Illuminating Visual Identity in Universal Multimodal Embeddings

Jiawei Cao[1,2*], Junyi Feng[2*], Jiashen Hua[2†‡], Ziheng Huang[2],
Bing Deng[2], Kaijie Wu[1‡], Chaochen Gu[1‡], Jieping Ye[2]
[1] Shanghai Jiao Tong University [2] Alibaba Group
{cjw333, kaijiewu, jacygu}@sjtu.edu.cn,
{felix.fjy, jiashen.hjs, huangziheng.hzh, dengbing.db, yejieping.ye}@alibaba-inc.com

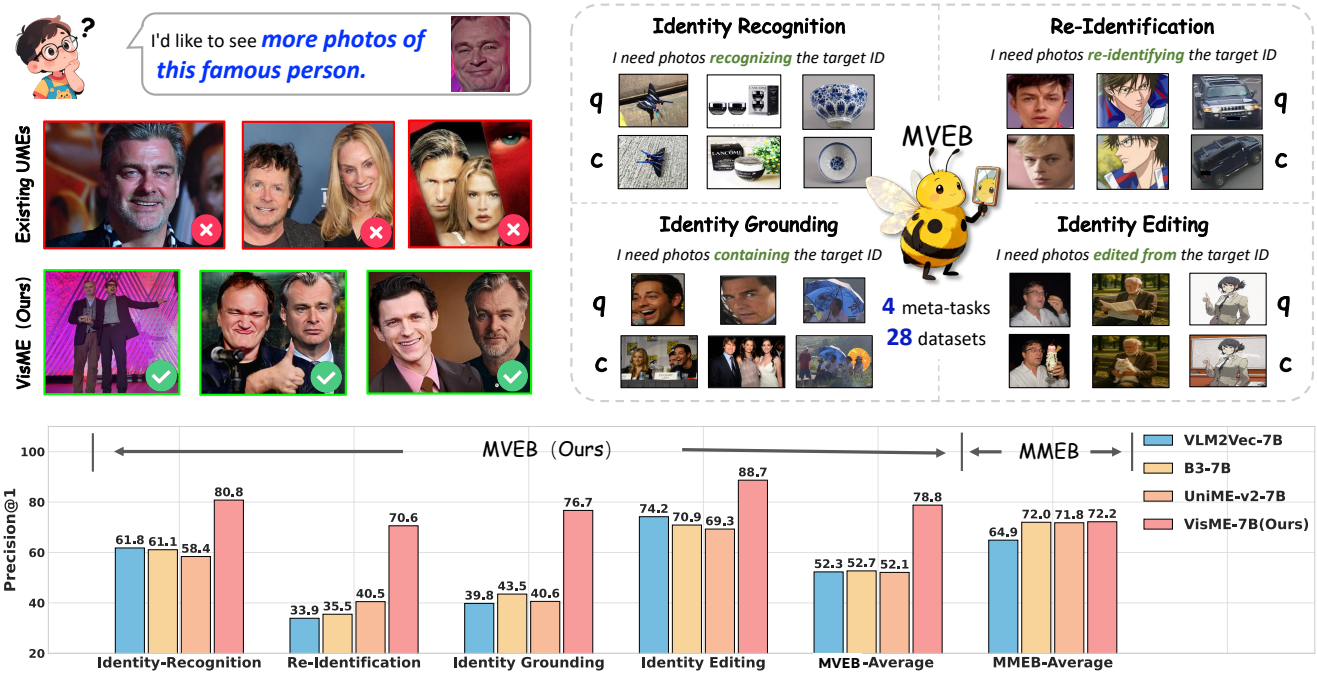

Figure 1. **Overview of the proposed VisME for identity-level multimodal embeddings.** (a) Qualitative comparison with existing UMEs on visual identity grounding tasks; (b) Illustration of four meta-tasks in MVEB (Identity Recognition, Re-Identification, Identity Grounding, and Identity Editing), covering 28 datasets; (c) Quantitative evaluation on MVEB and MMEB benchmarks.

Abstract

*Universal Multimodal Embeddings (UMEs) aim to unify various modalities and tasks into a shared representation space. In recent years, this field has witnessed substantial progress driven by the development of Multimodal Large Language Models (MLLMs). However, a crucial capability, visual identity discrimination, remains underexplored in existing UME methods, despite its critical role in a wide range of tasks, including instance retrieval, re-identification, and identity preservation in AI-generated content. To bridge this gap, we propose a unified formulation for visual identity discrimination (VisID) and introduce **MVEB** (**M**ultimodal **V**isual Identity **E**mbedding **B**enchmark), a large-scale benchmark curated from both real-world and synthetic datasets to support evaluation and training. Furthermore, we present a simple yet effective learning framework that jointly optimizes general multimodal and visual identity representations through a carefully designed identity-aware sampling mechanism. Extensive experiments demonstrate that our approach successfully endows UMEs with strong identity discrimination capability and maintains competitive general multimodal performance. We believe this work not only illuminates a critical yet neglected capability, but also takes a step toward more holistic universal multimodal embeddings. Code and data are available at MVEB.*

*Equal contribution. † Project Leader. ‡Corresponding author.

