ReID5o: Achieving Omni Multi-modal Person Re-identification in a Single Model

Jialong Zuo [1] **Yongtai Deng** [1] **Mengdan Tan** [1] **Rui Jin** [1] **Dongyue Wu** [1]
Nong Sang [1] **Liang Pan** [2] **Changxin Gao** [1*]

[1] National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology,
[2] Shanghai AI Laboratory.
{jlongzuo, cgao}@hust.edu.cn

Code & Dataset: https://github.com/Zplusdragon/ReID5o_ORBench

Abstract

In real-word scenarios, person re-identification (ReID) expects to identify a person-of-interest via the descriptive query, regardless of whether the query is a single modality or a combination of multiple modalities. However, existing methods and datasets remain constrained to limited modalities, failing to meet this requirement. Therefore, we investigate a new challenging problem called Omni Multi-modal Person Re-identification (OM-ReID), which aims to achieve effective retrieval with varying multi-modal queries. To address dataset scarcity, we construct **ORBench**, the first high-quality multi-modal dataset comprising 1,000 unique identities across five modalities: RGB, infrared, color pencil, sketch, and textual description. This dataset also has significant superiority in terms of diversity, such as the painting perspectives and textual information. It could serve as an ideal platform for follow-up investigations in OM-ReID. Moreover, we propose **ReID5o**, a novel multi-modal learning framework for person ReID. It enables synergistic fusion and cross-modal alignment of arbitrary modality combinations in a single model, with a unified encoding and multi-expert routing mechanism proposed. Extensive experiments verify the advancement and practicality of our ORBench. A range of models have been compared on it, and our proposed ReID5o gives the best performance.

1 Introduction

Person re-identification (ReID), as a fine-grained retrieval task, aims to search for person images of the target identity based on a given descriptive query [51]. Due to the broad applications of ReID in fields such as urban security, many researchers have conducted in-depth studies on it. It can be classified into single-modal retrieval [58, 45] and cross-modal retrieval [26, 46, 34]. The former involves the mutual search of RGB images, while the latter usually use queries from other modalities to retrieve RGB images. Considering the frequent lack of original RGB queries in real scenarios, the latter has increasingly drawn attention in recent years.

However, despite the considerable progress in this technology [61, 62, 24, 13], a challenging and practical problem, which aims to achieve effective retrieval with varying multi-modal queries and their combinations in the ReID model, remains largely unexplored in previous researches.

We propose that exploring such a challenging problem, termed Omni Multi-modal Person Re-identification (OM-ReID), is of great significance. On the one hand, due to the diversity of information acquisition, there is often a need for multi-modal queries in practical applications [50]. Users expect the model to be able to effectively process multi-modal information and provide comprehensive and accurate retrieval results. If the model can handle these varying modalities, it will significantly

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

boost its practicality. On the other hand, theoretically, different modalities provide abundant and complementary features [52, 5, 24]. For instance, RGB images are vulnerable to environmental light variations [46], and infrared as well as sketch images lack essential color information critical for ReID tasks. Fusing multi-modal features is akin to assembling an information jigsaw from diverse perspectives and dimensions. This enables a more comprehensive portrayal of person characteristics, thereby minimizing misjudgments stemming from single-modal limitations.

Considering the above aspects, and given that the related researches are constrained by datasets involving only few modalities and whose quality is not satisfactory [58, 46, 26, 34, 52], this paper makes the following contributions at the levels of both the dataset and the methodology.

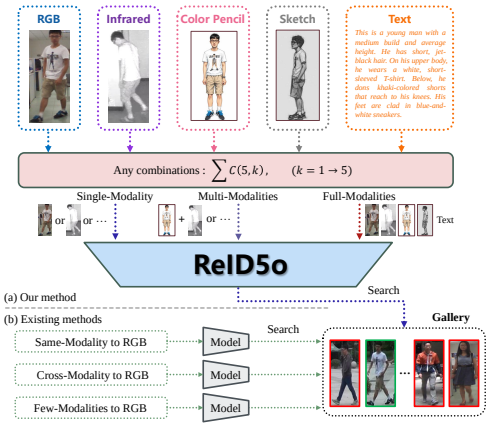

Figure 1: Our **ReID5o** can effectively conduct retrieval with any combinations of five modalities, adapting to various queries with different uncertain modalities in real scenarios. However, existing methods [46, 26, 52, 5, 24] are constrained to few modalities and are unable to achieve arbitrary retrieval with five modalities.

We collect the first high-quality person ReID dataset with 5 modalities, named **ORBench**. As shown in Fig. 2, for the same person, our dataset simultaneously contains RGB images, infrared images, colored pencil drawings, sketch drawings, and textual descriptions, providing a comprehensive portrayal of the person. Moreover, our dataset exhibits outstanding characteristics in terms of diversity. In addition to the cross-camera perspectives inherent in the real-captured data [46, 56], the contained paintings simultaneously reflects the frontal, dorsal, and lateral appearance features of the same person. Furthermore, the text component adopts an highly detailed and unrestricted descriptive style. By performing elaborate manual annotation and recurrent corrections, ORBench has significant superiority in terms of quality and diversity, compared to existing datasets. It contains 45,113 RGB images, 26,071 infrared images, 18,000 color pencil drawings, 18,000 sketches and 45,113 textual descriptions of 1,000 identities in total.

In addition, as shown in Fig. 1, to tackle the new challenge of retrieving with multiple modalities, we further propose a novel unified model named **ReID5o**, by which the portrayals of the same person across different modalities and their combinations are explicitly correlated to deeply mine the identity-invariance. ReID5o first encodes the inputs from diverse modalities into a shared embedding space via a multi-modal tokenizing assembler. Subsequently, a multi-expert router is devised to facilitate effective modality-specific representation learning within a unified feature extractor. Then, a feature mixture and a simple alignment strategy [16] are utilized to implement efficient fusion of multi-modal representations and their correlation learning, respectively. Compared with existing methods, ReID5o achieves the best performance and can serve as a simple but effective baseline method for our ORBench.

We conduct comprehensive experiments to verify the advancement and effectiveness of our dataset and method. It can be concluded that multi-modal combination querying brings significant performance improvements, and it serves as a more practical approach in real scenarios. As a research focusing on person ReID with arbitrary modalities and their combinations, we believe that our first high-quality dataset, ORBench, and our method, ReID5o, can inspire the subsequent development in this field.

2 ORBench Dataset

2.1 Construction Motivation

Multi-modal learning [44, 60, 22] has become a hot topic in the computer vision field. However, when directing to person ReID, researchers often confine themselves to datasets with few modalities and unsatisfactory quality [26, 34, 52, 5, 24]. Therefore, we construct ORBench, the first high-quality dataset with five modalities, driven by the following two major motivations.

