# Illuminating Visual Identity in Universal Multimodal Embeddings

## Supplementary Material

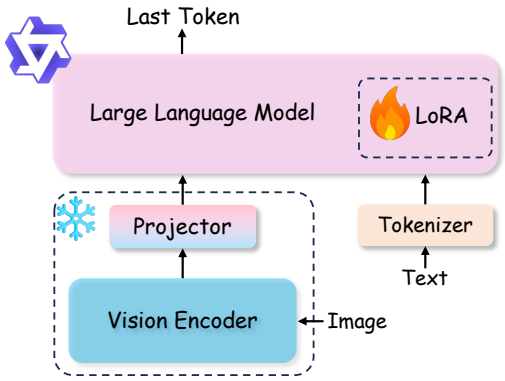


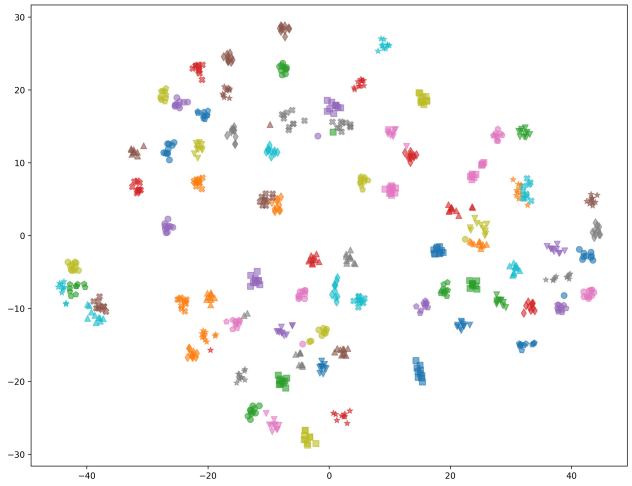

Figure 6. Our VisME model, based on the QwenVL series, freezes the vision components and fine-tunes only the LLM component with LoRA.

## A. Model Architecture

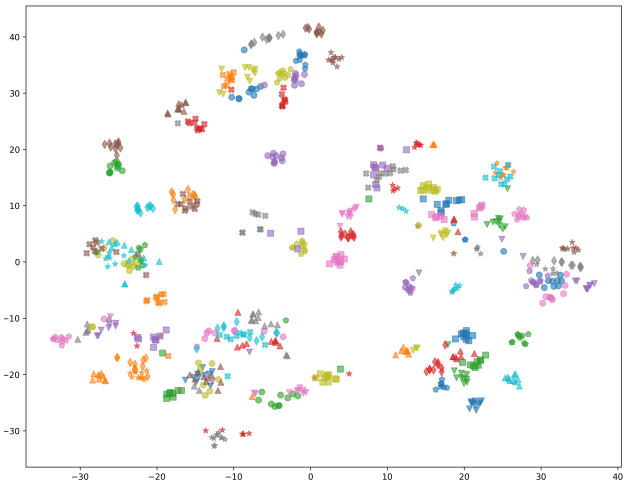
Our VisME utilizes the QwenVL series as the foundational model. The embedding is derived from the hidden states of the last layer corresponding to the last token. During training, we freeze the vision tower and only fine-tune the LLM component with LoRA while maintaining the causal mask.

## B. Analysis on Identity Discriminability

To demonstrate identity discriminability of VisME, we visualize the feature distributions of VisME and the baseline model VLM2VEC on a subset of sampled identity (ID) data using t-SNE. Specifically, we randomly sample 80 identities (800 samples) from the enlarged iNat evaluation set (used in Table 7) and employ t-SNE to visualize feature distributions of VisME-7B and VLM2Vec-7B. As shown in Figure 7, VisME produces significantly more cohesive clusters for each identity, whereas VLM2Vec's embeddings are notably more dispersed. This provides qualitative evidence that our model effectively discriminates various identities.



(a) VisME



(b) VLM2Vec

Figure 7. t-SNE of VisME-7B and VLM2Vec-7B embeddings on the extended iNat dataset. Each color–shape combination corresponds to a distinct identity, with 10 samples per identity.

## C. Detailed Performance of MIEB

This section presents a detailed comparison of VisME against previous UME models, with full results provided in Table 6.

The curation pipeline of MIEB guarantees that there is no overlap of specific images between the training and evaluation sets for any sub-dataset, except for the sub-datasets whose original splits contain overlaps. This evaluation protocol is specifically designed to assess the model's representation robustness on unseen identities.

