

MuCo: Multi-turn Contrastive Learning for Multimodal Embedding Model

Geonmo Gu^{1,3} Byeongho Heo¹ Jaemyung Yu¹ Jaehui Hwang¹ Taekyung Kim¹
 Sangmin Lee³ HeeJae Jun² Yoohoon Kang^{‡,2} Sangdoo Yun^{†,1} Dongyoon Han^{†,1}

¹NAVER AI Lab

²NAVER AI Search Platform

³Korea University

Abstract

*Universal Multimodal embedding models built on Multimodal Large Language Models (MLLMs) have traditionally employed contrastive learning, which aligns representations of query-target pairs across different modalities. Yet, despite its empirical success, they are primarily built on a “single-turn” formulation where each query-target pair is treated as an independent data point. This paradigm leads to computational inefficiency when scaling, as it requires a separate forward pass for each pair and overlooks potential contextual relationships between multiple queries that can relate to the same context. In this work, we introduce Multi-Turn Contrastive Learning (**MuCo**), a dialogue-inspired framework that revisits this process. **MuCo** leverages the conversational nature of MLLMs to process multiple, related query-target pairs associated with a single image within a single forward pass. This allows us to extract a set of multiple query and target embeddings simultaneously, conditioned on a shared context representation, amplifying the effective batch size and overall training efficiency. Experiments exhibit **MuCo** with a newly curated 5M multimodal multi-turn dataset (M3T), which yields state-of-the-art retrieval performance on MMEB and M-BEIR benchmarks, while markedly enhancing both training efficiency and representation coherence across modalities. Code and M3T are available at <https://github.com/naver-ai/muco>*

1. Introduction

Universal multimodal embeddings aim to encode diverse modalities and tasks into a unified vector space, enabling a single model to generalize across a wide range of applications without task-specific adaptation. As real-world scenarios increasingly demand flexible and scalable multimodal understanding, Multimodal Large Language Models (MLLMs) have emerged as a powerful foundation for learn-

ing such universal representations. By jointly modeling visual and textual inputs, MLLM-based embedding frameworks have demonstrated remarkable performance across classification, Visual Question Answering (VQA), and multimodal retrieval tasks [5, 6, 10, 14, 16, 17, 23, 25, 35].

Despite this progress, most existing methods [5, 6, 10, 14, 16, 21–23, 25, 28, 35] still rely on “**single-turn**” **contrastive learning** [27], which processes one query-target pair per image and relies on a large batch size to enhance their embedding quality. We argue that this paradigm has two critical limitations. First, by treating each pair as an isolated data point, it overlooks the rich contextual interdependence among the potential multiple queries that can be derived from the same image. While a semantically rich image can support diverse text queries (e.g., object attributes, spatial relationships) [8], processing them independently may miss the opportunity to build a contextually coherent representation. Second, scaling up contrastive learning by enlarging batch size introduces substantial computational overhead. This is because a large batch simply means processing more images. Unlike text, each image requires processing by a visual encoder and generates a substantially larger number of tokens. This inefficiency hinders the scalability of universal multimodal embedding models that depend on large-batch contrastive learning.

We introduce **Multi-turn Contrastive Learning (MuCo)** – a novel method that extends traditional single-turn contrastive learning (Fig. 1a) into a multi-turn setting (Fig. 1b). Instead of processing isolated pairs, **MuCo** processes multiple query-target pairs for a single image in a dialog format. By leveraging the causal and conversational capability of MLLMs, **MuCo** models the sequential query–target interactions across turns. This enables the model to progressively refine its embeddings through subsequent turns, yielding context-rich representations that generalize effectively across diverse tasks. Furthermore, our method is efficient: the computationally heavy visual input is processed only once, while subsequent lightweight text-only turns are used to extract multiple distinct embeddings. This strategy dramatically increases the effective

[‡]Provided general support for the project. [†]Corresponding authors.

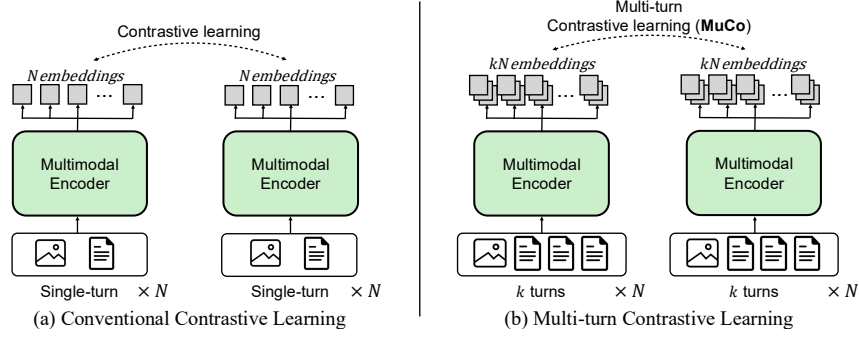


Figure 1. **Comparison of conventional single-turn vs. our multi-turn contrastive learning.** (a) Conventional contrastive learning employs a single query-target pair per image, using negative targets from other images to learn discriminative representations. (b) Our multi-turn contrastive learning (**MuCo**) generalizes this paradigm by using *multiple* query-target pairs per image, with expanded negative targets corresponding to *expanded* targets (from other images), which enables the model to learn more discriminative embeddings. Notably, for the same number of encoder forward passes, **MuCo** provides k -times larger effective batch size than conventional contrastive learning.

batch size with minimal additional training cost.

