve their inherent relational structure during projection. We define the distance-wise structure preservation loss as:

$$\mathcal{L}_{distance} = \sum_{(i,j)\in\mathcal{P}} \left| Dist(h(\mathbf{x}_i^{new}), h(\mathbf{x}_j^{new})) - Dist(\mathbf{x}_i^{new}, \mathbf{x}_j^{new}) \right|, \tag{3}$$

where $\mathcal{P}$ denotes anchor–reference pairs (i, j), with i sampled from the mini-batch and j chosen either randomly within the batch (global) or from the m nearest neighbors of i (local). The distance function is defined as $Dist(\mathbf{a}, \mathbf{b}) = 1 - \cos(\mathbf{a}, \mathbf{b})$, where $\cos(\cdot, \cdot)$ denotes cosine similarity.

[Angle-wise preservation loss] To further preserve the relative similarity structure, we incorporate an angle-wise loss inspired by relational knowledge distillation (RKD) [33]. While distance-based losses are defined by pairs of embeddings, angle-wise losses incorporate triplets. By leveraging the higher-order structural information provided by these additional embeddings, we hypothesize that angle-wise regularization complements distance preservation in maintaining the overall manifold structure. As with the previous losses (Eqs. (2) and (3)), our formulation is self-supervised; unlike RKD, which uses a strong teacher as the target, we preserve the angles defined by the new embeddings themselves after projection. Given a triplet of embeddings, the angle formed at the vertex $\mathbf{x}_j$ is defined as:

$$\psi_A(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \cos \angle \mathbf{x}_i \mathbf{x}_j \mathbf{x}_k = \langle \mathbf{e}^{ij}, \mathbf{e}^{kj} \rangle, \tag{4}$$

where $\mathbf{e}^{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}$ and $\mathbf{e}^{kj} = \frac{\mathbf{x}_k - \mathbf{x}_j}{\|\mathbf{x}_k - \mathbf{x}_j\|_2}$ are normalized directional vectors. This measures the angle at $\mathbf{x}_j$ formed by the triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$.

The angle-wise preservation loss then maintains these angular relationships after projection:

$$\mathcal{L}_{angle} = \sum_{(i,j,k)\in\mathcal{T}} \left| \psi_A(h(\mathbf{x}_i^{new}), h(\mathbf{x}_j^{new}), h(\mathbf{x}_k^{new})) - \psi_A(\mathbf{x}_i^{new}, \mathbf{x}_j^{new}, \mathbf{x}_k^{new}) \right|. \tag{5}$$

where $\mathcal{T}$ denotes the set of embedding triplets formed using global (random) and local (nearest-neighbor) sampling.

Together, the multi-level preservation loss function is defined as:

$$\mathcal{L}_{multi-preservation} = \mathcal{L}_{point} + \mathcal{L}_{distance} + \mathcal{L}_{angle} \tag{6}$$

4.3 Focal re-weighting strategy

For more effective training, we employ a focal re-weighting strategy into the multi-level preservation losses (Eq. (6)). Preservation difficulty might vary across modalities and semantics, with some samples being easily aligned while others remain challenging. Without re-weighting, easy samples can dominate the training loss, preventing the adapter from adequately learning difficult cases. Inspired by focal loss in object detection [21], we re-weight the loss term based on errors between original and reconstructed embeddings to emphasize challenging examples. For example, the focal-weighted point-level loss is defined as:

$$\begin{aligned} &\mathcal{L}_{point}^{focal} = \frac{1}{N}\sum_{i=1}^{N} w_i \Delta_i, \\ &\text{where} \quad \Delta_i = ||g(h(\mathbf{x}_i^{new})) - \mathbf{x}_i^{new}||_2^2, \quad w_i = \left(\frac{\Delta_i}{\bar{\Delta}}\right)^{\gamma}, \quad \bar{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\Delta_i. \end{aligned} \tag{7}$$

where γ is the focusing parameter that controls the strength of re-weighting, with $\gamma = 0$ corresponding to no re-weighting and larger values placing more emphasis on harder samples. We apply the same focal re-weighting strategy to the distance-wise and angle-wise preservation losses defined in Eqs. (3) and (5); formal definitions are provided in S.M.