Filling the modal gap and improving data quality. In the field of person ReID, previous datasets only covered few modalities, making it difficult to comprehensively depict the characteristics of persons. In addition, the quality of some multi-modal datasets [52, 5] is obviously poor, and only

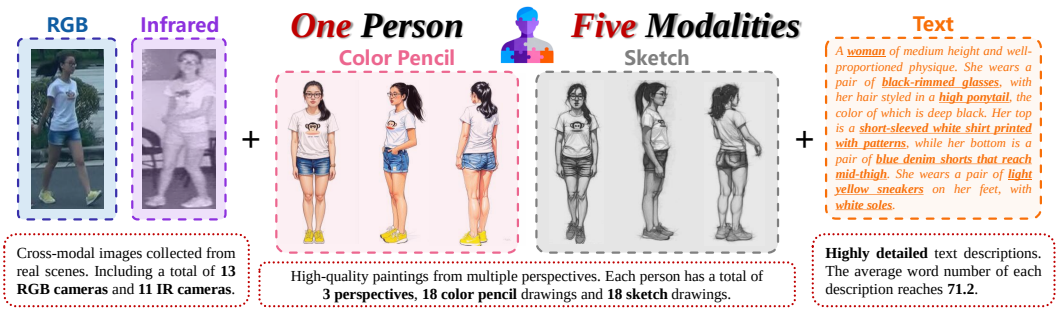

Figure 2: **The overview of our proposed ORBench dataset.** Our dataset is remarkable for containing rich, high-quality and diverse five-modal descriptive data for the same person, offering a comprehensive and in-depth resource for person ReID research.

rough generation strategies are utilized to obtain unrealistic multi-modal data. Therefore, to fill such a gap, we are determined to construct a new high-quality dataset that encompasses five modalities in a pioneering way, faithfully covering the characteristics of persons.

Promoting the development of multi-modal technologies and driving innovation. Multi-modal fusion and retrieval technologies [59, 42] have great potential in person ReID and there is an urgent need for their practical applications. However, due to the lack of suitable multi-modal datasets, their development is severely restricted. Therefore, to build an ideal platform for the development of this technology and drive innovation, we construct our ORBench, serving as the first high-quality multi-modal dataset in this field.

2.2 Dataset Collection

Based on the existing RGB-infrared datasets [46, 56], our ORBench dataset, with the cooperation of manual annotation and image generation experts, achieves the supplementation of additional modalities including color pencil drawings, sketch drawings, and textual descriptions. In addition, a manual correction and review mechanism has been incorporated to ensure the high quality.

For the RGB and infrared data, we manually selected persons with representative appearances from the existing datasets SYSU-MM01 [46] and LLCM [56]. That is to say, we filtered out those person images with poor imaging conditions and weak clothing representativeness. For example, if the appearance features of a person can hardly be seen in both the dual-modal images, the data of this person will be removed. Eventually, we obtained 45,113 RGBs and 26,071 IRs of 1,000 persons.

For the color pencil data, we utilized an online image generation model [1] as our painting agent, and incorporated extensive manual supervision to ensure the drawing quality. Specifically, for each person, we first selected some multi-perspective RGB images with high quality, and manually annotated them with detailed textual descriptions. Then, we required the agent to refer to these descriptions and images to paint color pencil drawings of the person's front, back, and side views. In addition, with many color pencil drawings painted, we manually selected six most satisfactory drawings for each view, to further ensure the quality. Eventually, we obtained 18,000 color pencil drawings.

For the sketch data, considering the maturity of the sketch synthesis technology, after conducting extensive investigations, we decided to utilize the Meitu application software [2] to directly convert the obtained colored pencil drawings into realistic sketch drawings. For the text data, we hired a few unique workers to be involved in the text annotation task and instructed them to describe the important characteristics for each RGB person image in detail. Through the above two procedures, we obtained 18,000 sketch drawings and 45,113 textual descriptions in total.

2.3 Dataset Statistics

Benefiting from our meticulous and labor-intensive manual annotation, ORBench is full of rich and high-quality multi-modal data. Compared with existing ReID datasets, our ORBench enjoys the following advantages:

Most Modalities to Date. Most previous multi-modal datasets [46, 26, 34] have merely two modalities, while for few tri-modal datasets [52, 5], the extra modalities are acquired via rough data synthesis strategies. As shown in Tab. 1, ORBench is the person ReID dataset boasting the greatest number of modalities so far, offering a more comprehensive and detailed portrayal of persons.

Table 1: **Comparisons with other ReID datasets.** ORBench is the first to have rich five-modal data and is the only dataset covering color pencil drawings. Only relatively high-quality datasets with manual annotations are listed, and low-quality synthetic datasets are excluded.

Datasets	Year	Modality	#Identities	#RGB Imgs	#IR Imgs	#CP Imgs	#SK Imgs	#Texts
Market1501 [58]	2015	R	1,501	32,668	-	-	-	-
CUHK-PEDES [26]	2017	R,T	13,003	40,206	-	-	-	80,412
MSMT17 [45]	2018	R	4,101	126,441	-	-	-	-
PKU-Sketch [34]	2018	R,S	200	400	-	-	200	-
SYSU-MM01 [46]	2020	R,I	491	30,071	15,792	-	-	-
ICFG-PEDES [8]	2021	R,T	4,102	54,522	-	-	-	54,522
TriReID [52]	2022	R,S,T	200	5,600	-	-	200	5,600
LLCM [56]	2023	R,I	1,064	25,626	21,141	-	-	-
MaSk1K [30]	2023	R,S	1,501	32,668	-	-	4,763	-
UFine6926 [63]	2024	R,T	6,926	26,206	-	-	-	52,412
ORBench	**2025**	**R,I,C,S,T**	**1,000**	**45,113**	**26,071**	**18,000**	**18,000**	**45,113**

High Quality. Existing tri-modal datasets often have poor quality due to rough data synthesis. For instance, Tri-CUHK-PEDES [6] just converts low-res RGB images to sketches, resulting in blurriness, missing parts, and distortion. TriReID [52] utilizes limited-capacity caption models for rough texts. However, our ORBench ensures high-quality data through careful, labor-intensive manual annotation. More examples are shown in the supplement.

Competitive Dataset Scale. As shown in Tab. 1, compared with existing datasets, the scale of our dataset is highly competitive, containing 45,113 RGB images, 26,071 infrared (IR) images, 18,000 color pencil (CP) drawings, 18,000 sketch (SK) drawing and 45,113 texts of 1,000 persons. Notably, our scale especially outshines those datasets with paintings. For example, compared with the previous largest sketch dataset, MaSk1K [30], we have approximately 4 times the number of sketches.

2.4 Evaluation Protocol

There are 1,000 valid identities in ORBench dataset. We have a fixed split using 600 identities for training and 400 identities for testing. During training, all multi-modal data for the 600 persons in the training set can be applied.

During the testing phase, the samples from the RGB modality are regarded as the gallery set, while the samples from the remaining four modalities and their combinations are treated as the corresponding query sets. Specifically, we design four search modes: single-modal search (MM-1), dual-modal search (MM-2), tri-modal search (MM-3), and quad-modal search (MM-4). For single-modal search, the samples of each modality form their respective query sets. For the remaining three multi-modal searches, we first select a primary modality. Then, for each sample of this modality, we randomly select samples of the same identity from the remaining modalities as supplementary ones to form a query tuple. Totally, there are 4 query sets for MM-1, 12 for MM-2, 12 for MM-3, and 4 for MM-4. The performance of each mode is measured as the average of retrieval results across its query sets.

As a common practice, we adopt Cumulative Matching Characteristic (CMC) and mean average precision (mAP) to evaluate the performance. Given a query, all gallery images are ranked according to their affinities with the query. A successful search is achieved if any image of the corresponding person is among the rank-k images.