Overall, VisME achieves substantial performance gains over prior UMEs across both IND and OOD settings. The most significant improvements are observed in the Re-Identification and Identity Grounding meta-tasks. This suggests that previous UMEs possess a less explicit capacity for perceiving identity-related information, especially in domains like faces. In contrast, our proposed Visual Identity Grounding task requires models to not only represent a face but also to perceive its specific identity within a global im-

Table 6. Performance of various models on the full MIEB benchmark. Rows shaded in red indicate OOD evaluation sets; rows shaded in gray correspond to meta-task or overall scores.

| Task | ID-Unseen? | VLM2Vec | GME | LamRA | LLaVE | B3 | VisME (Qwen2.5-VL-3B) | VisME (Qwen2.5-VL-7B) |
|---|---|---|---|---|---|---|---|---|
| **Identity Recognition** | | | | | | | | |
| Cars196 [29] | ✗ | 63.5 | 72.1 | 36.4 | 55.7 | 65.7 | 89.2 | 91.8 |
| CompCars [68] | ✔ | 58.1 | 74.2 | 38.7 | 61.1 | 61.9 | 78.3 | 81.6 |
| Inshop [44] | ✔ | 61.4 | 75.5 | 12.7 | 80.2 | 66.8 | 91.6 | 92.2 |
| Rp2k [49] | ✔ | 29.3 | 42.2 | 13.8 | 47.3 | 32.6 | 67.7 | 70.2 |
| SOP [48] | ✔ | 66.1 | 67.0 | 25.6 | 71.9 | 73.6 | 76.5 | 77.7 |
| MET [71] | ✔ | 64.5 | 68.2 | 22.3 | 71.0 | 67.8 | 76.6 | 79.7 |
| GLDV2 [62] | ✔ | 70.8 | 73.3 | 55.8 | 68.1 | 68.3 | 93.2 | 95.0 |
| iNat [57] | ✗ | 29.8 | 40.9 | 30.7 | 33.2 | 31.9 | 72.2 | 75.8 |
| Product1m [78] | ✔ | 43.0 | 48.1 | 30.2 | 42.0 | 76.3 | 80.9 | 81.9 |
| ShopProduct | ✔ | 60.3 | 71.6 | 2.7 | 47.6 | 66.7 | 60.7 | 61.8 |
| All ID-Rec | | 61.1 | 63.3 | 26.9 | 57.8 | 55.3 | 78.7 | 80.8 |
| **Re-Identification** | | | | | | | | |
| MS-Celeb [19] | ✔ | 21.4 | 29.1 | 18.7 | 24.3 | 19.2 | 58.8 | 68.5 |
| iCartoonFace [87] | ✔ | 29.7 | 49.2 | 6.9 | 58.5 | 37.0 | 72.5 | 75.8 |
| DukeMTMC [52] | ✔ | 22.3 | 35.7 | 5.4 | 26.7 | 30.9 | 77.7 | 78.7 |
| IUST [46] | ✔ | 33.7 | 35.0 | 12.1 | 33.0 | 37.6 | 57.8 | 59.4 |
| VeRi776 [42] | ✔ | 67.7 | 79.0 | 42.7 | 79.4 | 68.3 | 82.8 | 83.7 |
| Market1501 [85] | ✔ | 34.3 | 57.9 | 22.1 | 56.1 | 43.5 | 81.0 | 81.3 |
| CasiaFace [70] | ✔ | 12.1 | 17.2 | 12.6 | 15.9 | 11.8 | 43.0 | 47.6 |
| All Re-ID | | 31.6 | 43.3 | 17.2 | 42.0 | 35.5 | 67.7 | 70.6 |
| **Identity Grounding** | | | | | | | | |
| PIPA [81] | ✔ | 17.3 | 12.8 | 1.9 | 27.0 | 20.4 | 62.4 | 64.6 |
| MultiID [65] | ✔ | 10.0 | 12.5 | 2.9 | 15.4 | 12.3 | 77.7 | 80.5 |
| IDMR [39] | ✔ | 41.8 | 41.2 | 10.1 | 50.0 | 51.0 | 70.0 | 71.0 |
| WikiPerson [54] | ✔ | 21.4 | 33.6 | 3.8 | 30.5 | 26.3 | 78.3 | 83.4 |
| FORB [64] | ✔ | 81.4 | 84.9 | 35.2 | 82.0 | 87.2 | 90.8 | 89.7 |
| IDMROOD [39] | ✔ | 58.1 | 53.8 | 23.7 | 64.3 | 63.5 | 71.5 | 71.3 |
| All ID-Grd | | 38.3 | 39.8 | 12.9 | 44.9 | 43.5 | 75.1 | 76.7 |
| **Identity Editing** | | | | | | | | |
| GPTImageEdit [60] | ✔ | 77.8 | 84.7 | 27.6 | 91.1 | 85.5 | 94.6 | 95.0 |
| SEEDMultiTurn [15] | ✔ | 53.2 | 73.6 | 19.3 | 63.7 | 59.8 | 89.6 | 91.1 |
| SynCPR [40] | ✔ | 49.5 | 62.9 | 5.8 | 74.4 | 76.0 | 94.1 | 94.0 |
| COCOEdit [37] | ✔ | 52.2 | 65.7 | 22.0 | 65.8 | 56.5 | 71.7 | 75.9 |
| OpenGPT4o [9] | ✔ | 71.6 | 86.2 | 24.9 | 90.9 | 79.8 | 86.8 | 87.5 |
| All ID-Edit | | 60.9 | 74.6 | 19.9 | 77.2 | 70.9 | 87.4 | 88.7 |
| **Final Scores** | | | | | | | | |
| All IND | | 46.0 | 54.7 | 20.6 | 54.9 | 51.0 | 77.7 | 80.1 |
| All OOD | | 47.8 | 56.7 | 19.4 | 53.7 | 56.9 | 74.1 | 75.6 |
| All | | 46.5 | 55.3 | 20.2 | 54.5 | 52.7 | 76.7 | 78.8 |