Finally, we obtain our model, dubbed MKP-Adapter, by optimizing the following training objective, which combines all proposed loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda \mathcal{L}_{multi-preservation}^{focal}, \tag{8}$$

where $\mathcal{L}_{\text{multi-preservation}}^{focal}$ denotes the combined multi-level preservation losses with focal re-weighting, and λ controls the trade-off between projection onto the old embedding space and preservation of the new embedding capabilities.

5 Experimental Results

We evaluate the compatibility performance of MKP-Adapter across diverse upgrade scenarios and modalities. We first demonstrate its effectiveness on standard image-to-text (I2T) and text-to-image (T2I) cross-modal retrieval tasks. Beyond retrieval, we also assess performance on complex MMEB benchmark [17], which includes a range of embedding-dependent tasks such as classification, visual grounding, and visual question answering. We further extend our evaluation to additional modalities such as visual document and video. We consider two representative model upgrade scenarios from old to new embedding models: (1) Dual-tower $\rightarrow$ MLLM, upgrading from traditional dual-tower architectures to MLLMs, where CLIP ViT-L/14 [12, 36] is used as the old model, and UniME (7B) [8] (or GME (7B) [51]) as the new model; and (2) MLLM $\rightarrow$ MLLM, upgrading between MLLM variants, where E5-V (7B) [16] is used as the old model, and UniME (7B) [8] as the new model.

5.1 Experimental settings

[Training setup] To obtain training embeddings for image and text retrieval, we use the LLaVA-LCS image–text paired dataset [25], which contains approximately 558K image–text pairs. From this dataset, we construct three types of training inputs: (1) images only, (2) texts only, and (3) image+text compositions where paired image–text samples are provided jointly as input to the MLLM. The composition inputs are included to better support instruction-following scenarios in the MMEB benchmark, where queries combine both textual and visual information. As described in Sec. 4, we aggregate all inputs into a unified training set for adapter training. During training, inputs are randomly sampled to form mini-batches across diverse modalities. Additional details and ablations, including results without composition data, are provided in S.M. We further experiment with larger datasets, SBU [32] (1M pairs) and CC3M [40] (3M pairs), to analyze the effect of training data scale on compatibility performance. For visual document retrieval, we use the ColPali training set [6], which contains approximately 118K paired samples. Each visual document is treated as a single image during both training and evaluation. For video retrieval, we use a subset of the LLaVA-Hound video dataset [50] with approximately 300K paired samples. To obtain video embeddings, we uniformly sample 8 frames from each video, extract embeddings for each frame independently, and then average them to form a single video representation, following prior work [45].

We use a 2-layer MLP as our adapter architecture with the following structure: LayerNorm-Linear-GELU-Dropout(0.1)-Linear. The default hidden dimension is set to be equal to the input embedding dimension. We train the adapter using the AdamW optimizer [27] with a weight decay of 10^{-4}, a batch size of 32, and an initial learning rate of 10^{-3}. A cosine annealing learning rate schedule is applied, decaying the learning rate from the initial value to zero, without warmup. All experiments are conducted on 8 A100 (40GB) GPUs. We use the $L2$ loss as the default base projection loss defined in Eq. (1), although our preservation regularization can be combined with alternative base losses. Both γ and λ are set to 1.0 by default. Additional training details are provided in S.M. Results under diverse hyperparameter choices are reported in Tabs. S3 and S10.

[Evaluation setup] We evaluate the backward compatibility performance of MKP-Adapter across multiple benchmarks, where query embeddings are computed as $h(\mathbf{x}^{new})$ using the adapter trained by each method, while database embeddings remain fixed as $\mathbf{x}^{old}$ from the old model. We follow the official evaluation protocols and metrics of each benchmark for comparability with prior works. First, we demonstrate its effectiveness on standard I2T and T2I retrieval using MS-COCO [22], Flickr30K [35], and Urban1K [49], reporting Recall@1. Additional Recall@10 results are provided in Tab. S14, and large-scale evaluation results on the full COCO validation set are reported in Tab. S15. Second, we conduct a comprehensive evaluation on MMEB, a more challenging benchmark comprising 36 datasets across four task categories (classification, VQA, retrieval, and visual grounding). We note this evaluation assesses the generalization ability of the adapters across diverse tasks and domains, as MMEB in-