2.5 Discussion on Realism and Idealized Assumptions.

It is important to acknowledge the idealized assumptions underpinning our dataset construction. We operated under the premise that comprehensive eyewitness information enables the creation of a complete textual description and, subsequently, a high-fidelity visual rendering of the target. Consequently, the identity consistency across modalities in ORBench is intentionally high.

In the taxonomy of [20], our sketches and color paintings are best characterized as "viewed sketches"—precise depictions derived from ample information—rather than incomplete or inaccurate "forensic sketches" that result from the fallibility of human memory. This design choice was necessitated by the feasibility of large-scale dataset construction. While ORBench provides a foundational, high-quality benchmark for studying multi-modal alignment in a controlled setting, it does not fully capture the noise and variance inherent in real-world forensic scenarios. We posit that this clean dataset serves as a crucial starting point and a powerful pre-training resource.

3 ReID5o: Method

In this section, we propose a unified multi-modal learning framework for person ReID, named ReID5o. We utilize a multi-modal tokenizing assembler to encode the multi-modal inputs into a shared embedding space. Then, a unified encoder is used to extract modality-shared features, and a multi-expert router is designed to facilitate modality-specific representation learning. Also, a feature mixture fuses the multi-modal features. Finally, we use the SDM [16] and identity classification losses to implement cross-modal alignment. The whole schematic is shown in Fig. 3.

3.1 Multi-modal Tokenizing Assembler

To encode varying-modal data inputs into a shared embedding space, we design a straightforward multi-modal tokenizing assembler. This assembler is composed of five sub-tokenizers, which are respectively used to tokenize the RGB, IR, CP, SK, and text data. In addition, it will also provide discrete control signals for the subsequent multi-expert router according to the modality difference.

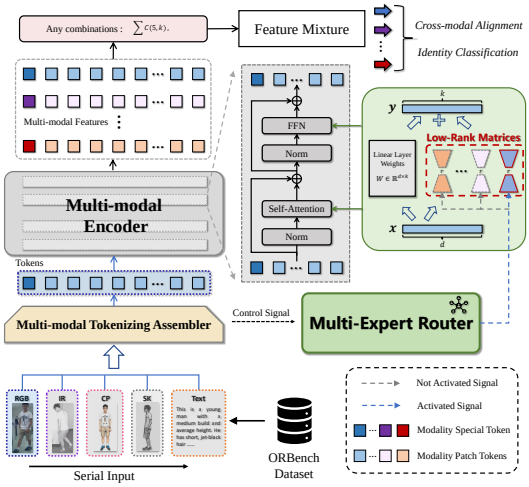

Figure 3: **The schematic of our proposed ReID5o framework.** As the unified multi-modal encoder extracts the modality-shared features, our specially designed multi-expert router can effectively promote the modality-specific representation learning.

Specifically, for these four types of visual inputs, we use non-shared visual tokenizers [38] with the same structure to encode them respectively. This visual tokenizer consists of a simple convolutional layer and an LN layer. In addition, we add an additional class token and positional embeddings as a common practice. For text input, following standard convention [38], we first add special flags at the beginning and end of the text respectively. Then, we employ a vocabulary mapping layer and positional embeddings to project the input text into a high-dimensional embedding space. Formally, for the input data $\mathbf{X}^{mod}$, the process of obtaining its corresponding embeddings can be formulated as follows.

$$\mathbf{E}^{mod}, c^{mod} = \gamma^{mod}(\mathbf{X}^{mod}), \tag{1}$$

where $\mathbf{E}^{mod} \in \mathbb{R}^{n \times D^e}$, representing that the input $\mathbf{X}^{mod}$ is encoded into n tokens with dimension D^e. c^{mod} represents a binary signal for modality activation, which is used to control the subsequent multi-expert router. Here, γ^{mod} represents the tokenizer of each modality and $mod \in \{R, I, C, S, T\}$, which respectively represent the RGB, Infrared, Color pencil, Sketch, and Text modalities.

3.2 Multi-Expert Router

We utilize a unified multi-modal encoder to extract modality-shared features, which inherits the rich multi-modal pre-trained knowledge from CLIP [38] to ensure a good initial starting point for subsequent training. As a Transformer, it contains multiple linear layers. For a certain linear layer $W \in \mathbb{R}^{d \times k}$, its function can be reduced as $\mathbf{y} = W(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^k$ represent the input and output feature for this layer, respectively. It can be seen that if the same linear layer is used directly for different modalities, it is hard to extract modality-specific representations.

Therefore, we propose a simple yet effective multi-expert router to explicitly promote the modality-specific representation learning based on the modality categories. Specifically, we incorporate multiple modality-specific low-rank matrices [15] into each linear layer as efficient feature extraction experts. In the absence of data input, these experts stay inactive. Once the assembler encodes data from a particular modality, the router will receive the associated control signal and activates the corresponding expert, enabling modality-specific parameter updates. Formally, for the input feature $\mathbf{x}^{mod}$ of a certain linear layer W, its output feature can be formulated as follows.

$$\begin{aligned} \mathbf{y}^{mod} &= W\mathbf{x}^{mod} + c^{mod}\Delta W^{mod}\mathbf{x}^{mod} \\ &= W\mathbf{x}^{mod} + c^{mod}B^{mod}A^{mod}\mathbf{x}^{mod}, \end{aligned} \tag{2}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and $r \ll \min(d, k)$. c^{mod} is a binary control signal from the assembler, indicating whether the expert ΔW^{mod} is activated. We can observe that for any modal input $\mathbf{x}^{mod}$, while W is used to extract the modal shared features, ΔW^{mod} will be employed to extract the modal unique features. In this way, we can effectively uncover the modal invariance and discrimination.

3.3 Feature Mixture and Learning Strategy

Now, for the input data $\mathbf{X}^{mod}$, through the aforementioned tokenization and expert routing mechanisms, we can obtain its output features $\mathbf{Z}^{mod} \in \mathbb{R}^{n\times D}$. Then, to achieve full complementary information interaction between different modalities, we employ an efficient feature mixture to fuse the features of varying modality combinations. The feature mixture consists of a multi-head self attention (MSA) layer, 1-layer transformer block and a multi-layer perception (MLP). Then, the combinatorial traversal is performed on the entire set of multimodal query features $\{Z^I, Z^C, Z^S, Z^T\}$. We generate all possible combinations from single-modal to quad-modal. These combinations are then concatenated and processed by the Feature Mixture module.

Taking the combination of $\mathbf{Z}^I$ and $\mathbf{Z}^T$ as an example, the feature fusing process can be formulated as follows.

$$\mathbf{z}^{I,T} = MLP(TF(MSA(LN(CAT(\mathbf{Z}^I, \mathbf{Z}^T))))), \tag{3}$$

where $\mathbf{z}^{I,T} \in \mathbb{R}^D$, indicating the IR and Text fused representation. $CAT(\cdot)$, $LN(\cdot)$ and $TF(\cdot)$ denote the concatenation, layer normalization and transformer block, respectively. The last layer of the MLP is an average pooling layer.