age context. This poses a significant challenge for previous UMEs, as evidenced by their performance on the PIPA, MultiID, and WikiPerson datasets.

Furthermore, scaling our model from 3B to 7B parame-

| Sub-Dataset | $N_{tgt}$ | $N_{cand}$ | VLM2Vec | VisME |
|---|---|---|---|---|
| iNat [57] | 2.0 | 3k | 29.8 | 75.8 |
| | 9.0 | 100k | 17.5 ↓41.3% | 56.9 ↓24.9% |
| SOP [48] | 1.9 | 3k | 66.0 | 77.7 |
| | 3.3 | 100k | 50.6↓23.3% | 67.3↓13.4% |
| ShopProduct [36] | 1.0 | 3k | 60.3 | 61.8 |
| | 1.0 | 44k | 46.5↓22.9% | 48.8↓21.0% |
| MS-Celeb [19] | 2.0 | 3k | 21.4 | 68.5 |
| | 4.0 | 50k | 11.4↓46.7% | 57.5↓16.1% |
| SynCPR [40] | 1.0 | 3k | 49.5 | 94.0 |
| | 1.0 | 100k | 16.9↓65.9% | 67.9↓27.8% |

Table 7. Effect of scaling the candidate pool from 3k to up to 100k on retrieval performance for VLM2Vec and VisME across MIEB sub-datasets. Relative performance drops are marked in red.

ters yields consistent improvements. The performance uplift is most pronounced on face recognition datasets, such as MS-Celeb and CasiaFace. This indicates that larger-scale models have a stronger capability for capturing fine-grained identity representations, a crucial attribute for tasks involving human faces.

## D. Analysis on Larger Candidate Pool

The MIEB benchmark employs a candidate pool of approximately 3k items per sub-dataset, a size selected to balance evaluation diversity with computational feasibility. To better mimic real-world deployment settings that involve large-scale candidate galleries, we study the effect of candidate pool size and report performance comparisons on selected sub-datasets with an enlarged candidate set.

As shown in Table 7, expanding the candidate pool leads to a performance decline for both VLM2Vec and VisME. Notably, VisME exhibits a smaller relative performance drop (in percentage terms) compared to VLM2Vec. This indicates that VisME not only surpasses existing UME models in absolute performance but also demonstrates superior robustness in large-scale candidate scenarios. Nevertheless, the observed performance degradation underscores a key challenge: current models still struggle with retrieval tasks involving large candidate pools. This highlights a substantial direction for future improvement. To this end, in addition to the standard evaluation configuration used in the main manuscript, we will release a larger-scale version of the MIEB to support the community in conducting more thorough evaluations.

## E. Examples of MIEB Dataset

In this section, we present one representative sample from each sub-dataset of our MIEB dataset, as summarized in Tables 8, 9, 10, 11. The columns in these tables are defined as follows:

- **OOD?**: Indicates whether the dataset is out-of-distribution.
- **Query Image**: The input query image.
- **Query Text**: The input query text.
- **Target Image**: The ground-truth target image corresponding to the query.
- $N_{tgt}$: The average number of target items per query within the candidate set.
- $N_{cand}$: The total size of the candidate pool.

These representative examples illustrate the diversity and structure of the MIEB dataset, providing a clear reference for understanding the characteristics and scale of each sub-dataset.

Table 8. Examples of the identity recognition task from MIEB.