Our framework is strengthened through successive pre-training and fine-tuning stages. For pretraining, we construct a **5M**-scale multimodal multi-turn dataset featuring diverse multi-turn conversational pairs that explicitly model various tasks from a single image, synthesized via MLLMs [2] and LLMs [1]. Pretraining on the M3T dataset allows **MuCo** to benefit from both a large effective batch size and richer contextual signals from sequential turns, leading to a substantially more powerful and discriminative shared embedding space. Subsequently, in the fine-tuning stage, we adopt the multi-turn paradigm to standard single-turn datasets. This is achieved by simulating multi-turn reasoning via a prompt-based in-context reconstruction task (akin to masked modeling [9]), which further enhances the embedding’s discriminative power.

Extensive experiments demonstrate that **MuCo** substantially outperforms existing MLLM-based embedding models across multiple benchmarks [26, 34], achieving state-of-the-art performance with improved efficiency and generalization. **MuCo** achieved new SOTA on the MMEB benchmark [17], improving zero-shot performance (Precision@1) by +3.0%p and fine-tuning performance by +1.6%p over the previous SOTA [6, 31]. On the M-BEIR benchmark [34], **MuCo** also secures SOTA in both 2B and 7B, improving overall Recall by +1.6%p and +1.7%p, respectively. Notably, **MuCo** demonstrates remarkable efficiency: while increasing the effective batch size 7 times ($1024 \rightarrow 7168$), the required FLOPs increase by a mere $\leq 3\%$ (from 17.5 to 18.0 PFLOPs). This is substantially more efficient than standard contrastive learning, which requires 122.7 PFLOPs for the same batch size expansion, and achieves +0.7%p higher performance.

2. Related Work

MLLM-based universal multimodal embedding model. Recent studies on universal multimodal embedding can

be grouped into two main paradigms: CLIP-based methods [21, 22, 28, 36] introduced architectural or objective variations to improve visual-text alignment. In contrast, MLLM-based approaches leveraged the strong multimodal understanding and instruction-following capabilities of large multimodal language models (MLLMs) to produce more semantically aligned and instruction-aware embeddings. Within this paradigm, one line of work focused on scaling and curating data for embedding learning, such as VLM2Vec [17], MMRet [38], mmE5 [6]. A complementary line of research improved the training procedure itself. E5-V [16] unified representations through MLLM distillation, while UniME [14] and LLaVE [18] improved supervision through refined negative sampling. B3 [31] enhanced scalability via optimized batch construction, and M3Task-UEM [30] introduced task-adaptive representation learning to improve embedding effectiveness across diverse tasks. While these models advance the field, they predominantly rely on single-turn supervision.

Training data for the universal multimodal embeddings. Advances in data construction have strongly shaped universal multimodal embedding. Early web-crawled datasets like DataComp [11], Conceptual Captions [3, 29], and YFCC [32] enabled large-scale alignment but suffered from noise and lacked semantic richness. The next generation [13, 37, 38] introduced synthetic captioning for existing images [37, 38] or synthesized images [13], improving linguistic diversity and grounding quality. Later efforts focused on large-scale aggregation. MMEB benchmark [17] unified 36 datasets across four meta-tasks, and M-BEIR [34] consists of 16 retrieval datasets across 8 retrieval tasks. These two benchmarks are widely used to evaluate universal multimodal embeddings. More recently, mmE5 [6] moved beyond aggregation by automatically generating aligned supervision, addressing data insufficiency, and extending to multilingual domains. Despite these advances, most data still relies on single query-target pairs. Recent MLLM-based embedding models focus on one-shot

descriptions or isolated instructions, leaving their multi-turn conversational nature underexploited. To address this gap, our work explores how multi-turn, dialogue-inspired supervision can enhance contextual alignment and instruction grounding in universal multimodal embedding learning.

3. Preliminary

Multimodal Large Language Models (MLLMs) [2, 19, 20, 24, 33] are generative models that extend beyond text-only inputs, processing a combination of images and text to produce sequential text tokens. Employing MLLMs as the backbone for multimodal embedding models is highly advantageous, as this enables the creation of unified embeddings from free-form multimodal samples. However, a fundamental challenge arises from the inherent asymmetry between modalities. The multimodal embedding models still rely on a visual encoder that converts each image into a dense sequence of tokens. For instance, on the MMEB benchmark [17], an image with the average resolution (509×456 , 294 tokens) requires 2.24 TFLOPs to process, while an average-length text (25 tokens) requires only 0.12 TFLOPs¹. This token length imbalance makes the visual input far more computationally costly to process, creating a significant bottleneck for training efficiency.

Contrastive learning. Previous methods [5, 6, 14, 17, 23, 31, 38] for training multimodal embedding models typically rely on a contrastive learning objective, InfoNCE loss [27], applied to single query-target pairs, where query and target consist of multimodal data such as images and texts:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(q_i, p_i) \in \mathcal{B}} -\log \frac{\phi(q_i, p_i)}{\sum_{p \in \mathcal{N} \cup \{p_i\}} \phi(q_i, p)}, \quad (1)$$

where q_i is a query and p_i is its positive target. The similarity score is given by $\phi(q_i, p_i) = \exp(f(q_i)^\top f(p_i)/\tau)$, where $f(\cdot)$ denotes an encoder producing ℓ_2 -normalized embeddings and τ is the temperature. $\mathcal{B}$ denotes the mini-batch, and $\mathcal{N}$ represents the set of valid negative targets within the batch. While using Eq. (1) has become the de facto standard for contrastive learning, its effectiveness depends heavily on the batch size [4, 7, 12, 28, 36], as larger batches provide more negative samples, which further leads to severe scalability issues in multimodal training.

4. Method

This section introduces our proposed method **Multi-turn Contrastive learning (MuCo)**. **MuCo** capitalizes on the inherent causal processing capability of MLLMs, extending it to sequential multi-turn input processing such that each embedding (for each query) reflects the contextual dependencies of all prior turns. In effect, the model is compelled to

internalize richer and coherent entangled embeddings that extend beyond conventional single-query alignment.