1. Introduction

Multimodal embedding models aim to encode heterogeneous modalities, such as images, text, and videos, into a unified representation space, enabling practical applications like multimodal search [63, 79] and multimodal retrieval-augmented generation [7, 75]. Early approaches, pioneered by CLIP [51], adopt a dual-encoder architecture [22, 23, 56, 74, 77], where visual and textual inputs are encoded separately and aligned by contrastive learning on large-scale image–text pairs. While effective for cross-modal alignment, these methods typically struggle with complex instructions [16] and fused modalities [79]. More recently, universal multimodal embeddings (UMEs) [16, 18, 25, 26, 38, 43, 45, 55, 66, 82] powered by multimodal large language models (MLLMs) [1, 3, 32, 58, 89] have been proposed. By encoding inputs into deep hidden states and performing contrastive fine-tuning on instruction-aware multimodal retrieval datasets [6, 25, 88], UMEs successfully inherit key capabilities from MLLMs: *(i)* a unified embedding space that natively supports fused modalities, *(ii)* strong capabilities in long-context reasoning and complex instruction following.

Despite the advances, a fundamental yet overlooked capability remains underexplored in existing UMEs: **Vis**ual **I**dentity **D**iscrimination (VisID), *i.e.*, the ability to identify images that correspond to the same visual identity as a query image, conditioned on a natural language task instruction (e.g., "find other photos of this person" or "retrieve images containing the car of the same brand"). VisID is a critical capability for both traditional vision-centric representation tasks (instance retrieval [27, 64, 72, 73], person re-identification [20, 69, 86], face recognition [11, 41]) and the recently prominent identity-preserving AI-generated content [33, 34, 59, 76]. However, current SoTA UMEs often fail to distinguish the desired identity, as illustrated in Figure 1. This is primarily because VisID has not been explicitly addressed in their design. For example, the most widely used benchmark, MMEB [25], includes only a single image-to-image subset (NIGHTS [13]) among its 36 subsets. To bridge this gap, we revisit VisID from the perspective of UMEs and make the following contributions:

- **Problem formulation**. We formally define *Visual Identity Discrimination* as a core capability for UMEs and decompose it into four practical meta-tasks: *(i) Identity Recognition*, *(ii) Re-Identification*, *(iii) Identity Grounding*, and *(iv) Identity Editing*. Some representative cases are shown in Figure 1.
- **Benchmark**. We introduce the Multimodal Visual Identity Embedding Benchmark (MVEB), a large-scale dataset designed to comprehensively evaluate and enhance VisID capabilities. We carefully design a data curation pipeline that integrates both real-world and AI-generated data. As a result, MVEB covers all four VisID meta-tasks across 28 datasets, of which 20 are allocated for training and 8 are reserved for out-of-domain evaluation.
- **Training framework**. We propose a simple yet effective joint training framework that unifies VisID learning with standard contrastive objectives. Our method employs an identity-aware sampling strategy and a tailored contrastive loss to enforce intra-identity consistency. Experiments show that it improves VisID performance by $\sim 25\%+$ on MVEB, while maintaining general multimodal retrieval accuracy on MMEB, demonstrating strong compatibility with existing UMEs.

We will release data, models, and code to support the development of identity-aware UMEs, a key step toward more capable and general-purpose multimodal representations.

2. Related Work

2.1. Dual-Encoder Multimodal Embeddings

CLIP [51] established a foundational paradigm for multimodal representation learning by employing a dual-encoder architecture within a contrastive framework to achieve cross-modal semantic alignment. Subsequent works refined this paradigm from three main perspectives: *(i)* Enhancing the text encoder. The original CLIP text encoder lacks multilingual capability and is limited to short contexts (maximum 77 tokens). Chinese-CLIP [67], jina-clip-v2 [28] and SigLIP2 [56] introduce multilingual capabilities. Recently, with the advent of LLM-based embeddings [4, 31, 83], LLM2CLIP [22], SAIL [80] and FLAME [5] utilize LLM-based text towers [4, 31] that support longer sequences and multiple languages. *(ii)* Scaling the vision encoder. Works such as EVA-CLIP [53] and InternViT [8] explore larger Vision Transformer (ViT) architectures to improve visual representation capacity. *(iii)* Improving training objectives. BLIP [35] and CoCa [74] incorporate image captioning as an auxiliary training signal, while SigLIP [56, 77] advances contrastive learning through optimized training objectives and effective scaling strategies. Another challenge for the dual-encoder paradigm is the lack of cross-modal fusion ability. To tackle composed retrieval tasks, ALIGN [23], UniIR [61] and MagicLens [79] either simply fuse features by linear addition, or incorporate additional cross-attention modules to realize cross modal interaction. However, the limited capacity of fusion modules still constrains their performance on modality-compositional tasks.

2.2. Universal Multimodal Embeddings

In response to the limitations above, a growing body of research has focused on universal multimodal embeddings (UMEs) [6, 10, 16, 24, 25, 38, 82, 88]. This emerging paradigm aims to develop a single, unified model capable of processing diverse modalities and generating embeddings

for a wide spectrum of tasks. From a modeling perspective, most UME approaches fine-tune a multimodal large language model (MLLM) [3, 32, 58, 89] using contrastive learning objectives. Several notable efforts have been made in training data construction. M-BEIR [61] aggregates 10 source multimodal datasets to cover 16 diverse retrieval scenarios. VLM2Vec [25] introduces MMEB, a comprehensive benchmark comprising numerous in-domain and out-of-domain retrieval tasks for systematic evaluation. MegaPairs [88] employs LLMs and VLMs to generate pseudo instructions for curated image pairs, yielding 26M samples. mmE5 [6] curates 560K high-quality multilingual synthetic examples. Recent works also extend UMEs to more tasks. GME [82] targets visual document retrieval. UNITE [26] and VLM2Vec-v2 [45] further incorporate the video modality. Other approaches focus on performance improvement through mining hard negative samples [30, 38, 55, 66] or introducing specially-designed training stages [16, 26]. While these methods support richer modalities and tasks, they remain focused on semantic-level alignment and largely neglect fine-grained visual identity discrimination. In contrast, our approach builds upon the UME paradigm but explicitly prioritizes this critical capability.