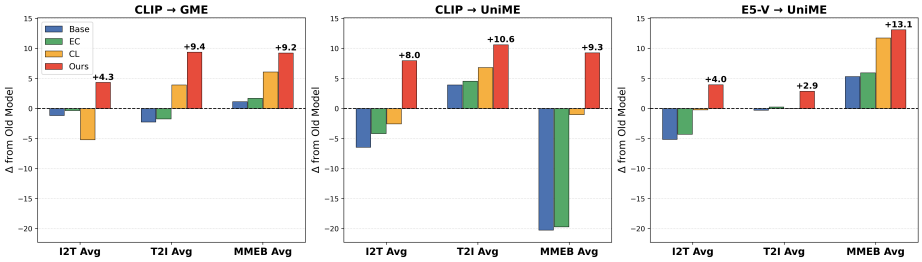


Fig. 2: Histogram of main results. The y-axis shows the performance gap relative to the old model; the black dashed line at zero indicates the old model's performance. "I2T/T2I AVG" and "MMEB AVG" correspond to the average scores computed over the benchmarks listed in Tab. 1 and Tab. 2, respectively.

cludes sub-datasets drawn from domains substantially different from the training data. Moreover, the queries contain explicit instructions, allowing us to evaluate instruction-following capability as well. Following standard protocols, we report Precision@1 as the metric. Third, we evaluate text-to-visual-document retrieval on ViDoRe (V1) [6], a benchmark of 10 sub-datasets spanning various content types (text, figures, infographics, tables) and domains (medical, business, scientific, administrative). Following prior work [6], we report nDCG@5 as the evaluation metric. Finally, we assess video retrieval performance on three benchmarks: MSR-VTT [46], MSVD [3], and VATEX [44], using Recall@1 as the metric. We report results only for CLIP→GME on document and video retrieval benchmarks, as UniME is not suitable as the new model due to its weak performance on both benchmarks (see Sec. B.9).

[Baselines] To the best of our knowledge, no prior work has explored adapter-only approaches for backward compatibility in multimodal retrieval. We therefore establish three adapter-only baselines for comparison, adapted from related literature, in which only the adapter is trained while both the old and new embedding models remain frozen.

(a) Base adapter [37]. This adapter is trained using the $L2$ loss defined in Eq. (1). We adopt $L2$ instead of $L1$, as it consistently yields superior performance in our experiments (See Tab. S5).
(b) EC-style adapter [47]. Embedding Converter (EC) [47] achieves state-of-the-art performance in forward-compatible training (FCT) for text retrieval (old → new) by distilling relative distance information between embeddings in addition to the $L2$ loss. We adopt this method to the reverse direction (new → old) to establish a backward-compatibility baseline.
(c) CL-based base adapter [14, 38]. This adapter is trained using an InfoNCE-style contrastive objective [4, 31]:

$$\mathcal{L}_{CL} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(h(\mathbf{x}_i^{new}), \mathbf{x}_i^{old})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(h(\mathbf{x}_i^{new}), \mathbf{x}_j^{old})/\tau)}, \tag{9}$$

Table 1: Compatibility results on standard I2T/T2I benchmarks (MS-COCO [22], Flickr30K [35], Urban1K [49].) $\mathbf{x}^{new}$ and $\mathbf{x}^{old}$ denote embeddings from the new and old models, respectively. h denotes the adapter function trained by each method. We report Recall@1 as the evaluation metric. **Bold** indicates the best performance among BCT methods.