We implement **MuCo** through a two-stage training process. In the first pretraining stage (§4.1), the model is trained on our newly constructed Multi-Modal Multi-Turn (**M3T**) dataset. This large-scale, dialogue-oriented corpus, containing multiple query-target pairs per image, enables the model to learn contextually coherent representations from conversational data. Subsequently, the fine-tuning stage extends the proposed **MuCo** framework on standard single-turn-based benchmarks, ensuring its effectiveness under conventional evaluation protocols (§4.2).

4.1. Multi-turn Contrastive Learning (MuCo)

Our method reframes the embedding extraction process as a multi-turn dialogue. By placing a special embedding token (`<|emb|>`) in the response of the `assistant` within a prompt structure, as shown in Fig. 2, our framework can extract multiple distinct embeddings from a single input sequence at once. Subsequent turns per image consist of appending *text-only* queries or positive targets.

We first instantiate **MuCo** in pretraining by compounding supervisory signals across sequential turns. Our goal is to learn context-rich embeddings for earlier turns – particularly for the initial turn² – that generalize effectively across tasks such as classification, retrieval, and visual question answering. We believe this could be achieved by training on multiple query-target pairs, processed sequentially via our dialogue template (Fig. 2); this design ensures that earlier embeddings become effective through training subsequent turns, as they are continually referenced by later queries (e.g., the n -th embedding is formed after attending to the previous $n - 1$ queries).

Let the set of training data be denoted as $\mathcal{D} = \{((\mathcal{I}_i, q_i^j), p_i^j)\}$ where $\mathcal{I}_i$ is an i -th image and q_i^j, p_i^j represent j -th query text and positive target text. For a mini-batch $\mathcal{B} \subset \mathcal{D}$, employing the traditional contrastive learning loss Eq. (1) gives

$$\frac{1}{|\mathcal{B}|} \sum_{((\mathcal{I}_i, q_i^j), p_i^j) \in \mathcal{B}} -\log \frac{\phi((\mathcal{I}_i, q_i^j), p_i^j)}{\sum_{p \in \mathcal{N} \cup \{p_i^j\}} \phi((\mathcal{I}_i, q_i^j), p)}. \quad (2)$$

We believe this naive approach has two drawbacks. First, it often fails to leverage the full context of multiple query-target pairs for a given single image, as trained exclusively in a single-query setting, despite the fact that these pairs are not mutually exclusive (i.e., crucial cues may lie in other questions and answers). Second, when increasing the number of images by enlarging the batch size, it usually fails to mitigate the computational overhead incurred by the image

¹Based on Qwen2-VL-2B [33]

²We refer to the position of the first embedding token (`<|emb|>`) as the initial or earliest turn, which is only used for inference later.

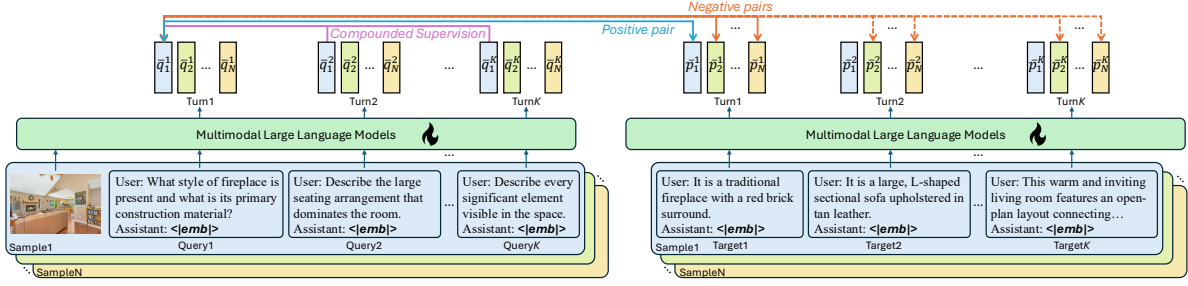


Figure 2. **Overview of MuCo.** With a multiple query-target paired dataset, **MuCo** intuitively structures the input by sequentially arranging the pairs as distinct dialogue turns. Lines are drawn between extracted embeddings, where **blue lines** denote a positive pair, **orange lines** denote negative pairs, and **purple lines** denote compounded supervision from subsequent turns to the initial turn. For clarity, lines are shown originating only from the earliest turn of the first sample in the batch. **Solid lines** represent the pair set used by conventional methods, while **dotted lines** represent the augmented pairs (*i.e.*, more learning signals) contributed by **MuCo** framework. For visual clarity, we omit the embedding function notation (*e.g.* $f(\cdot)$).

parts. To address these drawbacks, we propose concatenating queries and positive targets that share the same image using the prompt template shown in Fig. 2:

$$\bar{q}_i^j = (\mathcal{I}_i, (q_l^j)_{l \leq j}), \quad \bar{p}_i^j = (p_l^j)_{l \leq j}. \quad (3)$$

This cumulative structure enables processing multiple queries and positive targets in a single forward pass. By placing $\langle | \text{emb} | \rangle$ tokens at the end of each turn, we simultaneously extract all embeddings without re-encoding the image, significantly reducing computational overhead. Furthermore, gradients from all subsequent turns could flow back to the initial turn. The loss function for multiple queries and targets is changed as

$$\mathcal{L}_{\text{MuCo}} = \frac{1}{|\mathcal{B}|} \sum_{(\bar{q}_i^j, \bar{p}_i^j) \in \mathcal{B}} -\log \frac{\phi(\bar{q}_i^j, \bar{p}_i^j)}{\sum_{\bar{p} \in \mathcal{N}_i \cup \{\bar{p}_i^j\}} \phi(\bar{q}_i^j, \bar{p})}, \quad (4)$$

$$\text{s.t. } \mathcal{N}_i = \{\bar{p}_k^l \mid (\cdot, \bar{p}_k^l) \in \mathcal{B}, k \neq i\}. \quad (5)$$