2.3. Visual Identity Discrimination

Several traditional visual recognition tasks are closely related to visual identity discrimination (VisID), including instance-level image retrieval (e.g., cars [29, 68], landmarks [50, 62], e-products [48, 49]), person re-identification [52, 85], and face recognition [11, 19]. Classical approaches [41, 47, 86] typically learn domain-specific representations. Although recent works [2, 12, 27, 72, 73] aim to build a universal visual embedding model for general-purpose image-to-image search, they remain confined to the *unimodal* setting. A very recent MLLM-based embedding approach, IDMR [39], explores grounded instance retrieval. However, it focuses on a single subtask. Universal multimodal embeddings still lack a comprehensive formulation and systematic investigation of these identity-centric tasks.

Visual identity has gained significant attention in AIGC, where ID-preserving image and video generation methods [33, 34, 59, 76, 84] aim to retain key attributes of a reference subject during generation. In contrast to identity preservation in generation, our work focuses on identity discrimination in representations. We believe that robust identity-aware embeddings can play important roles in AIGC, enabling higher-quality data curation and providing a principled basis for evaluation.

3. Preliminary: Problem Formulation

Before introducing our method, we revisit the formulation of universal multimodal embeddings (UMEs) and formally define visual identity discrimination (VisID) within this framework. Then, we systematically categorize related tasks into four essential meta-tasks that encompass the core aspects of VisID.

UMEs take as input image–text pairs from the multimodal space $\mathcal{X} = \mathcal{I} \times \mathcal{T}$, where $\mathcal{I}$ and $\mathcal{T}$ denote the spaces of images and text respectively. The UME model $f_\theta(\cdot)$ encodes each input sample $x = (i, t)$ into a universal embedding space: $f_\theta(x) = y$, where $y \in \mathbb{R}^D$ is a D-dimensional embedding vector. Within this framework, prior multimodal embedding models [22, 51] focus on cross-modal alignment between text-only and image-only inputs, i.e., matching (ϵ, t) with (i, ϵ), where ϵ denotes a null (empty) placeholder indicating the absence of a modality. In contrast, traditional visual recognition tasks [11, 69, 72] are inherently image-to-image, operating solely on pairs of the form (i, ϵ).

Building upon the UME framework, we provide a more comprehensive and formal definition of visual identity discrimination (VisID). From a capability perspective, VisID refers to the ability to identify images that correspond to the same visual identity as a query image, conditioned on a natural language instruction. We formalize VisID as a multimodal matching task where both the query and candidate items are represented as (i, t) pairs:

Query $q = (i_q, t_q)$: The image i_q serves as the identity reference, while the text t_q specifies the target visual content to be identified. The text t_q can be a task instruction (e.g., "find other photos of this person"), a descriptive caption, or a composition of both.

Candidate $c = (i_c, t_c)$: The image i_c is the primary subject of matching, accompanied by optional text t_c that provides generic context (*e.g.*, "Represent the image") or may even be absent ($t_c = \epsilon$).

Based on this formulation, we categorize VisID into four distinct meta-tasks according to the characteristics of the query–candidate pairs.

- **Identity Recognition.** This task requires determining whether two images belong to the same identity, where identity may be defined either as a unique visual instance or as a fine-grained semantic category. It covers tasks such as product recognition, landmark recognition, artifact matching, and fine-grained species recognition.
- **Re-Identification (Re-ID).** Re-ID is a specialized identity recognition task that focuses on matching the same individual entity (e.g., a person, face, or vehicle) across distinct visual observations or depictions.
- **Identity Grounding.** Going beyond conventional visual grounding, this task uses a reference image depicting a known identity as the query, either a cropped instance from a scene or an image of the same entity captured in a different context (*e.g.*, another photo of the same person).
- **Identity Editing.** This task addresses the robustness

of identity representation against generative transformations. By applying text-prompted edits to a source image, the identity of the subject in the target image is preserved but other attributes (e.g., style, background) are altered.

Each of the above meta-tasks presents unique challenges. Representative examples are shown in Figure 1.

4. Methodology

4.1. MVEB: Data Curation

To facilitate the training and evaluation of VisID, we construct the *Multimodal Visual Identity Embedding Benchmark* (MVEB), a comprehensive dataset spanning the full spectrum of identity-level tasks defined in Sec. 3. To ensure both the diversity and the data quality, we carefully design a three-step curation pipeline, illustrated in Figure 2.

Step 1: Collection and Screening. We aggregate raw data from diverse public sources and apply a multi-tier screening process. This includes manual verification of relevance to VisID tasks, followed by dataset-level quality assessment. We exclude entire datasets that fail to meet either the relevance or quality criteria.

Step 2: Refinement and Mining. We perform fine-grained balancing and cleaning in this step: For existing identity-based datasets, we remove long-tail samples via identity-aware re-sampling to promote domain balance, *e.g.*, for the 4.1M GLDv2 [62], we sample 7.5K identities with at most four images each, yielding ∼30K training samples.