In addition, we directly regard the output of special token as the single-modal representation $\mathbf{z}^{mod}$. As a simple learning strategy, we adopt the commonly used SDM loss [16] and identity classification (IC) loss to supervise the learning process. Considering that RGB modality is usually utilized as the gallery set and contains more original and complete information, we take it as the core target for feature alignment. We explicitly conduct the alignment of diverse modality combinations. To realize this, the overall learning objective can be written as follows.

$$\mathcal{L} = \sum_{R\notin c_i} SDM(\mathbf{z}^R, \mathbf{z}^{c_i}) + \alpha \sum_{c_i} IC(\mathbf{z}^{c_i}), \tag{4}$$

where $c_i \in \{A|A \subseteq S, A \neq \varnothing\}$ and $S = \{R, I, C, S, T\}$. α is a manually set fixed hyper-parameter, which is used to control the importance of the identity classification loss.

4 Experiments

There is no existing person ReID methods specifically designed to achieve the retrieval of any combinations of these five modalities. We investigate a wide range of possible solutions based on some leading general multi-modal models [38, 54, 11] and cross-modal person ReID models [5, 61, 24, 16, 36], and compare these solutions with our proposed method. We also conduct experiments to verify the data quality of our proposed ORBench dataset. Extensive experiments and comparisons demonstrate the superiority of our ORBench dataset and ReID5o method.

4.1 Experimental Settings

Datasets. The main experiments are conducted on our proposed multi-modal ORBench dataset. This dataset is divided into a training set with 600 persons and a testing set with 400 persons. If there is no special statement, the performance is evaluated with the four proposed search modes. In addition, CUHK-PEDES [26], ICFG-PEDES [8] and UFine6926 [63] are used for comparisons.

Implementation Details. We employ a pre-trained multi-modal encoder [38], *i.e.*, CLIP-B/16, as the unified encoder, which is pre-trained on billions of image-text pairs with contrastive learning. In the actual implementation, to fully leverage pre-trained knowledge, we employ an independent text encoder for the textual modality, while adopting the aforementioned strategy for other visual modalities. For the multi-modal tokenizing assembler (MTA), the tokenizers for all modalities are initialized simultaneously with the pre-trained parameters in CLIP [38], and will not be shared during the subsequent training. For the multi-expert router (MER), each modality will be assigned an independent expert, and the low-rank matrices [15] of a certain expert will be installed on each linear layer of the encoder, where r is set to 4. For each layer of the feature mixture (FM), the hidden size and number of heads are set to 512 and 8. The hyper-parameter α for the ID loss is set to 1.0. During

training, all images are uniformly resized to 384×128 and the maximum length of the textual tokens is set to 77. Also, random erasing, horizontally flipping and crop with padding are employed for image augmentation. Random masking and replacement is employed for text augmentation. Our ReID5o is trained with Adam [19] for 60 epochs with an initial learning rate $1e^{-5}$. We spend 5 warm-up epochs linearly increasing the learning rate from $1e^{-6}$ to $1e^{-5}$. For the random-initialized experts and feature mixture, the initial learning rate is set to $5e^{-5}$. We use a single A100 80GB GPU.

4.2 Ablation Study and Analysis

Effectiveness of Each Component. To verify the contribution of each component in ReID5o, we conduct an ablation experiment on our ORBench. The results are reported in Tab. 2. In the baseline No.0, all visual modalities use the same shared visual tokenizer, and there is no expert router to extract modality-specific features. In addition, simple average pooling is employed for multi-modal feature fusion. As evident from Tab. 2, compared with the baseine method, each introduced component proves crucial for the overall performance of our ReID5o model, and combining all of them leads to the best performance.

Table 2: **Ablation study on each component of ReID5o.** We report the mAP results for each search mode, with the best in bold.

No.	Components			ORBench			
	MTA	MER	FM	MM-1	MM-2	MM-3	MM-4
0				49.24	66.36	74.21	78.42
1	✓			51.85	67.43	75.43	79.81
2	✓	✓		56.99	73.58	81.42	85.14
3	✓		✓	52.80	69.23	77.70	81.87
4	✓	✓	✓	**58.09**	**75.26**	**82.83**	**86.35**

Analysis of the Multi-Expert Router. To analyze the effectiveness and efficiency of the multi-expert router, we conduct ablation experiments on the setting of r in the low-rank matrices. In Tab. 3, we present the performance of each setting, as well as the introduced additional parameters and FLOPs compared to the baseline method (w.t. MER). We also analyze a simple and straightforward approach, that is, utilizing unique encoders to encode each modality separately (w. UE). We can observe that UE fails to bring significant improvement while introducing extremely large additional parameters. However, when $r = 4$, our multi-expert router achieves the best performance, and maintains a relatively low additional computational cost.

Table 3: **Different settings of the multi-expert router.** We report the mAP results for each search mode, with the best in bold.

Method	Params. (M)	FLOPs. (G)	ORBench			
			MM-1	MM-2	MM-3	MM-4
w.t MER	-	-	52.80	69.23	77.70	81.87
$r = 4$	2.36	3.64	**58.09**	**75.26**	**82.83**	**86.35**
$r = 8$	4.72	7.29	57.59	74.59	82.33	85.71
$r = 16$	9.44	14.57	57.81	74.90	82.59	86.21
$r = 32$	18.87	29.14	57.56	74.89	82.39	85.72
w. UE	256.4	-	52.42	69.13	76.36	80.15

Analysis of Multi-modal Complementarity. Taking the text modality as the primary modality, we explore the impact of the introduction of different modalities on its retrieval results, so as to investigate the complementarity among multi-modal data. As shown in Tab. 4, the introduction of any additional modality into the query can directly enhance the retrieval performance. Among them, the promoting effect of the C modality is the most obvious. Compared with the T modality, when using T and C modalities, the mAP has increased by 26.26%. In addition, regarding the C modality, the complementarity effect of the I modality is significantly better than that of the S modality. Compared with T + C, T + I + C can bring about a 5.61% increase in mAP, while T + C +S only brings a 1.91% increase.

Table 4: **Queries with different modality combinations.** The text modality is the primary modality. The best results are in bold.

Modality	ORBench			
	Rank-1	Rank-5	Rank-10	mAP
T	63.15	77.94	82.92	54.88
T+I	81.19	92.19	95.19	70.77
T+C	91.01	96.83	98.04	81.14
T+S	84.71	93.39	95.55	74.01
T+I+C	95.85	99.25	99.68	86.75
T+I+S	92.85	98.01	98.94	81.98
T+C+S	93.19	97.80	98.83	83.05
T+I+C+S	**96.79**	**99.37**	**99.78**	**87.46**

Different Learning Strategies. The different implementations of the loss functions for the learning strategy are shown in Tab. 5. We have implemented one instance-level loss function [38] and three identity-level loss functions [17, 53, 16] for the cross-modal alignment. As we can see, compared with other losses, the SDM loss [16] leads to the best performance. In addition, introducing the identity classification (IC) loss will further improve the model's performance to 58.09% mAP for the single-modal search mode.

Table 5: **Ablation study on different learning strategies.** We report the mAP results for each search mode, with the best in bold.