Case		Method	COCO		Flickr30K		Urban1K		Average	
query	database		I2T R@1	T2I R@1	I2T R@1	T2I R@1	I2T R@1	T2I R@1	I2T R@1	T2I R@1
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: GME-7B [51]										
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	56.32	36.50	85.10	65.00	68.00	55.90	69.80	52.47
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	68.68	57.35	90.80	80.90	90.10	88.60	83.20	75.61
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	58.38	34.93	85.50	64.74	62.00	51.00	68.63	50.22
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	59.52	35.88	85.30	65.74	63.50	50.60	69.44	50.74
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	54.38	37.06	76.00	68.00	63.40	64.10	64.59	56.38
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**66.18**	**38.78**	**87.50**	**69.16**	**68.70**	**77.60**	**74.13**	**61.85**
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: UniME-7B [8]										
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	56.32	36.50	85.10	65.00	68.00	55.90	69.80	52.47
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	70.08	53.71	93.40	82.34	96.10	95.40	86.52	77.15
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	53.74	37.44	81.90	66.22	54.50	65.50	63.35	56.39
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	56.22	38.50	82.20	67.70	58.40	64.80	65.61	57.00
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	53.00	38.42	79.40	68.76	69.30	70.70	67.23	59.29
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**66.68**	**42.22**	**91.20**	**72.00**	**75.40**	**75.00**	**77.76**	**63.07**
$\mathbf{x}^{old}$: E5-V [16], $\mathbf{x}^{new}$: UniME-7B [8]										
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	62.24	50.98	88.20	80.64	83.80	85.60	78.08	72.41
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	70.08	53.71	93.40	82.34	96.10	95.40	86.52	77.15
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	55.90	50.21	85.10	79.22	77.80	86.90	72.93	72.11
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	57.30	50.32	85.90	79.52	78.10	88.10	73.77	72.65
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	63.18	50.12	88.70	78.46	81.70	88.70	77.86	72.43
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**65.28**	**52.61**	**90.60**	**81.10**	**90.20**	**92.10**	**82.03**	**75.27**

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter (set to 0.07). This loss maximizes similarity between the projected new embedding $h(\mathbf{x}_i^{new})$ and its corresponding old embedding $\mathbf{x}_i^{old}$ (positive pair), while minimizing similarity with other old embeddings (negatives). Note that our proposed regularization techniques can also be combined with this contrastive base adapter. We include these results in Tab. S6.

We omit XBT [14] from the main tables for two reasons. First, XBT requires LoRA-tuning [9] of the new embedding model, making it an unfair comparison to our adapter-only setting. Second, as detailed in Sec. C, XBT performs worse in our MLLM compatibility settings, even compared to other adapter-only baselines, likely because it disrupts the pre-trained representations of well-trained MLLMs. Additional analyses and results are provided in S.M.

5.2 Main results

Fig. 2 provides an overview of backward compatibility performance across three representative model upgrade scenarios in image-text-based benchmarks; the y-axis shows the performance gap relative to the old model, with the black dashed line at zero indicating the old model's performance. Overall, MKP-Adapter (shown in red) consistently improves over the old model and all baseline adapters

Table 2: Compatibility results on MMEB benchmarks. The numbers in parentheses indicate the number of sub-datasets in each category. We report Precision@1 as the evaluation metric. Other details are the same as in Tab. 1.

Case		Method	Per Meta-Task Score				Average
query	database		Classification (10)	VQA (10)	Retrieval (12)	Grounding (4)	Overall (36)
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: GME-7B [51]							
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	52.80	8.88	51.45	55.52	40.29
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	57.12	34.87	72.10	59.95	56.25
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	53.99	19.28	51.18	36.15	41.43
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	54.44	19.35	52.01	37.33	41.98
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	53.43	23.55	57.61	52.08	46.37
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**54.88**	**27.93**	**61.26**	**54.85**	**49.52**
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: UniME-7B [8]							
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	52.80	8.88	51.45	55.52	40.29
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	60.54	52.94	68.08	85.30	63.69
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	27.86	8.48	23.08	20.15	20.03
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	26.99	8.46	24.63	22.58	20.57
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	39.80	27.42	44.60	51.65	39.27
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**47.72**	**37.11**	**56.93**	**63.18**	**49.56**
$\mathbf{x}^{old}$: E5-V [16], $\mathbf{x}^{new}$: UniME-7B [8]							
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	45.54	12.49	47.98	54.88	38.21
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	60.54	52.94	68.08	85.30	63.69
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	42.58	29.99	51.24	56.52	43.51
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	43.23	30.91	51.72	56.83	44.14
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	48.90	36.31	57.13	65.15	49.95
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**50.58**	**37.95**	**57.76**	**67.03**	**51.30**

Table 3: Compatibility results in visual document and video benchmarks. ViDore-V1 consists of 10 sub-datasets and is evaluated using nDCG@5, while video benchmarks are evaluated using Recall@1.