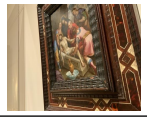
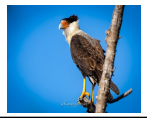
| Dataset | OOD? | Query Image | Query Text | Target Image | $N_{tgt}$ | $N_{cand}$ |
|---|---|---|---|---|---|---|
| Cars196 [29] | ✗ |  | Represent the image with the following text. Find a car of the same model, regardless of color. |  | 7.0 | 1.5k |
| CompCars [68] | ✗ |  | Represent the image with the following text. Find a car of the same model, regardless of color. |  | 2.0 | 2.3k |
| Inshop [44] | ✗ |  | Represent the image with the following text. Retrieve all images that the same outfit from different views. |  | 6.2 | 3.1k |
| Rp2k [49] | ✗ |  | Represent the image with the following text. Find an image of the same product as the one in the given image. |  | 1.9 | 5.4k |
| SOP [48] | ✗ |  | Represent the image with the following text. Find an image of the same product as the one in the given image. |  | 1.9 | 2.8k |
| MET [71] | ✗ |  | Represent the image with the following text. Find an image of the same artifact as the one in the given image. |  | 1.8 | 2.7k |
| GLDv2 [62] | ✗ |  | Represent the image with the following text. Retrieve all images that depict the same landmark. |  | 2.0 | 3.0k |
| iNat [57] | ✗ |  | Represent the image with the following text. Retrieve all images that depict the same species. |  | 2.0 | 3.0k |
| Product1m [78] | ✔ |  | Represent the product in the given image. |  | 14.2 | 2.2k |
| ShopProduct [36] | ✔ |  | Find an image of a model wearing the product from the given image. |  | 1.0 | 3.0k |

Table 9. Examples of the re-identification task from MIEB.

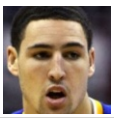
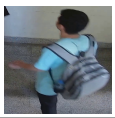
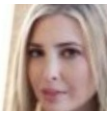
| Dataset | OOD? | Query Image | Query Text | Target Image | $N_{tgt}$ | $N_{cand}$ |
|---|---|---|---|---|---|---|
| MS-Celeb [19] | ✗ |  | Represent the face with the following text. Retrieve all images with the same facial identity. |  | 2.0 | 3.0k |
| iCartoonFace [87] | ✗ |  | Represent the face with the following text. Retrieve all images with the same cartoon character. |  | 2.0 | 1.8k |
| DukeMTMC [52] | ✗ |  | Represent the person with the following text. Re-identify the person in the given image. |  | 4.0 | 4.4k |
| IUST [46] | ✗ |  | Represent the person with the following text. Re-identify the person in the given image. |  | 4.0 | 2.6k |
| Veri776 [42] | ✗ |  | Represent the image with the following text. Retrieve vehicles with the same identity as the query image. |  | 10.0 | 2.6k |
| Market1501 [85] | ✔ |  | Represent the person with the following text. Re-identify the person in the given image. |  | 4.0 | 3.0k |
| Casiaface [70] | ✔ |  | Represent the face with the following text. Retrieve all images with the same facial identity. |  | 2.0 | 3.0k |

Table 10. Examples of the visual identity grounding task from MIEB.

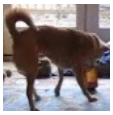
| Dataset | OOD? | Query Image | Query Text | Target Image | $N_{tgt}$ | $N_{cand}$ |
|---|---|---|---|---|---|---|
| PIPA [81] | ✗ |  | Represent the face in the given image. |  | 5.0 | 2.9k |
| MultiID [65] | ✗ |  | Represent the face in the given image. |  | 5.8 | 3.1k |
| IDMR [39] | ✗ |  | Find me an image containing the object in the given image with the following caption: The dog is standing on a blue rug in front of a television, with a person bending over nearby. |  | 1.0 | 0.1k |
| WikiPerson [54] | ✔ |  | Represent the face in the given image. |  | 1.0 | 3.0k |
| FORB [64] | ✔ |  | Represent the image with the following text. Find the same object as the one in the given image. |  | 1.0 | 20.0k |
| IDMROOD [39] | ✔ |  | Given the knife in the image, find an everyday image that contains the knife: |  | 1.0 | 1.0k |

Table 11. Examples of the identity editing task from MIEB.

| Dataset | OOD? | Query Image | Query Text | Target Image | $N_{tgt}$ | $N_{cand}$ |
|---|---|---|---|---|---|---|
| GPTImageEdit [60] | ✗ |  | Represent the image with the following editing text. Replace the pigeons with squirrels. |  | 1.0 | 2.9k |
| SEEDMultiTurn [15] | ✗ |  | Represent the image with the following editing text. Change the background to a street background. |  | 1.0 | 3.1k |
| COCOEdit [37] | ✗ |  | Represent the image with the following editing text. There are two zebras standing in a field.. |  | 1.0 | 2k |
| SynCPR [40] | ✗ |  | Find the same person described by the following text. Wearing white sneakers, sitting on a park bench. |  | 1.0 | 3.0k |
| OpenGPT4o [9] | ✔ |  | Represent the image with the following text. Add a futuristic visor to the character's eyes. |  | 1.0 | 20.0k |