Strategy	ORBench			
	MM-1	MM-2	MM-3	MM-4
ITC [38]	53.44	67.35	74.62	78.25
SupITC [17]	48.16	64.84	73.03	77.34
CMPM [53]	48.02	67.27	75.41	79.51
SDM [16]	57.65	74.51	82.14	85.60
SDM+IC [16]	**58.09**	**75.26**	**82.83**	**86.35**

Table 6: **Performance comparison with existing methods on ORBench.** These methods can be divided into cross-modal methods and multi-modal methods. The former performs alignment between two modalities, while the latter performs alignment among multiple modalities. Each method has been fine-tuned on our ORBench. The results* are the outcomes of our reproduction efforts.

Method	Single-modal Search			Dual-modal Search			Tri-modal Search			Quad-modal Search		
	R@1	R@10	mAP	R@1	R@10	mAP	R@1	R@10	mAP	R@1	R@10	mAP
Cross-modal Models												
CLIP [38]	59.84	77.88	48.73	80.31	92.80	65.79	87.67	96.78	73.17	90.71	98.25	76.69
PLIP [61]	60.57	78.24	48.79	80.78	92.82	65.63	88.14	96.98	74.60	90.95	98.12	77.87
IRRA [16]	61.17	79.30	51.42	82.50	94.17	68.83	89.70	97.39	75.86	92.65	98.49	79.15
RDE [36]	60.91	79.20	48.82	82.36	93.95	67.52	89.94	97.50	75.27	92.75	98.63	78.64
Multi-modal Models												
Meta-TF* [54]	6.25	19.46	4.24	10.81	31.26	6.68	14.82	40.45	8.65	18.32	47.66	10.45
ImageBind* [11]	60.18	78.32	48.61	81.78	93.18	67.02	88.73	97.23	74.20	91.05	98.36	77.36
UNIReID [5]	59.49	77.59	47.98	80.91	93.16	66.83	88.71	97.08	74.86	91.34	98.02	78.12
AIO* [24]	50.44	70.88	40.14	58.82	81.67	49.34	69.72	89.34	58.58	75.33	92.52	63.48
ReID5o	**68.12**	**86.36**	**58.09**	**86.15**	**96.76**	**75.26**	**93.53**	**99.15**	**82.83**	**96.25**	**99.70**	**86.35**

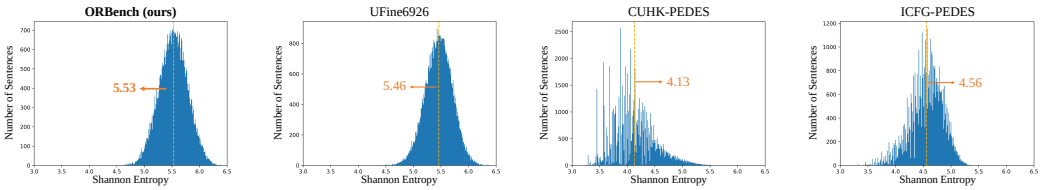

Figure 4: Comparisons of the Shannon entropy per textual description with existing person ReID datasets containing the text modality. The average entropy of our dataset reaches 5.53, representing the highest level of textual information richness among current datasets.

4.3 Comparison with Existing Methods

As shown in Tab. 6, we compared the performance of extensive possible methods on our ORBench. These methods can be divided into multi-modal methods in the general domain and cross-modal methods in the field of person ReID. The former usually aligns data from multiple modalities into the same feature space, such as images, audio, videos, and texts. For the former methods, we treat each visual modality in ORBench as a unique image modality for them. That is, we will utilize four different image tokenizers and then perform training on ORBench. For the latter, since most methods only achieve image-text alignment, for this type of methods, we directly regard the visual data of the four modalities as the image modality of these methods. However, for the AIO method [24] that achieves alignment among RGB, IR, SK, and Text, we additionally add a CP tokenizer to its framework and conduct fine-tuning on our ORBench using its default settings. In addition, since none of these methods achieve explicit alignment between any modality combinations and RGB, we implement an equivalent multi-modal search mode by superimposing the similarity lists of queries and gallery for each modality. We can observe that, compared with the existing methods, our ReID5o achieves the best performance in all four search modes. These results validate the effectiveness of our method in leveraging arbitrary modality combinations for person ReID, and it can serve as the first strong baseline method for our ORBench.

4.4 Dataset Quality Assessment

Considering that we additionally annotate data of three modalities to the existing RGB-infrared datasets, and the sketch data is converted from the color pencil data, we conduct experiments to explore the data quality of the text modality and the color pencil modality. For the text modality, we use Shannon entropy to measure the amount of information contained in the texts. Specifically, we can treat each word as an event, calcu-

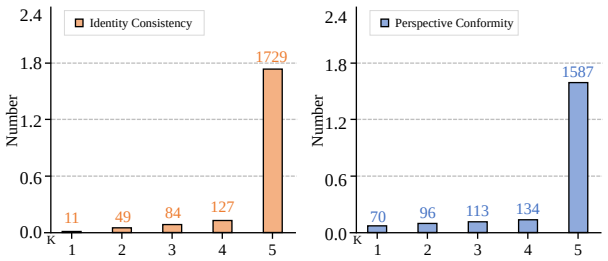

Figure 5: Public evaluation to assess the identity consistency and perspective comformity of the color pencil drawings in ORBench.

late the probability of each word occurring, and finally calculate the entropy of the entire text. We conduct a statistical comparison of the Shannon entropy per text in our ORBench with those in other datasets [63, 26, 8], as illustrated in Fig 4. We can see that our ORBench has richer textual information than other datasets, verifying the high quality of the text modality. For the color pencil modality, we adopt the method of public evaluation to assess the identity consistency and perspective conformity of the color pencil drawings. Specifically, we randomly pick 2,000 samples from the color pencil drawings and split them into 20 groups. Then, 20 evaluators are asked to objectively score these samples on identity consistency with the original RGB images and conformity of drawing perspectives. Higher scores mean better evaluations. Detailed evaluation processes can be found in the supplement. As shown in Fig 5, the overall quality of the color pencil modality is quite high.

4.5 Dataset Generalization Evaluation

We conduct experiments to compare the generalization performance of various models trained on different existing multi-modal datasets to other manually annotated datasets. The subtasks are divided into text-based retrieval and sketch-based retrieval. For the former, we use CUHK-PEDES [26] as the target dataset. For the latter, we used PKU-Sketch [34] as the target dataset. To ensure a fair comparison, we employ the same number of training samples for each training dataset, i.e., 40,000 training samples are randomly selected from each dataset. As the results shown in Tab. 7, for all models, compared with other multi-modal datasets, training on our ORBench dataset can achieve better generalization performance. This strongly demonstrates the high quality of our dataset and its generalizability to real-world application scenarios. Researchers can have full confidence in using our high-quality ORBench dataset.