Case		Method	Visual document	Video			Average
query	database		ViDoRe-V1 [6] (10) AVG	MSR-VTT [46]	MSVD [3]	VATEX [44]	
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: GME-7B [51]							
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	28.30	28.60	36.27	20.43	28.40
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	89.40	37.40	52.54	27.00	51.59
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter [37]	22.04	19.90	29.36	11.43	20.68
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	EC-style adapter [47]	24.01	20.90	28.81	12.08	21.45
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CL-based adapter [38]	38.47	30.30	41.04	19.47	32.32
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	**45.35**	**33.60**	**43.88**	**22.18**	**36.25**

on both standard I2T/T2I retrieval and the more challenging MMEB benchmark, with average (across scenarios) gains of +5.43% on I2T, +7.63% on T2I, and +10.53% on MMEB. In contrast, existing baselines often fail to preserve backward compatibility (*i.e.*, they underperform the old model itself). Specifically, on standard I2T/T2I retrieval (Fig. 2 and Tab. 1), CL-based adapters typically degrade I2T performance (*e.g.*, in the CLIP→GME and CLIP→UniME scenarios), while $L2$-based adapters (Base and EC-style) often degrade T2I performance. On the other hand, MKP-Adapter consistently improves performance on both tasks, achieving substantial gains in the CLIP→UniME scenario (+7.96%

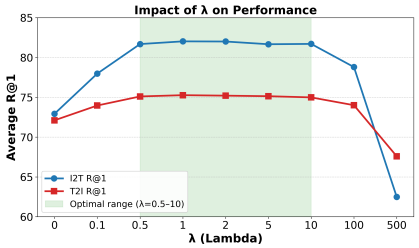
Table 4: Ablation study on proposed components.

Case query	Case database	Setting	I2T AVG	T2I AVG	MMEB AVG
$\mathbf{x}^{old}$: CLIP ViT-L/14 [12], $\mathbf{x}^{new}$: UniME-7B [8]					
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	77.76	63.07	49.56
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	w/o Point-wise preservation	77.11	63.22	47.43
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	w/o Distance-wise preservation	71.74	58.08	47.81
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	w/o Angle-wise preservation	76.83	62.26	41.55
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	w/o Focal re-weighting	77.43	62.98	46.61

Table 5: Analyses on preservation loss.

Case query	Case database	Method	Average I2T R@1	Average T2I R@1
$\mathbf{x}^{old}$: E5-V-7B [16], $\mathbf{x}^{new}$: UniME-7B [8]				
$\mathbf{x}^{old}$	$\mathbf{x}^{old}$	-	78.08	72.41
$\mathbf{x}^{new}$	$\mathbf{x}^{new}$	-	86.52	77.15
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Base adapter	72.93	72.11
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	MKP-Adapter	82.03	75.27
$h(\mathbf{x}^{new})$	$h(\mathbf{x}^{new})$	Base adapter	76.88	70.71
$h(\mathbf{x}^{new})$	$h(\mathbf{x}^{new})$	MKP-Adapter	86.30	77.11

Fig. 3: The impact of lambda. We omit the case without the base loss.



on I2T and +10.6% on T2I on average). On the MMEB benchmark (Fig. 2 and Tab. 2), CL-based adapters generally outperform $L2$-based ones, consistent with their stronger T2I performance in Tab. 1, as MMEB often involves text-heavy queries. However, they still fail to maintain backward compatibility in some scenarios (*e.g.*, CLIP→UniME). In contrast, MKP-Adapter consistently achieves the strongest performance across upgrade scenarios, with particularly large gains in the CLIP→UniME setting (averaging +9.30% over the old model), showing robust backward compatibility and out-of-domain generalization.

We further report results on text-to-visual-document and text-to-video retrieval in Tab. 3. MKP-Adapter achieves substantial gains on both tasks, demonstrating strong generalization (+17.05% on visual documents and +4.80% on videos). In this setting, $L2$-based adapters perform significantly worse than on other benchmarks; we therefore adopt a contrastive learning (CL) objective as the base adapter. Tab. S6 provides additional evidence that our method remains effective when combined with CL-based adapters, confirming the modularity and versatility of our approach. Further details are provided in S.M.