Note that to address the second drawback, the semantic overlap issue, we exclude all other positive targets originating from the same image $\mathcal{I}_i$ as the current query $\bar{q}_i^j$ from its negative set $\mathcal{N}_i$ (the denominator in Eq. (4)). This exclusion is practically achieved by adding negative infinity to the corresponding values in the logit matrix for the pairs to be masked, as shown in Fig. 3 (b). This approach is better than Eq. (2) in optimizing computation overheads. For example, with a batch size of 1,024 images and 7 pairs per image, our method provides each query with 7,161 effective negative samples ($1024 \times 7 - 7$).

In this way, **MuCo** inherently processes inputs as sequential turns, where each subsequent embedding is conditioned on the context of all preceding turns, enabling the model to learn diversified and context-aware relationships between multiple queries and a given image. From another perspective, gradients from later turns retroactively refine earlier ones, yielding a compounded supervisory signal that encourages the initial representations to remain informative across the entire dialogue.

Dialogue-templated data (M3T). We construct synthetic corpora, named M3T, to train **MuCo** effectively, as such dialogue-oriented multimodal structures are publicly unavailable. Our key design principles are: we prioritize a large scale and multiple query-target pairs for an image for versatility and diversity, and structure the data to reflect comprehensive visual information, ensuring a thorough interpretation of both global and local image contexts. Furthermore, to maximize computational efficiency during training, we restrict positive targets to text-only formats, while images are exclusively used as part of the queries.

We take 5 million images sampled from DataComp [11], processed through a two-step pipeline combining state-of-the-art MLLM [2] and LLM [1]. First, we employ the MLLM to generate dense captions, distilling rich visual information into text. Second, we let the LLM synthesize seven diverse query-target pairs per image using the dense captions, aligned with MMEB core categories [17]: one for classification, one for retrieval, and five distinct VQA pairs composed of two requiring holistic understanding, two focusing on localized details, and one demanding creative reasoning. This process yields 35 million query-target text pairs and 5 million images. Further details and prompts are provided in the supplementary material.

4.2. Single-turn Adaptation of MuCo

Our multi-turn pretraining generalizes to the conventional multimodal benchmarks [17, 34], having a single query-target pair per sample. To align with single-query data, we introduce an adaptive strategy that employs extra contextual dialogue from a given query, allowing **MuCo** to elicit its pre-trained multi-turn capability.

We repurpose **MuCo** to handle single-query pairs. As shown in Fig. 4, we first establish an initial turn to extract the initial embedding from the query q or positive target p . In the following turns, we provide the model with its masked counterpart (*e.g.*, the masked positive target for a given query). The model is then explicitly instructed to first reconstruct the masked parts and generate a subsequent em-

	p_1	p_2	p_3	p_4
q_1	+	-	-	-
q_2	-	+	-	-
q_3	-	-	+	-
q_4	-	-	-	+

(a) Logit matrix without MuCo (Batch size: 4, # turns: 1)

	$\tilde{p}_1^1$	$\tilde{p}_1^2$	$\tilde{p}_1^3$	$\tilde{p}_1^4$	$\tilde{p}_2^1$	$\tilde{p}_2^2$	$\tilde{p}_2^3$	$\tilde{p}_2^4$
$\tilde{q}_1^1$	+	$-\infty$	$-\infty$	$-\infty$	-	-	-	-
$\tilde{q}_1^2$	$-\infty$	+	$-\infty$	$-\infty$	-	-	-	-
$\tilde{q}_1^3$	$-\infty$	$-\infty$	+	$-\infty$	-	-	-	-
$\tilde{q}_1^4$	$-\infty$	$-\infty$	$-\infty$	+	-	-	-	-
$\tilde{q}_2^1$	-	-	-	-	+	$-\infty$	$-\infty$	$-\infty$
$\tilde{q}_2^2$	-	-	-	-	$-\infty$	+	$-\infty$	$-\infty$
$\tilde{q}_2^3$	-	-	-	-	$-\infty$	$-\infty$	+	$-\infty$
$\tilde{q}_2^4$	-	-	-	-	$-\infty$	$-\infty$	$-\infty$	+

(b) Logit matrix with MuCo (Batch size: 2, # turns: 4)

Figure 3. **Logit masking strategy in our MuCo framework.** (a) The conventional method with a batch size of $N = 4$ yields a $N \times N$ (i.e. 4×4) matrix. In contrast, **MuCo** (b) uses a batch size of $N = 2$ and $k = 4$ turns to construct a larger $Nk \times Nk$ (i.e. 8×8) matrix. Crucially, our method masks out pairs originating from the same image (gray, $-\infty$) to prevent a semantic overlap issue. True positives (blue, +) and true negatives (orange, -) are used for the loss. Crucially, other pairs originating from the same image (gray, $-\infty$) are masked to prevent a semantic overlap issue.

bedding capturing the context of the entire dialogue.