For AIGC datasets such as GPTImageEdit [60], where hard-negative samples are scarce, we further leverage an auxiliary embedding model trained following [2, 72, 73] to perform denoising and hard-negative mining. Specifically, we first identify candidate negative pairs that share the same editing instruction but originate from different source identities. From this candidate pool, we then employ an auxiliary embedding model to select the pairs whose generated images are closest in the embedding space. These samples are challenging because they are semantically congruent with the instruction yet identity-inconsistent, forcing the model to learn more fine-grained representations.

Step 3: Task Formatting and Splitting. Each sample is assigned a task-specific natural language instruction that explicitly encodes its retrieval objective. (e.g., "Find the same object as the one in the given image."). Finally, we partition the data into training (as well as in-domain validation), and out-of-domain validation sets. Crucially, whenever feasible, we enforce a strict *identity-aware split*: no identity present in training appears in any evaluation set. For sub-datasets with insufficient unique identities, we fall back to a standard random split at the sample level.

Through the above curation pipeline, we construct MVEB, a benchmark comprising four meta-tasks, 28 sub-datasets (20 for training and 8 for out-of-domain validation), and a total of 522K samples. Detailed statistics are provided in Figure 3.

4.2. Training framework

To integrate VisID into existing UME training paradigms, we design a simple yet efficient framework that is fully compatible with standard UME pipelines. Our approach jointly addresses data sampling and loss design, enabling optimization of identity-aware representations in the current framework.

4.2.1. Data Sampling

We consider data sampling from three perspectives, *i.e.*, semantic-aware sampling, identity-aware sampling, and hard-negative sampling to construct meaningful triplets $(q, \mathcal{C}^+, \mathcal{C}^-)$. Each mini-batch $\mathcal{B}$ is formed by aggregating samples from K sub-datasets, with each sub-dataset contributing exactly M samples:

$$\mathcal{B} = \bigcup_{k=1}^{K} \mathcal{B}_k, \quad |\mathcal{B}_k| = M. \tag{1}$$

This design ensures two key properties. First, it guarantees *inter-source diversity*, exposing the model to varied data distributions within a single batch. Second, it preserves *intra-source coherence*, which naturally promotes the formation of challenging hard negatives. Specifically, samples from the same sub-dataset often share high-level semantic or stylistic properties, making them ideal negative candidates.

Semantic-aware Sampling. For semantic-aware tasks, we follow the classic in-batch contrastive learning paradigm. For a given query q, the positive set is $\mathcal{C}^+ = \{c^+\}$, where c^+ denotes the unique sample in the batch that matches q. The negative set comprises all remaining samples in the batch: $\mathcal{C}^- = \mathcal{B} \setminus \{c^+\}$, which includes both intra-source and inter-source negatives.

Identity-aware Sampling. Unlike generic cross-modal datasets, each query q in identity-centric datasets $\mathcal{X}$ belongs to an identity group $\mathcal{G}_q^{\mathcal{X}}$, defined as

$$\mathcal{G}_q^{\mathcal{X}} = \{x \in \mathcal{X} \mid \mathrm{ID}(x) = \mathrm{ID}(q)\}. \tag{2}$$

Naive in-batch sampling is ill-suited for such data, as it may inadvertently include other members of $\mathcal{G}_q^{\mathcal{X}}$ in the same batch, leading to false-negative pairs that violate the contrastive assumption.

To address this, we adopt an offline sampling scheme: before training, we pre-generate all mini-batches for the entire training schedule, ensuring two key properties: (i) each identity appears at most once per batch across all datasets, and (ii) the sampling probability of each identity is proportional to the total number of its instances in the dataset.

During training, batches are read sequentially from the pre-scheduled stream. For any anchor q, a single positive

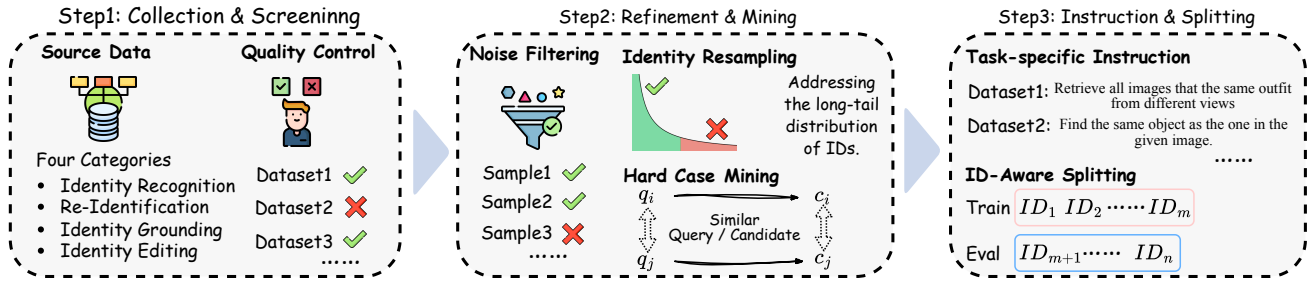

Figure 2. Overall dataset curation pipeline.