Table 7: **Evaluation comparison on dataset generalization.** "T", "C","I","P","A","O" denote Tri-CUHK-PEDES [5], CUHK-PEDES [26], ICFG-PEDES [8], PKU-Sketch [34], AIO [24] and our ORBench, respectively. The target datasets are all manually annotated real-world datasets. We report the rank-1 results

Method	$O \rightarrow I$	$T \rightarrow I$	$C \rightarrow I$	$O \rightarrow P$	$T \rightarrow P$	$A \rightarrow P$
Cross-modal Models						
PLIP [61]	**52.61**	49.23	49.23	-	-	-
IRRA [16]	**38.52**	32.57	32.57	-	-	-
SketchTrans [6]	-	-	-	**70.30**	66.30	62.70
Multi-modal Models						
UNIReID [5]	**36.62**	33.64	-	**77.80**	69.80	65.40
ImageBind [11]	**38.58**	35.33	-	**75.40**	68.60	61.80
ReID5o	**43.27**	38.65	-	**80.20**	71.80	67.00

4.6 Visualization

To demonstrate the role of multi-modal queries more clearly, we visualize retrieval results on ORBench using diverse modality combinations as queries, ranging from single-modal to multi-modal inputs. As shown in Fig. 6, the addition of supplementary modalities—such as integrating text with infrared or sketch data—significantly improves retrieval accuracy, with top-ranked results becoming more relevant. This highlights that multi-modal queries, by capturing complementary information (e.g., semantic descriptions and visual details under varied conditions), form a more complete depiction of search targets. For practical applications, this implies that leveraging multi-modal descriptive queries—rather than relying on single modalities—can enhance retrieval precision, enabling more accurate identification of targets in complex scenarios.

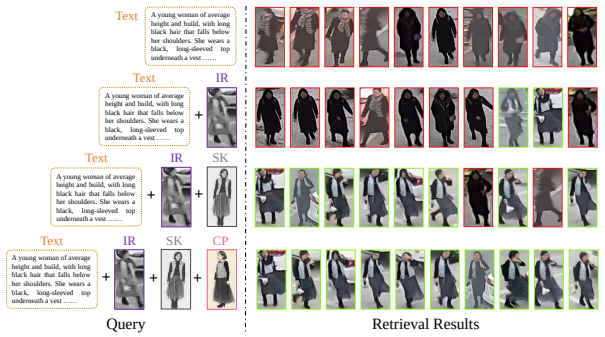

Figure 6: Visualization of the top-10 ranking lists for four queries with different modalities on ORBench.

5 Related Work

Multi-modal Learning. The goal of multimodal learning is to fully integrate heterogeneous information from multiple modalities to obtain more representative representations of features [29]. With the popularity of Transformers, researchers have proposed unified architectures to process multimodal inputs end-to-end [18, 9, 25]. Although most models [3, 43, 35] do not face the problem

of missing modalities, this problem frequently occurs in real-world applications. Recent works [4, 33, 40, 11, 32] have introduced various strategies to mitigate this impact. Zhang *et al.*[57] align information from different modalities by mapping them into a shared feature space. Lee *et al.*[21] employ missing-modality-aware prompt learning to deal with modality absence. Lian *et al.*[28] propose a graph-completion network to jointly optimize classification and reconstruction tasks in an end-to-end framework.

Cross-modal Person ReID. Compared to single-modal person ReID tasks [23, 31, 55, 7, 14], cross-modal ReID tasks [5, 46] typically use RGB images as retrieval targets while employing text descriptions [26, 39, 10, 48, 27], infrared images [37, 49, 47], attributes [41], or sketches [12, 6] as queries to simulate real-world scenarios with uncertain query modalities. Li *et al.*[26] pioneered text-based person ReID. Wang *et al.*[41] proposed a joint attribute-identity method to learn attribute-semantic and identity-discriminative features. Qiu *et al.*[37] introduced a high-order structure-based intermediate feature learning network to fully exploit structural information in IR features. Lin *et al.*[30] designed mechanisms to mitigate the impact of sketches painter's subjectivity on ReID performance. Subsequent work combines some modalities to address practical constraints in obtaining target modalities and using complementary information. Chen *et al.*[5] fused features of RGB, sketch, and text modalities for modality-agnostic retrieval. Li *et al.*[24] developed a unified framework handling RGB, infrared, sketch, and text modalities. However, existing works are still constraint to scenarios with few modalities and overlook the rich cross-domain information contained in different modality combinations.

6 Limitation

While the proposed ReID5o framework and the ORBench dataset establish a strong foundation for OM-ReID research, we acknowledge several limitations that point to future directions. First and foremost, ORBench is constructed under an idealized data generation assumption. The high fidelity of its sketch and color painting queries, which resemble "viewed sketches" rather than imperfect "forensic sketches," ensures tight cross-modal alignment but does not fully encapsulate the noise and variance (e.g., inaccuracies, omissions) prevalent in real-world scenarios. Consequently, while our model demonstrates robust performance on this clean benchmark and shows promising generalization in cross-dataset tests, its performance in truly noisy, in-the-wild settings requires further investigation. Secondly, our work is currently limited to a predefined set of five modalities. Other practically relevant modalities, such as thermal infrared or radar, are not included, and our framework is not yet designed for incremental learning of new modalities. Finally, the study focuses on a closed-set identification task, leaving open-world scenarios with unseen identities as an open challenge. We believe addressing these limitations—by introducing realistic noise, extending modality coverage, and moving towards open-set recognition—will be crucial steps for transitioning OM-ReID from a controlled benchmark to robust real-world applications.

7 Conclusion

This paper investigated the problem of Omni Multi-modal Person Re-identification (OM-ReID), where the goal is to retrieve a person using a query from any single modality or their arbitrary combinations. To support this research, we constructed ORBench, the first comprehensive dataset for OM-ReID, featuring 1,000 identities across five modalities (RGB, infrared, sketch, color pencil, and text). We also developed ReID5o, a flexible framework that enables unified encoding and effective retrieval for any modality input. Our experiments demonstrate that leveraging multi-modal queries substantially improves retrieval performance compared to single-modality approaches, underscoring their practical value. We believe that the ORBench dataset and the ReID5o framework establish a solid foundation and a strong baseline for future research in this emerging field.

8 Acknowledgments.

This work was supported by the National Natural Science Foundation of China No.62176097, and the Hubei Provincial Natural Science Foundation of China No.2022CFA055. We would like to sincerely thank the HPC Platform of Huazhong University of Science and Technology for providing computational resources.

References

[1] Doubao, text-to-image model, general intelligent drawing. https://www.volcengine.com/docs/6791/1354010.

[2] Meitu, structured sketch. https://ai.meitu.com/.

[3] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation With Multimodal Transformers. In *CVPR*, 2022.

[4] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep Adversarial Learning for Multi-Modality Missing Data Completion. In *KDD*, 2018.

[5] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards Modality-Agnostic Person Re-Identification With Descriptive Query. In *CVPR*, 2023.

[6] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch Transformer: Asymmetrical Disentanglement Learning from Dynamic Synthesis. In *ACMMM*, 2022.

[7] Jiyuan Chen, Changxin Gao, Li Sun, and Nong Sang. Ccsd: cross-camera self-distillation for unsupervised person re-identification. *Visual Intelligence*, 1(1):27, 2023.

[8] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020.

[10] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search, 2021.

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023.

[12] Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 2020.

[13] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *CVPR*, pages 17521–17531, 2024.

[14] Jiahao Hong, Jialong Zuo, Chuchu Han, Ruochen Zheng, Ming Tian, Changxin Gao, and Nong Sang. Spatial cascaded clustering and weighted memory for unsupervised person re-identification. *Image and Vision Computing*, 156:105478, 2025.

[15] 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.

[16] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2023.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, 2021.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Brendan Klare, Zhifeng Li, and Anil K Jain. Matching forensic sketches to mug shot photos. *TPAMI*, 33(3):639–646, 2010.