We argue this process functions as an in-context reconstruction task, compelling the model to reason about the relationship between the query and target to infer the masked content. We achieve this by leveraging LLMs’ capabilities through prompting (i.e., instructing the LLM to perform reconstruction) and a special mask token ($\langle \text{mask} \rangle$); interestingly, even simple tokens such as a space or ‘.’ work equally well. This prompt-only approach avoids auxiliary objectives, ensuring simplicity and direct application to standard contrastive learning. To maintain the efficiency of text-only subsequent turns, we convert any images in the counterpart to text via image captioning. This conversion is performed offline, prior to finetuning, using a image captioning model. This approach offers flexibility, as any suitable captioning model can be employed for this preprocessing step (See Tab 8). Crucially, we use only the initial embedding without the subsequent turn to align with standard single-query test scenarios [17, 34].

Details work as follows: first, the query q and the corresponding positive target p are augmented using prompts. The augmented query and target (q', p') is defined as $q' = (q, \pi_1, \tilde{p}, \pi_2)$, $p' = (p, \pi_1, \tilde{q}, \pi_2)$, where π_1 is a describing request such as *User: Please rewrite your last response in human-readable language*, $\tilde{p}$ and $\tilde{q}$ represents the masked query and target, and π_2 requests a new embedding such as *User: Reconstruct the previous response, acknowledge my query, and seamlessly integrate the answer.*

For original mini-batch $\mathcal{B}$, we utilizes augmented set $\tilde{\mathcal{B}}$:

$$\tilde{\mathcal{B}} = \bigcup_{(q,p) \in \mathcal{B}} \{(q, p), (q, p'), (q', p), (q', p')\}. \quad (6)$$

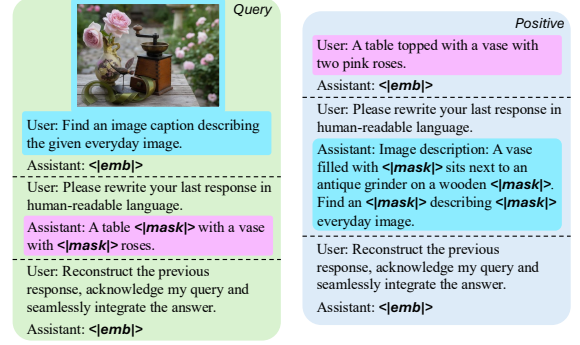


Figure 4. **Multi-turn template for fine-tuning MuCo on single-pair datasets.** We illustrate Query (left) and Positive (right) templates. The initial query (cyan) is reused as a masked target on the Positive side, and the positive target (pink) becomes a masked target on the Query side. This process simulates multi-turn interactions from a single pair, guiding the model to reconstruct its counterpart and enrich the learned embeddings.

The contrastive loss with the augmented set is defined as

$$\frac{1}{|\tilde{\mathcal{B}}|} \sum_{(q_i, p_i) \in \tilde{\mathcal{B}}} -\log \frac{\phi(q_i, p_i)}{\sum_{p \in \mathcal{N}_{p_i} \cup \{p_i\}} \phi(q_i, p)}. \quad (7)$$

Since the augmented samples are similar to the original ones, we exclude them from negative samples in contrastive loss. i.e., the negative set $\mathcal{N}_{p_i}$ is

$$\mathcal{N}_{p_i} = \{p \mid (\cdot, p) \in \tilde{\mathcal{B}}\} \setminus \{p_i, p_i^-\}, \quad (8)$$

where p_i^- denotes the opposite form of p_i : it is the augmented version if p_i is an original target in $\mathcal{B}$, and the original target if p_i is an augmented one. With this loss, the supervisory signal flows back to refine the initial embedding, enabling it to encapsulate richer relational information. In essence, our key intuition is to use simple prompting (without any further complicated elements) to leverage the MLLM’s capability to simulate multi-turn behavior on single-turn data, thereby exploiting its pretraining strength. Simple tweaks such as token masking are also believed to enhance the discriminative power of the learned embeddings. We examine several related design choices in the experimental section.

5. Experiment

5.1. Experimental Setup

Implementation details. Qwen2-VL [33] is utilized as the MLLM backbone. We apply LoRA [15] with a rank of 64 and a scaling factor of $\alpha = 64$ exclusively to the LLMs components of the model and freeze the visual encoder of the model. All training is conducted for 1 epoch on 32 NVIDIA A100 80GB GPUs. Unless otherwise specified, a global batch size of 1,024 is used. The contrastive learning temperature τ is set to 0.02, and a constant learning rate

Table 1. **Precision@1 (%) results on MMEB**, which includes 36 tasks across four categories: Classification, Visual Question Answering (VQA), Retrieval, and Visual Grounding. ID and OOD stand for the in-distribution average and out-of-distribution average metrics, respectively. **Boldface** denotes the best scores in the subset and the second-best scores are highlighted with underline.