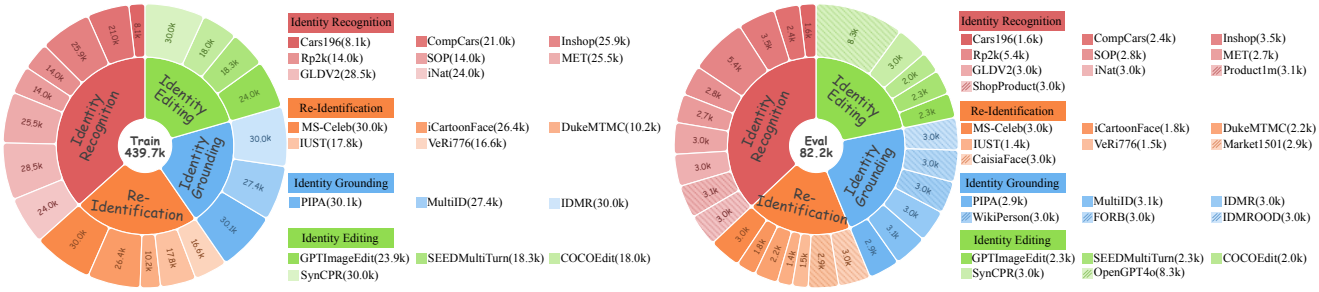

Figure 3. The distribution of MVEB dataset. Sectors shaded in light gray correspond to out-of-domain (OOD) validation sets.

candidate c^+ is then randomly sampled from $\mathcal{G}_q^{\mathcal{X}}$ to form the training triplet.

Structured hard-negative sampling. To enhance the discrimination of models, we introduce a mechanism for structured hard-negative mining. For certain sub-datasets that contain pre-defined hard-case relationships, we ensure these hard negatives are actively sampled. Specifically, when an anchor q is drawn from such a dataset, we simultaneously sample k of its designated hard negatives into the same mini-batch. This forces the model to learn more fine-grained boundaries beyond what random sampling provides.

Consequently, by combining these strategies, both identity-centric and semantic-aware tasks are optimized under a single objective $(q, \mathcal{C}^+, \mathcal{C}^-)$, enabling a unified and effective training framework.

4.2.2. Unified Contrastive Loss

To realize our unified learning objective, we adopt a single contrastive loss that jointly optimizes semantic alignment and identity discrimination through triplets $(q_i, \mathcal{C}_i^+, \mathcal{C}_i^-)$ generated by our sampling strategy. Here, q_i is a query, $c_i^+ \in \mathcal{C}_i^+$ is a positive candidate, and $\mathcal{C}_i^-$) is the set of negative candidates drawn from the same mini-batch.

Let $f_\theta(\cdot)$ denote the universal embedding function. The similarity between two inputs x_i and x_j is defined as the scaled cosine similarity:

$$Sim(x_i, x_j) = \frac{f_\theta(x_i)^\top f_\theta(x_j)}{\tau} \tag{3}$$

where $\tau > 0$ is a learnable temperature parameter.

For each query q_i, the loss encourages its embedding to be closer to that of its positive c_i^+ than to any negative $c_i^- \in \mathcal{C}_i^-$. Leveraging the natural partitioning of negatives into intra-source and inter-source groups, as ensured by our batch construction, we formulate the loss as:

$$\mathcal{L}_i = -\log \frac{e^{Sim(q_i, c_i^+)}}{e^{Sim(q_i, c_i^+)} + \sum_{c_j^- \in \mathcal{C}_i^-} e^{Sim(q_i, c_j^-)}}. \tag{4}$$

The total batch loss is the average over all queries:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum\nolimits_{i \in \mathcal{B}} \mathcal{L}_i. \tag{5}$$

This formulation is inherently universal: the same loss (Eq. 4) governs both paradigms, with task specificity emerging solely from how triplets are constructed. In semantic-aware tasks, c_i^+ is a caption aligned with q_i. In identity-discrimination tasks, it is another view or edited variant of the same entity. Meanwhile, structured hard negatives ensure fine-grained discrimination, while inter-source negatives provide broad semantic diversity.

5. Experiments

5.1. Experimental Setup

Datasets and Metrics. Our training corpus is constructed by combining the MMEB [25] and our proposed MVEB.

Table 1. Comparison of methods on benchmark datasets. **Bold** indicates the best result, underline the second best.

Models	MMEB				MVEB				Average score	
	Cls	VQA	Ret	Grd	ID-Rec	Re-ID	ID-Grd	ID-Edit	MMEB_{avg}	MVEB_{avg}
Dual-Encoder Models										
CLIP (ViT-large-14) [25]	42.8	9.1	53.0	51.8	55.3	44.3	37.4	52.5	37.8	48.2
SigLIP2 (so400m-14-384) [56]	55.7	9.1	43.6	62.2	73.1	45.6	36.8	65.4	39.4	57.1
LLM2CLIP (EVA-02) [22]	54.1	14.5	61.3	63.5	65.4	47.5	38.1	72.1	46.5	56.2
∼2B VLM Models										
VLM2Vec (Qwen2-VL-2B) [25]	58.7	49.3	65.0	72.9	50.1	24.1	32.5	53.3	59.7	40.4
GME (Qwen2-VL-2B) [82]	56.9	41.2	67.8	53.4	62.2	39.1	33.7	72.9	55.8	52.2
UniME (Phi-3.5-V) [16]	54.8	55.9	64.5	81.8	44.4	31.8	33.8	70.3	64.2	43.6
LLaVE (Aquila-VL-2B) [30]	62.1	60.2	65.2	84.9	29.1	26.6	29.4	72.9	65.2	36.4
VLM2Vec-V2.0 (Qwen2-VL-2B) [25]	62.9	56.3	69.5	77.3	61.8	33.9	39.8	74.2	64.9	52.3
B3 (Qwen2-VL-2B) [55]	**67.0**	61.2	**70.9**	79.9	53.7	13.8	31.7	59.5	68.1	40.1
VisME (Qwen2-VL-2B)	66.1	62.5	69.5	78.4	74.7	54.4	62.0	87.3	67.6	69.1
VisME (Qwen2.5-VL-3B)	63.3	**67.4**	70.5	**87.8**	**78.7**	**67.7**	**75.1**	**87.4**	**69.6**	**76.7**
∼7B VLM Models										
VLM2Vec (Qwen2-VL-7B) [25]	62.6	57.8	69.9	81.7	61.8	33.9	39.8	74.2	65.8	52.3
GME (Qwen2-VL-7B) [82]	57.7	34.7	71.2	59.3	63.3	43.3	39.8	74.6	56.0	55.3
UniME-v2 (LLaVA-OneVision-7B) [17]	65.6	68.7	73.1	90.9	58.4	40.5	40.6	69.3	71.8	52.1
LamRA (Qwen2.5-VL-7B) [43]	51.7	34.1	66.9	56.7	26.9	17.2	12.9	19.9	52.4	20.2
LLaVE (Llava-OV-7B) [30]	65.7	65.4	70.9	**91.9**	57.8	42.0	44.9	77.2	70.3	54.5
B3 (Qwen2-VL-7B) [55]	**70.0**	66.5	**74.1**	84.6	61.1	35.5	43.5	70.9	72.0	52.7
VisME (Qwen2-VL-7B)	68.9	69.1	72.8	86.1	78.6	60.6	68.6	**90.2**	72.1	74.0
VisME (Qwen2.5-VL-7B)	66.4	**70.2**	72.6	90.7	**80.8**	**70.6**	**76.7**	88.7	**72.2**	**78.8**

