# All in One Framework for Multimodal Re-identification in the Wild

Supplementary Material

## Differences between UNIReID and AIO

There are three distinctions between UNIReID and AIO:
1) **Divergent Goals:** UNIReID and AIO fundamentally differ in their objectives. UNIReID aims to construct a multimodal model for intra-domain retrieval with the descriptive query. At the same time, AIO is explicitly crafted for universal retrieval in real-world scenarios, with four arbitrary modalities or their combinations. Notably, all experiments in this paper follow a zero-shot generalizable setting, which is inapplicable for UNIReID.
2) **Different Challenges:** UNIReID demands paired multimodal data. In comparison, AIO confronts even more challenging scenarios, involving unpaired heterogeneous multimodal data, with imbalanced and missing modalities. Thus, we introduce synthesized modalities and build connections among imbalanced modalities.
3) **Disparate Approach:** UNIReID incorporates multiple tasks to accommodate uncertain multimodal input. The number of optimization objectives of UNIReID grows exponentially with the number of modalities, making it hard to extend to more modalities and hindering its scalability. Conversely, AIO designs a flexible solution, treating uncertain multimodal input as variable input lengths. It leverages the adaptable nature of the transformer architecture, simplifying the integration of additional modalities. Furthermore, UNIReID employs separate encoders for various modalities, resulting in a lack of synergy between distinctive modalities. Different from UNIReID, AIO leverages a shared foundation model as the backbone to collaboratively learn comprehensive knowledge from heterogeneous multimodal data to complement each other and enhance its generalizablity in real-world scenarios.
All these differences make AIO more robust and generalizable than UNIReID in real scenarios.

## Limitation

1) The computational complexity of AIO, necessitating $\mathcal{O}(n^2 \times D)$ operations for processing token embeddings $E^A, E^R, E^I, E^S, E^T$, particularly in the context of multimodal input, imposes a substantial memory cost and computational burden. This complexity poses challenges in scalability for incorporating additional modalities and deployment on resource-constrained edge devices. We assess the inference speed across varying numbers of modalities. Tab. 9 shows that the computation complexity escalates exponentially with the increase in the number of modalities, as anticipated.
2) Furthermore, it is worth noting that the implementation of multimodal ReID on synthetic data may not perfectly align with real-world scenarios, but also brings valuable insights for future works.
3) Moreover, the learnable parameters within the tokenizer are constrained compared to approaches that fine-tune the entire backbone, presenting a double-edged sword. While AIO is lightweight and user-friendly, it may not capture as much detailed knowledge as some alternatives. To address this challenge, a promising way is to selectively unfreeze a subset of deep layers within the backbone model, a direction we plan to investigate in future work.

| Number of Modalities | Inference Speed (ms) |
|---|---|
| 1 | 10.23 |
| 2 | 47.66 |
| 4 | 181.32 |

Table 9. **Computation complexity in the different number of input modalities.** All results are calculated with 700 samples.