Models	# Params	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	ID	OOD	Overall
Zeroshot setting (pretrained) on MMEB benchmark								
CLIP [28]	0.4B	42.8	9.1	53.0	51.8	–	–	37.8
MagicLens [37]	0.6B	38.8	8.3	35.4	26.0	–	–	27.8
E5-V [16]	8B	21.8	4.9	11.5	19.0	–	–	13.3
MMRet [38]	7B	47.2	18.4	<u>56.5</u>	62.2	–	–	44.0
mmE5 [6]	11B	60.6	55.7	54.7	72.4	–	–	<u>58.6</u>
MuCo-2B	2B	53.6	<u>59.9</u>	55.2	<u>74.6</u>	–	–	58.2
MuCo-7B	7B	<u>56.0</u>	64.7	58.9	75.7	–	–	61.6
Fine-tuning on MMEB benchmark (< 7B Models)								
CLIP [28]	0.4B	55.2	19.7	53.2	62.2	47.6	42.8	45.4
VLM2Vec [17]	4B	54.8	54.9	62.3	79.5	66.5	52.0	60.1
LLaVE [18]	2B	62.1	60.2	65.2	84.9	69.4	59.8	65.2
UniME [14]	4B	54.8	55.9	64.5	81.8	68.2	52.7	64.2
B3-2B [31]	2B	67.0	61.2	70.9	79.9	72.1	<u>63.1</u>	<u>68.1</u>
MoCa-3B [5]	3B	59.8	<u>62.9</u>	<u>70.6</u>	88.6	<u>72.3</u>	61.5	67.5
MuCo-2B	2B	<u>66.2</u>	65.6	70.1	<u>85.8</u>	72.9	65.0	69.5
Fine-tuning on MMEB benchmark (≥ 7B Models)								
VLM2Vec [17]	7B	61.2	49.9	67.4	86.1	67.5	57.1	62.9
MMRet [38]	7B	56.0	57.4	69.9	83.6	68.0	59.1	64.1
mmE5 [6]	11B	67.6	62.7	71.0	89.7	72.4	66.6	69.8
LLaVE [18]	7B	65.7	65.4	70.9	<u>91.9</u>	75.0	64.4	70.3
UniME [14]	7B	66.8	<u>66.6</u>	70.6	90.9	74.6	65.8	70.7
B3-7B [31]	7B	70.0	66.5	<u>74.1</u>	84.6	<u>75.9</u>	67.1	<u>72.0</u>
MoCa-7B [5]	7B	65.8	64.7	75.0	92.4	74.7	<u>67.6</u>	71.5
MuCo-7B	7B	<u>68.3</u>	71.9	73.7	90.9	77.3	69.1	73.6

Table 2. **Recall(%) Results on M-BEIR**. *S*: Single modality (text or image), *M*: Multi-modality (text and image). The arrow denotes ‘query → target’. Full per-dataset results are in the supplementary material. **Boldface** denotes the best scores in the subset and the second-best scores are highlighted with underline.

Models	# Params	<i>S</i> → <i>S</i>	<i>S</i> → <i>M</i>	<i>M</i> → <i>S</i>	<i>M</i> → <i>M</i>	Overall
UniIR [34]	0.4B	51.0	69.1	32.9	52.4	48.9
LamRA-Ret [25]	2B	47.6	66.2	<u>41.0</u>	<u>61.0</u>	<u>50.0</u>
MuCo-2B	2B	<u>49.1</u>	<u>68.2</u>	41.6	64.8	51.6
MM-Embed [23]	7B	50.9	76.9	40.0	60.9	52.7
LamRA-Ret [25]	7B	<u>52.9</u>	71.8	<u>45.2</u>	<u>65.0</u>	<u>54.9</u>
M3Task-UEM [30]	7B	54.0	<u>74.9</u>	41.9	55.7	53.9
MuCo-7B	7B	54.0	71.6	47.4	70.4	56.6

of $5e^{-5}$ is employed. For our pretraining (§4.1), multiple query-target pairs are randomly shuffled per batch instance to prevent positional bias. In our adaptive prompting strategy (§4.2), we split the text sentence into words based on spaces and then randomly mask 50% of the words using a uniform sampling strategy. If not specified, we employ Qwen2-VL-7B for image captioning in §4.2

Datasets. In Tab. 1, we use the proposed M3T dataset of 5M samples to pretrain **MuCo**. For pre-training dataset analysis in Tab. 3, we construct two randomly sampled subsets, M3T (20%) with 1M and M3T (60%) with 3M samples.

We also utilize mmE5 [6] dataset with 560K and the Mega-Pairs [38] dataset with 26M samples for comparison. The details of these datasets are described in the supplementary material. For the fine-tuning stage, models are trained on two benchmarks: MMEB [17] and M-BEIR [34]. MMEB is a benchmark containing 36 individual datasets across four categories: classification, visual question answering (VQA), retrieval, and visual grounding. MMEB reframes all four categories into ranking problems with a maximum of 1,000 candidates. The benchmark provides training samples from 20 of the 36 datasets (considered In-Distribution), while the remaining 16 are reserved exclusively for evaluation (Out-of-Distribution). M-BEIR is a benchmark focusing specifically on multimodal retrieval performance. It consists of 16 retrieval datasets across 8 retrieval tasks. In our experiments, we fine-tune our models on the respective training set for each benchmark and evaluate them on the corresponding test set.

5.2. Main Results

We provide the evaluation results on MMEB in Tab. 1 and M-BEIR in Tab. 2. **MuCo** achieves state-of-the-art (SOTA) performance on the MMEB and M-BEIR benchmark across different model scales. This superior performance stems

Table 3. **Pretraining dataset variants.** ‘None’ represents fine-tuning without pretraining. We compare against other pretraining datasets (mmE5 and MegaPair) and our M3T at various scales. The ‘mmE5’ entry under the ‘Multi-turn’ section is generated using images from the single-turn mmE5 synth dataset.

Pre-training	Dataset	Samples	ZS MMEB	FT MMEB
Single-turn	None		–	68.5
	mmE5 [6]	0.6M	55.6	68.6
	MegaPairs [38]	26M	41.5	68.7
Multi-turn MuCo -§4.1	mmE5 [6]	0.6M	57.0	69.0
	M3T (20%)	1M	57.1	69.0
	M3T (60%)	3M	57.7	69.2
	M3T	5M	58.2	69.5

from our multi-turn dialogue structure, which allows the model to learn rich context and compounded supervisory signals, thereby effectively improving the embeddings. In the zero-shot setting, our 7B model (61.6) surpasses the previous SOTA mmE5-11B [6] (58.6). In the fine-tuning setting, **MuCo** establishes new SOTA scores in both sub-7B and 7B-and-above categories. Specifically, **MuCo**-2B achieves 69.5, outperforming the previous best B3-2B [31] (68.1). Similarly, **MuCo**-7B establishes a new top score of 73.6, surpassing B3-7B [31] (72.0) and achieving the highest OOD score, which indicates superior generalization to unseen tasks. This strong performance trend continues on the M-BEIR benchmark. **MuCo**-2B (51.6) and **MuCo**-7B (56.6) outperform all competing methods. We note their robust performance in the complex multi-modal to multi-modal ($M \rightarrow M$) setting, achieving top scores of 64.8 (2B) and 70.4 (7B).