This unified collection comprises 40 datasets totaling 1.1M pairs. The MMEB portion contributes 20 in-distribution datasets (662K pairs) covering four general tasks: Classification (Cls), VQA, Retrieval (Ret), and Grounding (Grd). The MVEB portion adds another 20 in-distribution datasets (439K pairs) focusing on four identity-centric tasks: Identity Recognition (ID-Rec), Re-Identification (Re-ID), Visual Identity Grounding (ID-Grd), and Identity Editing (ID-Edit). For evaluation, we assess performance on the MMEB (20 in-distribution and 16 out-of-distribution test sets) and the MVEB benchmark (20 in-distribution and 8 out-of-distribution test sets). We report Precision@1 (P@1) as the primary metric for all tasks.

Implementation Details. Our model, based on Qwen2.5-VL [3] and Qwen2-VL [58], is trained on 8 NVIDIA A800 GPUs. We use LoRA (rank=16, alpha=32) [21] and GradCache [14] for memory optimization. We process images at a maximum resolution of 200,704 pixels, corresponding to 256 visual tokens, and limit the maximum sequence length at 768. The model is trained for 3,000 steps using a global batch size of 1024 and a learning rate of $1e^{-4}$. The contrastive loss uses a learnable temperature τ, initialized to 0.02, and $k = 5$ hard negatives same as the paper [55].

5.2. Main Results

We compare our method against a range of baselines, including dual-encoder models and recent UMEs. The results are summarized in Table 1.

- *Dominant Performance Across Benchmarks.* Our method establishes best results on both MMEB and MVEB benchmarks. On the MMEB, our 7B model achieves an average score of 72.2, outperforming the best (B3). The superiority is even more pronounced on our proposed identity-centric MVEB benchmark. Our 2B model alone scores 69.1, significantly surpassing all baselines, and our 7B model extends this lead, reaching an impressive 78.8. This demonstrates our model's robust general multimodal capabilities and its exceptional strength in fine-grained identity understanding.
- *Analysis of Dual-Encoder Models.* To accommodate the architectural limitations of dual-encoder models, we modified all tasks except Identity Editing into pure image-to-image retrieval tasks by eliminating the text. For Identity Editing, we generate the final representation by averaging the corresponding visual and text embeddings. Under this adapted protocol, SigLIP2 distinguishes itself as the best-performing dual-encoder on MVEB, especially

on the Identity Recognition task. This interesting phenomenon suggests that the dual-encoder architecture also holds promise for VisID tasks.

- *Shortcomings of Existing UMEs.* Among the UMEs, GME-7B demonstrates the best results. We contend that its performance stems from its diverse UMRB dataset, which likely fosters a more optimal feature space distribution. This view is supported by the VLM2Vec-V2.0 model, whose superior performance over other 2B-scale models can be linked to its training on more extensive data. Nevertheless, a clear pattern emerges: all these UME models lag significantly behind our approach on identity-level benchmarks. This performance discrepancy reinforces our central claim that visual identity understanding represents a critical, yet previously neglected, capability in the multimodal representation learning.

5.3. Ablations Analysis

In the following ablation studies, our default setting is the Qwen2.5-VL-3B model trained on a combined dataset of MMEB and MVEB.

Effectiveness of Sampling Strategies. To evaluate the effectiveness of our sampling strategy, we conduct all experiments using LoRA with rank 16 and alpha 32. The results are presented in Table 2. Our baseline model is trained using interleaved batch sampling without any constraints on the in-distribution dataset; specifically, each interleaved batch has a size of 64, yielding a total batch size of 1024. By incorporating ID-aware sampling, we observe substantial performance gains: MVEB's in-distribution accuracy improves by 11.9 points and its out-of-distribution accuracy by 12.8 points compared to the variant without ID-aware sampling. This improvement stems from the fact that, under standard interleaved sampling, each batch is drawn exclusively from a single sub-dataset, which increases the risk of false negatives, particularly when samples sharing identical IDs appear across different batches. Our ID-aware sampling explicitly mitigates this issue by ensuring that such samples are not treated as negative pairs, thereby preventing the severe performance degradation previously observed in MVEB. Furthermore, by integrating a controlled proportion of hard negative examples into the training pipeline, both MMEB and MVEB consistently outperform random sampling across both IND and OOD settings. Collectively, these two strategies allow for the unified training of both normal and ID-centric data.