5.3 Ablation studies

We evaluate the effectiveness of our proposed components in Tab. 4, including the multi-level preservation losses (point-, distance-, and angle-wise) and the focal re-weighting strategy. All adapters are trained on LLaVA-LCS under the same setup as in Tabs. 1 and 2. We focus on the CLIP ViT-L/14 (old) →

Table 6: Analyses on adapters and training sets. Details follow Tab. 1.

Case		Variants	COCO		Flickr30K		Urban1K		Average	
query	database		I2T R@1	T2I R@1	I2T R@1	T2I R@1	I2T R@1	T2I R@1	I2T R@1	T2I R@1
$\mathbf{x}^{old}$: E5-V [16], $\mathbf{x}^{new}$: UniME-7B [8]										
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Original	65.28	52.61	90.60	81.10	90.20	92.10	82.03	75.27
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Adapter ×2	65.72	52.70	90.20	81.08	91.40	92.30	82.44	75.36
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Adapter ×4	65.92	52.91	90.80	81.46	91.10	92.20	82.61	75.52
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	Adapter ×8	66.42	52.94	90.70	81.24	91.40	92.10	82.84	75.43
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	SBU (1M pairs)	66.22	52.82	90.00	81.12	91.50	92.00	82.57	75.31
$h(\mathbf{x}^{new})$	$\mathbf{x}^{old}$	CC3M+SBU (4M pairs)	66.02	52.66	90.60	81.44	91.00	91.60	82.54	75.23

UniME (new) scenario, where the first row corresponds to the full model. Removing any component consistently degrades performance across metrics, indicating complementary contributions. In particular, removing the "distance loss" degrades standard I2T/T2I performance, whereas removing other components more strongly affects MMEB metrics. Additional results in Tab. S1 and Tab. S2 analyze the standalone impact of each component. Each preservation loss yields positive gains individually and complementary improvements when combined (Tab. S1). Focal re-weighting consistently improves performance across different loss combinations (Tab. S2), demonstrating its robustness.

5.4 Analyses

[Impact of preservation loss] We validate our core hypothesis: whether our method preserves the representational quality of the new embeddings. To this end, we evaluate retrieval when both the query and database use the projected embeddings $h(\mathbf{x}^{new})$ (rows 5-6), instead of using the original old embeddings in the database as in the standard BCT setting (rows 3-4). If the similarity structure of the new embeddings is preserved after projection, the resulting performance should remain close to that of the new model and considerably exceed that of the old model. As shown in the last row of Tab. 5, the projected embeddings from ours (row 6) achieve performance within 0.2 points of the new embedding model (row 2), indicating that MKP-Adapter preserves retrieval quality in the projected space and thus enables effective backward compatibility (row 4).

Next, we analyze the trade-off between projection and preservation by varying λ. When $\lambda = \infty$, the base projection loss is removed, leading to near-zero performance, as expected. Performance peaks around $\lambda \in [0.5, 10]$, while overly large values ($\lambda > 100$) degrade results by over-emphasizing preservation. Our method remains stable across a broad moderate range, indicating robustness to λ. Sensitivity to other hyperparameters, including the relative weights among preservation losses and the focal parameter γ, is reported in Tab. S10.

[Adapter sizes and dataset scales] Larger adapters yield slightly better results, but the overall trends remain consistent. Increasing the training data from 558k pairs (LLaVA-LCS) to 4M pairs (CC3M+SBU) provides only marginal gains, suggesting that our adapter design and a moderate-scale dataset are already sufficient to achieve strong and robust performance across benchmarks.

[Computational overhead] We first measure the inference cost of our adapter on a single A100 (40GB) with a batch size of 32. In the CLIP ViT-L/14 [12] to UniME (7B) [8] setting, our adapter adds only 0.25 ms of latency per batch (0.01% latency to the original backbone forward pass). Training is also efficient, requiring less than 10 minutes per epoch on a 1M-sample dataset, whereas LoRA-tuning a 7B model typically takes over a day per epoch even on H100 GPUs. Although we use 50 epochs by default, 10 epochs already achieve comparable performance (see S.M.), enabling complete training in 1–2 hours on 8 A100 GPUs. In contrast, backfilling even a moderate 5M-scale database takes about one day.