5.3. Empirical Analysis

This section analyzes our design choices. Unless specified, all experiments use **MuCo**-2B on MMEB. In tables, ZS *MMEB* denotes zero-shot performance (Precision@1) after **MuCo** pretraining (§4.1), and FT *MMEB* denotes the performance (Precision@1) after **MuCo** fine-tuning (§4.2) on the MMEB training set.

Impact of pretraining dataset. We analyze the impact of different pretraining datasets in Tab. 3 by varying pretraining method and dataset. Note that we fix **MuCo** fine-tuning (§4.2) to all cases for FT MMEB. The results reveal two key insights. First, **MuCo** pretraining is highly effective. Training on mmE5 after converting them to our multi-turn format outperforms the original single-turn pretraining (69.0 vs. 68.6 FT MMEB), demonstrating the significant advantage of our richer, multi-turn signal. Second, M3T is a high-quality dataset that enables large-scale **MuCo** pretraining with clear gains. **MuCo** pretraining performance improves monotonically from 0.6M to 5M. It implies that M3T offers mmE5-level quality at roughly $\times 10$ scale, while MegaPairs [38] degrades performance with massive data. It is noteworthy that the model without any pretraining

Table 4. **Ablation study for data composition.** ‘All’ means the full dataset. Subsequent rows show performance after sequentially removing task categories: Classification (– CLS), followed by global VQA (– global VQA; holistic understanding), local VQA (– local VQA; localized details), and creative VQA (– creative VQA; creative reasoning). All results are in the zero-shot setting on MMEB. GRD stands for visual grounding. The final row is the result of training only on the remaining Retrieval (RET) data.

Setup	CLS	VQA	RET	GRD	Overall
All	53.6	59.9	55.2	74.6	58.2
– CLS	51.7	59.2	54.6	72.5	56.9
– global VQA	51.1	57.6	54.6	71.5	56.3
– local VQA	50.8	56.1	54.1	70.2	55.5
– creative VQA	50.2	55.7	53.7	69.8	55.1

Table 5. **Analysis of scaling batch size vs. turns.** We compare the baseline method (#turns=1) with increasing batch sizes (1024 to 8192) against our proposed **MuCo** method (#turns ≥ 2) with an increasing number of turns at a fixed batch size 1024 in a zero-shot setting on MMEB. For the 2-turn and 4-turn settings, pairs were randomly sampled from the 7 available pairs. The ‘effective batch’ refers to the total number of query-target pairs (#batch $\times$ #turns), which amplifies the learning signal by providing more positive and negative pairs for the contrastive loss.

#turns	#batch	#effective batch	PFLOPs	MMEB
1	1024	1024	17.5	57.1
1	2048	2048	35.1	57.3
1	4096	4096	70.2	57.4
1	7168	7168	122.7	57.5
1	8192	8192	140.4	57.8
2	1024	2048	17.6	57.4
4	1024	4096	17.7	57.7
7	1024	7168	18.0	58.2

(None) achieves 68.5, which is competitive with the mmE5-pretrained model (68.6). This shows that **MuCo** fine-tuning alone is strong enough to surpass the previous SOTA B3-2B (68.1, Table 1) even without pre-training.

We analyze each pretraining task’s contribution in Tab. 4. Sequentially excluding tasks (– CLS, global/local/creative VQA) causes a performance drop in the corresponding task. Notably, excluding local VQA also degrades Visual Grounding (GRD), and removing creative VQA further lowers overall performance. This confirms each component of our synthesized data is beneficial.

Scaling batch-size vs. turns. We compare scaling the batch size and the number of turns in Tab. 5. For the conventional method (#turns=1), increasing the batch size from 1024 to 8192 yields performance gains (57.1 to 57.8) at a prohibitively high computational cost. In contrast, our method scales the number of turns (#turns 2 to 7) at a fixed 1024 batch size. Notably, our 7-turn model (7168 effective batch size) achieves a higher score than the 1-turn baseline with a 7168 batch size (58.2 vs. 57.5) and even surpasses the larger 8192-batch baseline (57.8) while the 7168-batch baseline incurs a massive computation overhead. This

Table 6. **Impact of compounded supervision.** Preventing causal attention to attend earlier contexts leads to a performance drop, supporting the importance of compounded supervision.

Setup	ZS MMEB	FT MMEB
w/o compounded supervision	57.3	68.4
w/ compounded supervision	58.2	69.5

Table 7. **Impact of logit masking.** While beneficial in pre-training due to its diverse multi-turn data, this strategy is critical for fine-tuning, as it prevents a severe performance collapse caused by incorrectly treating self-augmented pairs as negatives

Logit masking		ZS MMEB	FT MMEB
Pretraining	Fine-tuning		
-	-	57.7	31.1
-	✓	57.7	69.2
✓	-	58.2	30.9
✓	✓	58.2	69.5

demonstrates that our method provides a far more efficient path to achieving the benefits of a larger effective batch.