[21] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, 2023.

[22] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

[23] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal Transformer with Conv-Patchify for Person Re-identification. In *ACMMM*, 2022.

[24] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *CVPR*, pages 17459–17469, 2024.

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 2021.

[26] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017.

[27] Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels. In *AAAI*, 2023.

[28] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *TPAMI*, 2023.

[29] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, 2023.

[30] Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond Domain Gap: Exploiting Subjectivity in Sketch-Based Person Retrieval. In *ACMMM*, 2023.

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and A Strong Baseline for Deep Person Re-identification. In *CVPRW*, 2019.

[32] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI*, 2021.

[33] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-Image-Specific Learning for Diagnosis-Oriented Neuroimage Synthesis With Incomplete Multi-Modality Data. *TPAMI*, 2021.

[34] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *ACM MM*, pages 609–617, 2018.

[35] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric Multimodal Contrastive Representation Learning. In *ICML*, 2022.

[36] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, pages 27197–27206, 2024.

[37] Liuxiang Qiu, Si Chen, Yan Yan, Jing-Hao Xue, Da-Han Wang, and Shunzhi Zhu. High-Order Structure Based Middle-Feature Learning for Visible-Infrared Person Re-Identification, 2024.

[38] 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 *ICML*, pages 8748–8763, 2021.

[39] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In *ACMMM*, 2022.

[40] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling. In *CVPR*, 2023.

[41] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification. In *CVPR*, 2018.

[42] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[43] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: BERT Pretraining of Video Transformers. In *CVPR*, 2022.

[44] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023.

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[46] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[47] Bin Yang, Jun Chen, Xianzheng Ma, and Mang Ye. Translation, Association and Augmentation: Learning Cross-Modality Re-Identification From Single-Modality Annotation. *TIP*, 2023.

[48] Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. In *ACMMM*, 2023.

[49] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic Tri-Level Relation Mining With Attentive Graph for Visible Infrared Re-Identification. *TIFS*, 2022.

[50] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *IJCV*, pages 1–31, 2024.

[51] 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 TPAMI*, 44(6):2872–2893, 2021.

[52] Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *ICMR*, pages 63–71, 2022.

[53] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018.

[54] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

[55] Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Contrastive Learning. In *NeurIPS*, 2021.

[56] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023.

[57] Yunhua Zhang, Hazel Doughty, and Cees Snoek. Learning Unseen Modality Interaction. In *NeurIPS*, 2023.

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[59] Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE TKDE*, 36(1):239–260, 2023.

[60] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. *IEEE TPAMI*, 2024.

[61] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. Plip: Language-image pre-training for person representation learning. In *NeurIPS*, 2024.

[62] Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. In *NeurIPS*, 2024.

[63] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *CVPR*, pages 22010–22019, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction of our paper accurately reflect our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We conduct an in-depth discussion of the limitations of our work, which can be found in Sec. 6 of the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work focuses on computer vision applications, not including theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have fully provided all the details of our proposed methods in the main text and supplementary materials, to guarantee the reproducibility of our paper's main experimental results. Meanwhile, we will open-source our work after our paper is accepted. These details and open-source initiatives will ensure the reproducibility of our work and make a significant contribution to the entire community.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To ensure the proper use of the technology, we would not provide open access to our data and code during the paper submission process.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper specifies all the training and test details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to comparably expensive computational cost, the error bars are not reported in the paper like nearly all related work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our paper provides sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Our paper discusses both potential positive societal impacts and negative societal impacts of the work in the supplementary materials.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our paper describes safeguards that have been put in place for responsible release of data and models in the appendix.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We communicated with the dataset creators in writing, obtained permission to use and extend their data, and strictly complied with their licenses. Additionally, we have specified the research license (CC BY-NC-SA 4.0) on OpenReview.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our work does introduce new assets by extending the existing dataset with new annotations. We can confirm that all relevant guidelines have been followed.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our external manual annotation work did involve human participation. We confirm that all annotators were fairly compensated, and our research strictly adhered to all relevant ethical guidelines.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Our external manual annotation work did involve human participation. We confirm that all annotators were fairly compensated, and our research strictly adhered to all relevant ethical guidelines.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Appendix

A.1 More Specific Samples

We have shown and compared some specific typical samples of our proposed ORBench dataset and two existing tri-modal datasets in Fig 7. As we can observe, due to the adoption of some rough data synthesis strategies to achieve the supplementation of additional modalities, the quality of existing tri-modal datasets is extremely low. For Tri-CUHK-PEDES [5], since it is based on the existing CUHK-PEDES [26] and directly stylizes its RGB image data into a sketch style, we can notice that the sketch images in this dataset have a very low resolution and are significantly different in style from actual sketch drawings. For AIO [24], it is based on the existing SYNTH-PEDES [61] and directly uses some data augmentation methods to obtain the infrared and sketch modality. We can observe that the infrared and sketch images in this dataset are very rough and coarse-grained, failing to effectively cover the appearance characteristics of persons. However, for our proposed ORBench dataset, benefiting from meticulous manual annotation and repeated corrections, it possesses very high-quality five-modal data. The color pencil drawings and sketch drawings can maintain a high degree of identity consistency with the persons, and they have diverse painting angles and outstanding detail presentation capabilities. The text descriptions are extremely detailed and can be on a par with the existing finest text-based person ReID dataset UFine6926 [63]. Our ORBench, with its excellence in aspects such as the number of modalities, data diversity, and data quality, can make a very prominent contribution to the person ReID community and profoundly inspire subsequent researches.

A.2 Public Evaluation Procedure

In order to further evaluate the quality of our color pencil drawing data, we adopt the method of public evaluation to assess the identity consistency and perspective conformity. We randomly pick 2,000 samples from the color pencil drawings and split them into 20 groups. Then, 20 evaluators are asked to objectively score these samples on identity consistency with the original RGB images and conformity of drawing perspectives.

Regarding its identity consistency, the scores can be divided into five levels. The higher the score, the better the consistency. The meanings represented by each score are as follows:

Score 1: There are two or more features in the color pencil drawing that are inconsistent with the original RGB image, such as the color of the shoes, whether wearing a hat or not, etc.

Score 2: There is one feature in the color pencil drawing that is inconsistent with the original RGB image.

Score 3: There are no features in the color pencil drawing that are inconsistent with the original RGB image, but there are relatively obvious painting errors, such as the orientation of the feet, etc.

Score 4: There are no features in the color pencil drawing that are inconsistent with the original RGB image, but there are few less obvious painting errors, such as the unnatural twisting of the hands, etc.

Score 5: There are no features in the color pencil drawing that are obviously inconsistent with the original RGB image, and there are no painting errors at all.

Regarding its perspective conformity, the scores can be divided into five levels. The higher the score, the better the conformity. The meanings represented by each score are as follows:

Score 1: The perspective does not match. The features of a certain perspective cannot be reflected in the color pencil drawing at all. For example, the label of the color pencil drawing is "back view", but the content of the drawing is actually a front view.

Score 2: The perspective partially does not match. The features of a certain perspective cannot be partially reflected in the color pencil drawing. For example, the label of the color pencil drawing is "front view", but the content of the drawing is a side view that shows only a small part of the front features.

