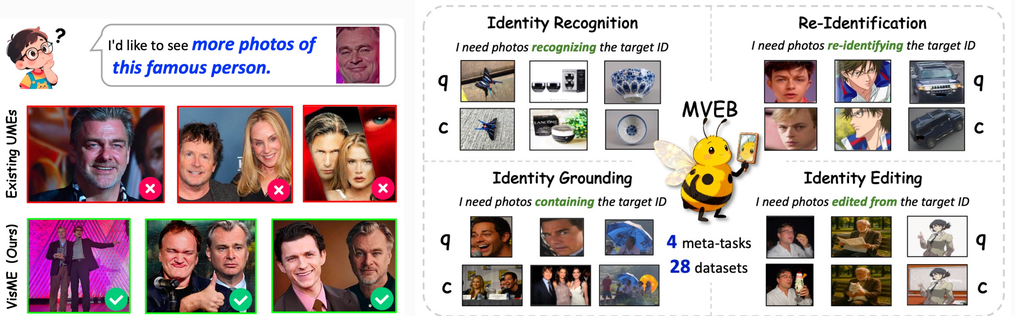
Illuminating Visual Identity in Universal Multimodal Embeddings

Jiawei Cao[1]; 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

Project Page | WeChat

Overview



Motivation :

- Universal Multimodal Embeddings (UMEs) powered by MLLMs excel at semantic alignment but systematically fail at Visual Identity Discrimination (VisID).
- VisID requires identifying images with the same visual identity, conditioned on natural language instructions.
- Existing benchmarks and training data provide limited support for identity-centric tasks.

Contributions :

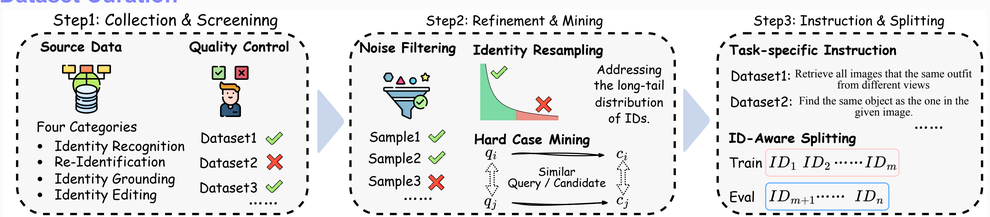
- Formulate Visual Identity Discrimination (VisID) tasks for UMEs.
- Introduce MVEB (Multimodal Visual Identity Embedding Bnchmark) .
- Propose VisME, a joint training framework for both semantic and identity discrimination .

Methods and Materials

Formulating VisID (4 Meta-Tasks)

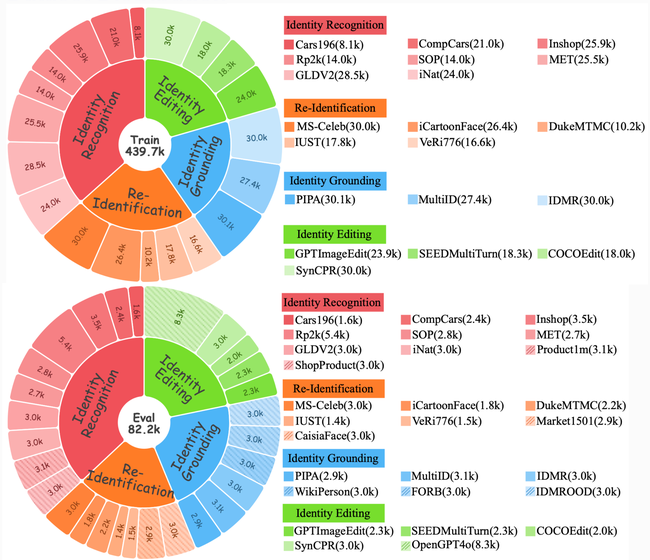
1.**Identity Recognition**: Recognizing unique instance and fine-grained categories (artifacts, products, landmarks).
2.**Re-Identification**: Matching specific entities (faces, vehicles) across distinct visual observations or depictions.
3.**Identity Grounding**: Retrieving same identity from global and cropped views.
4.**Identity Editing**: Preserving identity attributes against text-prompted generative edits.

Dataset Curation



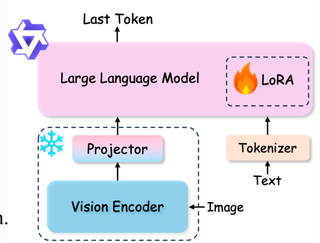
Methods and Materials

Dataset Distribution



Training Framework: VisME

- Base model: Qwen2-VL / Qwen2.5-VL as MLLM for UMEs
- Freeze vision encoder and projector, only finetune the LLM with LoRA.
- Structured hard negtative mining improves fine-grained discrimination.

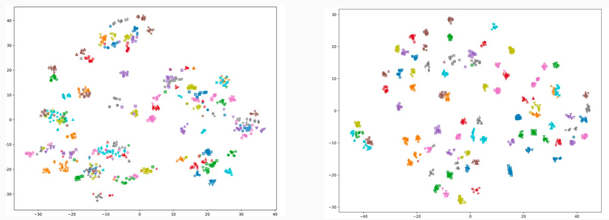


- Identity-aware sampling: one identity appears at most once in each mini-batch to avoid false negatives.
- Unified loss function for both semantic and identity tasks.

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

Results

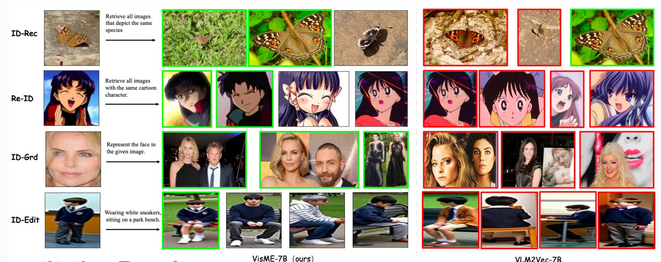
t-SNE Visualization (iNat Embeddings)



(a) VLM2Vec-7B (b) VisME-7B

VisME produces significantly more cohesive clusters for each identity, whereas VLM2Vec's embeddings are notably more dispersed.

Qualitative Retrieval Comparison



Quantitative Results

Models	MVEB				Average score	
	ID-Rec	Re-ID	ID-Grd	ID-Edit	$MMEB_{avg}$	$MVEB_{avg}$
VLM2Vec (Qwen2-VL-7B)	61.8	33.9	39.8	74.2	65.8	52.3
GME (Qwen2-VL-7B)	63.3	43.3	39.8	74.6	56.0	55.3
UniME-v2 (LLaVA-OneVision-7B)	58.4	40.5	40.6	69.3	71.8	52.1
LamRA (Qwen2.5-VL-7B)	26.9	17.2	12.9	19.9	52.4	20.2
LLaVE (Llava-OV-7B)	57.8	42.0	44.9	77.2	70.3	54.5
B3 (Qwen2-VL-7B)	61.1	35.5	43.5	70.9	72.0	52.7
VisME (Qwen2-VL-7B)	78.6	60.6	68.6	**90.2**	72.1	74.0
VisME (Qwen2.5-VL-7B)	**80.8**	**70.6**	**76.7**	88.7	**72.2**	**78.8**

- Compared with other models, VisME obtains the best **MVEB** score: 78.8
- Competitive **MMEB** score: 72.2, preserving general multimodal ability.

VisME achieves a significant improvement on Visual Identity Discrimination tasks and maintains competitive performance on general semantic retrieval tasks!