6 Conclusion

We propose a practical adapter-only approach for achieving backward compatibility with MLLMs. We devise multi-level preservation losses with focal reweighting to retain the representational quality of new embeddings after projection. Experiments across diverse modalities demonstrate strong backward compatibility with negligible latency and efficient training.

References

1. Byun, J., Hwang, T., Fu, J., Moon, T.: Grit-vlp: Grouped mini-batch sampling for efficient vision and language pre-training. In: ECCV (2022)
2. Byun, J., Kim, D., Moon, T.: Converting and smoothing false negatives for vision-language pre-training. arXiv preprint arXiv:2312.06112 (2023)
3. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: ACL (2011)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
5. Dong, Q., Dai, D., Song, Y., Xu, J., Sui, Z., Li, L.: Calibrating factual knowledge in pretrained language models. In: EMNLP (2022)
6. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449 (2024)
7. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. IJCV (2021)
8. Gu, T., Yang, K., Feng, Z., Wang, X., Zhang, Y., Long, D., Chen, Y., Cai, W., Deng, J.: Breaking the modality barrier: Universal embedding learning with multimodal llms. In: Proceedings of the 33rd ACM International Conference on Multimedia (2025)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
10. Hu, W., Bansal, R., Cao, K., Rao, N., Subbian, K., Leskovec, J.: Learning backward compatible embeddings. In: SIGKDD (2022)
11. Huang, C., Feng, D., Lei, W., Lv, J.: Dreditor: An time-efficient approach for building a domain-specific dense retrieval model. arXiv preprint arXiv:2401.12540 (2024)

12. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (july 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773
13. Jaeckle, F., Faghri, F., Farhadi, A., Tuzel, O., Pouransari, H.: Fastfill: Efficient compatible model update. arXiv preprint arXiv:2303.04766 (2023)
14. Jang, Y.K., Lim, S.n.: Towards cross-modal backward-compatible representation learning for vision-language models. In: ICCV (2025)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
16. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
17. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In: ICLR (2025)
18. Koepke, A.S., Oncescu, A.M., Henriques, J.F., Akata, Z., Albanie, S.: Audio retrieval with natural language queries: A benchmark study. IEEE Transactions on Multimedia (2022)
19. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems (2021)
20. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281 (2023)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
24. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NIPS (2023)
26. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: CVPR (2025)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
28. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
29. Meng, Q., Zhang, C., Xu, X., Zhou, F.: Learning compatible embeddings. In: ICCV (2021)
30. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., et al.: Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590 (2025)
31. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
32. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: NIPS (2011)
33. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)

34. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
35. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
37. Ramanujan, V., Vasu, P.K.A., Farhadi, A., Tuzel, O., Pouransari, H.: Forward compatible training for large-scale embedding retrieval systems. In: CVPR (2022)
38. Seo, S., Uzunbas, M.G., Han, B., Cao, S., Lim, S.N.: Metric compatible training for online backfilling in large-scale retrieval. In: WACV (2025)
39. Shah, S., Reddy, S., Bhattacharyya, P.: A retrofitting model for incorporating semantic relations into word embeddings. In: COLING (2020)
40. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
41. Shen, Y., Xiong, Y., Xia, W., Soatto, S.: Towards backward-compatible representation learning. In: CVPR (2020)
42. Wang, C.Y., Chang, Y.L., Yang, S.T., Chen, D., Lai, S.H.: Unified representation learning for cross model compatibility. arXiv preprint arXiv:2008.04821 (2020)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
44. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: ICCV (2019)
45. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
46. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016)
47. Yoon, J., Arik, S.O.: Embedding-converter: A unified framework for cross-model embedding transformation. In: ACL (2025)
48. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
49. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: ECCV (2024)
50. Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A.G., Bisk, Y., et al.: Direct preference optimization of video large multimodal models from language model reward. In: NAACL (2025)
51. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. In: CVPR (2025)
52. Zhou, Y., Li, Z., Shrivastava, A., Zhao, H., Torralba, A., Tian, T., Lim, S.N.: Bt^ 2: Backward-compatible training with basis transformation. In: ICCV (2023)
53. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)