Impact of Training Data. Under the proposed training framework, the formula of training data influences the model's performance preferences. We compare training on the MMEB and MVEB datasets individually versus on their combined mixture. Results are shown in Table 3. Training on a single benchmark only enhances the corresponding capability, failing to generalize to the other. In contrast, mixed training consistently improves performance on both datasets, with the gains for MVEB exceeding those for MMEB. This indicates that the two datasets complement each other, and that general mixed training particularly benefits MVEB.

Table 2. Ablation study on our proposed sampling strategy.

ID-aware Sampling	Hard Case Mining	MMEB		MVEB	
		IND	OOD	IND	OOD
✗	✗	74.2	59.4	64.3	61.2
✔	✗	74.2	59.8	76.2	74.0
✔	✔	75.1	60.3	77.7	74.1

Table 3. Ablation study on the composition of training data. Experiments are conducted on the Qwen2-VL-2B.

Data Formula \ Benchmarks	MMEB		MVEB	
	IND	OOD	IND	OOD
MMEB	71.4	59.0	51.7	51.2
MVEB	41.9	43.5	72.4	64.8
MMEB+MVEB	72.3	59.2	76.8	72.6

Table 4. Ablation study on interleave batch size.

Interleave Batch size	MMEB			MVEB		
	IND	OOD	Avg	IND	OOD	Avg
32	74.7	59.5	69.1	76.8	72.6	75.6
64	75.1	60.3	69.6	77.7	74.1	76.7
128	75.2	60.5	69.7	78.1	73.5	76.8

Influence of Interleave Batch Size. To further investigate the impact of interleave batch size on model performance, we conduct a controlled ablation study under the full sampling strategy described above, varying the interleaved batch size while keeping all other components fixed. Results are reported in Table 4. As shown in the table, increasing the interleaved batch size from 32 to 64 yields consistent performance gains across both MMEB and MVEB. However, further increasing the size from 64 to 128 leads to only marginal improvements, and even a slight degradation on the MVEB OOD split, suggesting diminishing returns and potential negative effects from excessive intra-batch homogeneity. Based on this analysis, we select an interleaved batch size of 64 as the optimal trade-off between cross-dataset mixing and preservation of ID-aware structure.

Sensitivity to Experimental Parameters. We further conduct ablation studies on key hyperparameters. As shown in Table 5, initializing the temperature in the loss function at 0.02 yields the best performance, though it is nearly indistinguishable from fixing it at 0.02 throughout training. We also evaluate the effects of batch size and LoRA rank.

Figure 4. Qualitative comparison of image retrieval results on four tasks (ID-Rec, Re-ID, ID-Grd, and ID-Edit) between our method (based on Qwen2.5-VL-7B) and VLM2Vec-7B. Green boxes denote correct or relevant results, red boxes denote failures or irrelevant outputs, and no boundary indicates that all target images have been retrieved.

Table 5. Ablation study on temperature for contrastive learning.

Temperature τ	MMEB			MVEB		
	IND	OOD	Avg	IND	OOD	Avg
0.01	74.5	60.0	69.2	77.4	72.9	76.1
0.02	75.0	60.3	69.5	77.6	74.0	76.6
0.03	74.5	60.0	69.1	77.5	72.0	75.9
0.02(learnable)	75.1	60.3	69.6	77.7	74.1	76.7

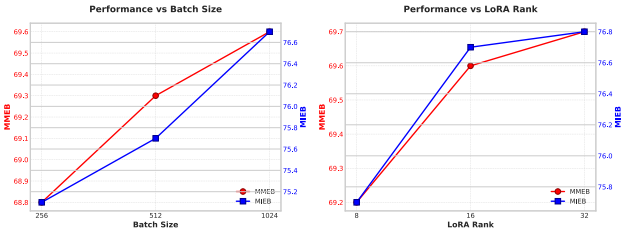

Figure 5. Performance variation with respect to batch size and LoRA rank.

Increasing the batch size from 256 to 1024 leads to modest performance gains, suggesting that hard-case mining reduces the model's sensitivity to batch size. Additionally, varying the LoRA rank reveals that rank 32 achieves optimal results.

5.4. Qualitative Analysis

Figure 4 presents a qualitative analysis of our model's performance across four meta-tasks, highlighting its superior identity-level capabilities. The model accurately identifies butterfly species by capturing subtle wing patterns—details that baseline models like VLM2Vec overlook. It also exhibits broad cross-domain versatility (Row 2). Notably, our VisME model introduces an identity grounding capability absent in prior UMEs (Row 3): given a cropped face as query, it retrieves both the original source image and other distinct images containing the same person, effectively linking partial observations to a complete identity. Finally, Row 4 shows the model's proficiency in handling compositional queries involving instruction-based editing, accurately retrieving a specific identity even when the query is modified by textual instructions.

6. Conclusion

In this work, we addressed a critical yet overlooked capability within the field of Universal Multimodal Embeddings (UMEs): visual identity discrimination (VisID). We argued that the absence of a focus on VisID limits the practical utility of current UMEs. To bridge this gap, we introduced a comprehensive framework encompassing a new problem formulation, a large-scale benchmark targeting this problem (MVEB), and an effective identity-aware training strategy. Extensive experiments show that our approach significantly enhances performance on our identity-centric MVEB by a large margin, while maintaining competitive performance on the general multimodal benchmark, *i.e.*, MMEB. Further, our work may help to establish a quantitative evaluation for identity similarity, boosting future studies about AI-generated content requiring identity preservation. We hope our work will be able to support diverse communities.