Impact of compounded supervision. We conduct an ablation study to validate the effectiveness of our compounded supervision (*i.e.*, subsequent turns provide a compounded supervisory signal that retroactively refines earlier embeddings), as shown in Tab. 6. We test a variant where this accumulation is disabled by modifying the causal attention mask in both pretraining and fine-tuning; this forces each turn to only attend to the initial image and its own tokens, isolating it from the context of preceding turns. The results demonstrate our hypothesis: disabling this compounded supervision leads to a performance drop (68.4) compared to **MuCo** (69.5). This confirms that our strategy of accumulating supervisory signals across turns is a superior approach for learning robust representations.

Impact of logit masking. We validate the importance of our logit masking strategy in Tab. 7, which is critical for fine-tuning. Disabling it (*e.g.* row (1,3) vs. row (2,4) in FT MMEB) causes a performance collapse. This is because, during fine-tuning, we construct subsequent turns using each counterpart. Without logit masking, the model treats these semantically overlapped pairs as negative pairs, leading to severe confusion that prevents learning. In contrast, the impact during pretraining is far less severe in ZS MMEB. This is because the pretraining data consists of diverse query-target pairs for each image. While this still causes a minor performance drop due to potential semantic overlap between the pairs, the diversity prevents the learning collapse seen in the fine-tuning stage.

Ablation study of the subsequent turn design for fine-tuning on single-pair dataset. We compare other design choices for the subsequent turn used in the fine-tuning stage as shown in Tab. 8. First, we analyze the counterpart masking ratio, finding that 50% yields the best performance. A lower ratio (*e.g.* 25%) is suboptimal as the masked input is

Table 8. **Ablation study of the subsequent turn design for fine-tuning on single-pair dataset.** We analyze the counterpart masking ratio, the importance of reconstruction guidance (compared to no guidance or simple rephrasing), and find that the choice of image captioning model has a minimal impact

Setup	MMEB
Counterpart masking ratio 25%	69.3
Counterpart masking ratio 50%	69.5
Counterpart masking ratio 75%	68.9
Rephrasing template	68.5
Without reconstruction guidance	69.0
Image captioning (BLIP Large [21])	69.4
Image captioning (Qwen2-VL-7B [33])	69.5

too similar to the original counterparts which provide only a minimal learning signal. Conversely, a higher ratio (*e.g.* 75%) also results in a larger performance drop because the reconstruction task becomes excessively difficult.

We also analyze the dialog template for the fine-tuning strategy. Interestingly, using just a simple rephrasing template (*User: Please rephrase your last response in embedding space \n Assistant: <|emb|>*) outperforms the **MuCo** without compounded supervision (68.4 in Tab. 6). This demonstrates our compounded supervision can improve the performance even though a simple subsequent is employed. When training without reconstruction guidance, the performance drops to 69.0 from 69.5, which confirms that the explicit reconstruction prompt is crucial for compelling the model to reason about the relationship between the pairs in Eq. 6. Finally, we observe that the choice of image captioning model (BLIP-Large vs. Qwen2-VL-7B) has a minimal impact on performance.

6. Conclusion

In this work, we propose Multi-turn Contrastive Learning (**MuCo**), a dialogue-inspired framework to overcome the limitations of conventional single-turn contrastive learning. By reframing representation learning as a multi-turn dialogue and modeling contextual dependencies, **MuCo** learns richer and more coherent embeddings than isolated pair alignment. Furthermore, **MuCo** tackles scalability bottlenecks by processing multi-turn query sequences in a single forward pass. This approach significantly reduces computational overhead while increasing the effective batch size. Supported by our new 5M-scale M3T corpus, **MuCo** achieves new state-of-the-art performance on universal multimodal embedding benchmarks. **MuCo** demonstrates that it is possible to simultaneously enhance model performance and training scalability, effectively redefining the efficiency-capacity trade-off in multimodal alignment. We believe this work opens new avenues for more efficient and context-aware multimodal representation learning systems.

Acknowledgment

This work utilized Bruno from the NAVER AI Search Platform to facilitate large-scale data processing tasks. We thank the platform for providing robust Model as a Service (MaaS) support.

References

- [1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. 2, 4
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4
- [3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 2
- [4] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. *Advances in Neural Information Processing Systems*, 35:33860–33875, 2022. 3
- [5] Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. *arXiv preprint arXiv:2506.23115*, 2025. 1, 3, 6
- [6] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8254–8275, 2025. 1, 2, 3, 6, 7
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3
- [8] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2
- [10] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datcomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 2, 4
- [12] Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (ReplANLP-2021)*, pages 316–321, 2021. 3
- [13] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yooheon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024. 2
- [14] Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2860–2869, 2025. 1, 2, 3, 6
- [15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5
- [16] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 1, 2, 6
- [17] Ziyang Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 4, 5, 6
- [18] Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*, 2025. 2, 6
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3
- [20] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3
- [21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 2, 8
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

- [23] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3, 6
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [25] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025, 2025. 1, 6
- [26] Rui Meng, Ziyang Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 2
- [27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6
- [29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2
- [30] Rohan Sharma, Changyou Chen, Feng-Ju Chang, Seongjun Yun, Xiaohu Xie, Rui Meng, Dehong Xu, Alejandro Mottini, and Qingjun Cui. Multi-modal multi-task unified embedding model (m3t-uem): A task-adaptive representation learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22783–22793, 2025. 2, 6
- [31] Raghuv eer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhui Chen, and Bhuwan Dhingra. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. In *The Thirtieth Annual Conference on Neural Information Processing Systems*, 2025. 2, 3, 6, 7
- [32] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2
- [33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 5, 8
- [34] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2024. 2, 4, 5, 6
- [35] Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6345–6356, 2025. 1
- [36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2, 3
- [37] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. Magiclens: self-supervised image retrieval with open-ended instructions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59403–59420, 2024. 2, 6
- [38] Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, 2025. 2, 3, 6, 7