Score 3: The perspective basically matches. The features of a certain perspective can be basically reflected in the color pencil drawing. For example, the label of the color pencil drawing is "front view", but the content of the drawing is a side view that basically shows the front features.

(a) A typical synthetic sample in Tri-CUHK.

(b) A typical synthetic sample in AIO.

(c) Some typical samples in **our ORBench**.

Figure 7: A visual overview of some typical samples in our proposed **ORBench** dataset and two existing multi-modal datasets [5, 24]. As we can see, the quality of existing datasets is quite rough. However, our ORBench not only encompasses a greater number of modalities but also demonstrates significant superiority in the data richness, quality and diversity.

Score 4: The perspective mostly matches. The features of a certain perspective can be mostly reflected in the color pencil drawing. For example, the label of the color pencil drawing is "front view", but the content of the drawing is a side view that mostly shows the front features (such as turning slightly sideways about 10 degrees).

Score 5: The perspective completely matches. The features of a certain perspective can be completely reflected in the color pencil drawing. For example, if the label of the color pencil drawing is "front view", then the content of the drawing is a completely frontal image of the person.

A.3 More Experimental Results

We have presented the retrieval results of our ReID5o in Tab 8, for 32 different query sets on our ORBench dataset. It can be seen that the supplementation of each additional modality can effectively improve the retrieval performance. This inspires us that when actually deploying the ReID technology in real-world scenarios, we should focus on the development in the direction of multi-modal retrieval.

A.4 Controlled Release

To address the privacy concerns associated with the application of person ReID technology, we will implement a controlled release of our dataset and code, thereby preventing privacy violations and ensuring information security. We will require that users adhere to usage guidelines and restrictions to access our dataset and code. We have drafted the following regulations that must be adhered to, which will be refined and elaborated in subsequent releases:

Table 8: The performance of our ReID5o via the queries with different modality combinations. The fIst modality of each group is regarded as the primary modality. The best results are in bold.

Modality	ORBench				
	Rank-1	Rank-5	Rank-10	mAP	mINP
I	52.288	69.170	75.674	43.644	16.694
C	87.222	95.139	96.903	75.093	37.599
S	69.833	85.389	89.944	58.719	23.935
T	63.147	77.939	82.920	54.884	19.053
I+C	93.909	98.475	99.348	84.082	43.937
C+I	93.514	98.292	99.347	82.962	44.468
I+S	84.403	94.101	96.671	72.757	32.008
S+I	84.264	93.931	96.694	72.540	33.701
I+T	79.530	91.453	94.609	68.994	30.017
T+I	81.191	92.188	95.186	70.773	29.020
C+S	86.875	95.250	97.250	75.156	37.562
S+C	86.319	94.722	96.986	75.107	37.349
C+T	87.694	95.292	97.278	76.687	38.634
T+C	91.014	96.831	98.039	81.141	38.563
S+T	80.431	91.319	94.153	68.896	30.848
T+S	84.709	93.391	95.546	74.005	30.868
I+C+S	95.127	99.070	99.693	85.028	44.344
C+I+S	94.778	98.875	99.597	83.941	45.096
S+I+C	95.153	98.778	99.583	83.983	45.127
I+C+T	95.031	98.839	99.597	85.655	45.728
C+I+T	95.056	99.083	99.639	84.643	46.331
T+I+C	95.850	99.247	99.684	86.750	43.969
I+S+T	91.799	97.640	98.791	80.963	38.979
S+I+T	91.764	97.486	98.944	79.991	40.322
T+I+S	92.853	98.006	98.942	81.975	37.373
C+S+T	90.903	96.833	98.250	78.983	40.496
S+C+T	90.792	97.014	98.194	79.007	40.467
T+C+S	93.191	97.795	98.825	83.048	40.155
I+C+S+T	95.751	99.127	99.645	86.501	46.628
C+I+S+T	96.444	99.236	99.694	85.686	47.192
S+I+C+T	96.000	99.083	99.694	85.742	47.156
T+I+C+S	96.792	99.374	99.784	87.460	44.498
MM-1 Aver.	68.123	81.909	86.360	58.085	24.320
MM-2 Aver.	86.154	94.604	96.759	75.258	35.581
MM-3 Aver.	93.525	98.222	99.145	82.831	42.366
MM-4 Aver.	96.247	99.205	99.704	86.347	46.368

1. Privacy: All individuals using the ORBench dataset and ReID5o model should agree to protect the privacy of all the subjects in it. The users should bear all responsibilities and consequences for any loss caused by the misuse.

2. Redistribution: The ORBench dataset and ReID5o model, either entirely or partly, should not be further distributed, published, copied, or disseminated in any way or form without a prior approval from the creators, no matter for profitable use or not.

3. Commercial Use: The ORBench dataset, ReID5o model, in full, part, or in derived formats, is not permitted for commercial use. Derivative works for commercial purposes are also prohibited. If a commercial entity wants to use them, a separate licensing agreement with the creators must be negotiated.

4. Modification: The ORBench dataset, in whole or part, cannot be modified. This includes changes to data elements, visual content, textual descriptions, and metadata, as well as structural changes. Modification is restricted to maintain dataset integrity for research. In rare cases where modification is needed, a request to the creators must be made, specifying the reasons and changes.

In parallel, we will require users to provide relevant information and will rigorously screen the submitted details to restrict access to our dataset and models by institutions or individuals with a history of privacy violations.

A.5 Broad Impact Discussion

This paper explores multi-modal person re-identification technology, which holds promise for applications in smart retail, intelligent transportation, and public safety systems by enabling cross-camera identification and tracking of individuals in urban environments.

However, the deployment of such technology must be approached with careful consideration of its ethical and societal implications. First, we explicitly acknowledge that the term "bias" in this context encompasses not only labeled sensitive attributes but also imbalances in data distribution. Our dataset, ORBench, is sourced primarily from university settings and does not reflect a globally uniform

demographic distribution. Certain demographic groups—such as the elderly, individuals with visible disabilities, or specific racial groups—are likely underrepresented. As a result, model performance may vary across populations, and end-users should validate the fairness and generalization of the system in their specific application contexts prior to deployment.

Furthermore, while we adopted a structured annotation protocol to maximize objectivity and consistency—focusing on descriptive visual attributes such as clothing and accessories—we recognize that the design of such guidelines inherently introduces a form of methodological bias. By prescribing which features are salient, the annotation process shapes the model's attention and may omit other potentially relevant characteristics.

The potential for misuse of this technology also warrants serious attention. Although our research is intended for academic and public safety purposes, the capability to retrieve individuals via textual descriptions could be repurposed to target people based on attributes associated with protected categories—such as disability, religious attire, or age. Like many dual-use technologies, person re-identification systems require strong governance, ethical oversight, and clear usage policies to prevent discriminatory or privacy-infringing applications. We oppose any such misuse.

In light of these considerations, we emphasize the importance of legal and regulatory frameworks to govern the application of ReID systems. Training often relies on surveillance data collected without explicit individual consent, raising privacy concerns. The research community should avoid using ethically problematic datasets and adopt responsible data practices. To mitigate potential harm, we will release ORBench under gated access, with usage restricted to non-commercial research purposes and subject to strict guidelines.

By addressing these issues transparently, we hope to encourage the responsible development and deployment of multi-modal person re-identification systems.