7. Acknowledgments

This work was supported by the National Natural Science Foundation of China (62227811, 62473255, 62273235). The authors are also affiliated with the Key Laboratory for System Control and Information Processing, Ministry of Education of China and Shanghai Key Laboratory for Perception and Control in Industrial Network Systems.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2

[2] Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. *arXiv preprint arXiv:2304.05884*, 2023. 3, 4

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 6

[4] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. 2

[5] Anjia Cao, Xing Wei, and Zhiheng Ma. Flame: Frozen large language models enable data-efficient language-image pretraining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4080–4090, 2025. 2

[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. *arXiv preprint arXiv:2502.08468*, 2025. 2, 3

[7] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, 2022. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[9] Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025. 2, 6

[10] Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Qi Guo, Ser-Nam Lim, et al. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025. 2

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 3

[12] Morris Florek, David Tschirschwitz, Björn Barz, and Volker Rodehorst. Efficient and discriminative image feature extraction for universal image retrieval. In *DAGM German Conference on Pattern Recognition*, pages 164–180. Springer, 2024. 3

[13] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 2

[14] Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983*, 2021. 6

[15] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 2, 6

[16] 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. *arXiv preprint arXiv:2504.17432*, 2025. 2, 3, 6

[17] Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *arXiv preprint arXiv:2510.13515*, 2025. 6

[18] Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, 2025. 2

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 3, 2, 5

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[22] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024. 2, 3, 6

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[24] 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. 2

[25] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*. 2, 3, 5, 6

[26] Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Fuzheng Zhang, Guorui Zhou, et al. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*, 2025. 2, 3

[27] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14777–14787, 2025. 2, 3

[28] Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024. 2

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3, 2, 4

[30] 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. 3, 6

[31] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. 2

[32] 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. 2, 3

[33] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. *arXiv preprint arXiv:2503.12689*, 2025. 2, 3

[34] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19406–19416, 2025. 2, 3

[35] 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. 2

[36] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 3, 4

[37] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2, 6

[38] 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*. 2, 3

[39] Bangwei Liu, Yicheng Bao, Shaohui Lin, Xuhong Wang, Xin Tan, Yingchun Wang, Yuan Xie, and Chaochao Lu. Idmr: Towards instance-driven precise visual correspondence in multimodal retrieval. *arXiv preprint arXiv:2504.00954*, 2025. 3, 2, 6

[40] Delong Liu, Haiwen Li, Zhaohui Hou, Zhicheng Zhao, Fei Su, and Yuan Dong. Automatic synthetic data and fine-grained adaptive feature alignment for composed person retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 3, 6

[41] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2, 3

[42] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2016. 2, 5

[43] 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. 2, 6

[44] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2, 4

[45] Rui Meng, Ziyan 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, 3

[46] Alireza Sedighi Moghaddam, Fatemeh Anvari, Mohammadjavad Mirshekari Haghighi, Mohammadali Fakhari, and Mo-

hammad Reza Mohammadi. A culturally-aware benchmark for person re-identification in modest attire. *Engineering Applications of Artificial Intelligence*, 158:111494, 2025. 2, 5

[47] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 3

[48] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 3, 2, 4

[49] Jingtian Peng, Chang Xiao, and Yifan Li. Rp2k: A large-scale retail product dataset for fine-grained image classification. *arXiv preprint arXiv:2006.12634*, 2020. 3, 2, 4

[50] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 3

[51] 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. 2, 3

[52] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 3, 2, 5

[53] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[54] Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Visual named entity linking: A new dataset and a baseline. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2403–2415, 2022. 2, 6

[55] Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, Bhuwan Dhingra, et al. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *arXiv preprint arXiv:2505.11293*, 2025. 2, 3, 6

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 6

[57] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 3, 4

[58] 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. 2, 3, 6

[59] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3

[60] Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025. 4, 2, 6

[61] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2024. 2, 3

[62] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 3, 4, 2

[63] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 2

[64] Pengxiang Wu, Siman Wang, Kevin Dela Rosa, and Derek Hu. Forb: a flat object retrieval benchmark for universal image embedding. *Advances in Neural Information Processing Systems*, 36:25448–25460, 2023. 2, 6

[65] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Towards controllable and id consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025. 2, 6

[66] Youze Xue, Dian Li, and Gang Liu. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. *arXiv preprint arXiv:2506.02020*, 2025. 2, 3

[67] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. 2

[68] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 3, 2, 4

[69] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 2, 3

[70] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 5

[71] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021. 2, 4

[72] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovskỳ, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 11290–11301, 2023. 2, 3, 4

[73] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, and Ondřej Chum. Udon: Universal dynamic online distillation for generic image representations. *Advances in Neural Information Processing Systems*, 37:86836–86859, 2024. 2, 3, 4

[74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[75] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 2

[76] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025. 2, 3

[77] 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

[78] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11782–11791, 2021. 2, 4

[79] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024. 2

[80] Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14604–14614, 2025. 2

[81] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4804–4813, 2015. 2, 6

[82] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. 2, 3, 6

[83] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. 2

[84] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14464–14474, 2025. 3

[85] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 3, 2, 5

[86] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2, 3

[87] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2264–2272, 2020. 2, 5

[88] 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

[89] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2, 3