Chat-based Person Retrieval via Dialogue-Refined Cross-Modal Alignment

Yang Bai[1,2], Yucheng Ji[3], Min Cao[3], Jinqiao Wang[2,4], Mang Ye[1*]
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, China
[2]Wuhan AI Research, Wuhan, China
[3]School of Computer Science and Technology, Soochow University, Suzhou, China
[4]Institute of Automation, Chinese Academy of Science, Beijing, China
{ybai.ai, yemang}@whu.edu.cn

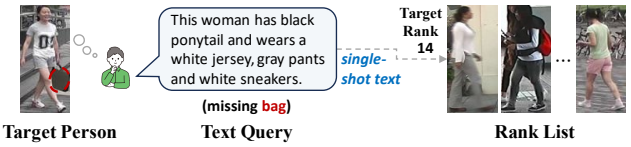
https://github.com/Flame-Chasers/DiaNA

Abstract

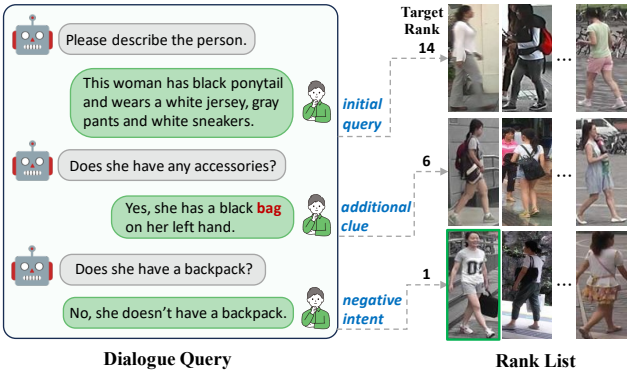
*Traditional text-based person retrieval (TPR) relies on a single-shot text as query to retrieve the target person, assuming that the query completely captures the user's search intent. However, in real-world scenarios, it can be challenging to ensure the information completeness of such single-shot text. To address this limitation, we propose chat-based person retrieval (**ChatPR**), a new paradigm that takes an interactive dialogue as query to perform the person retrieval, engaging the user in conversational context to progressively refine the query for accurate person retrieval. The primary challenge in ChatPR is the lack of available dialogue-image paired data. To overcome this challenge, we establish **ChatPedes**, the first dataset designed for ChatPR, which is constructed by leveraging large language models to automate the question generation and simulate user responses. Additionally, to bridge the modality gap between dialogues and images, we propose a dialogue-refined cross-modal alignment (**DiaNA**) framework, which leverages two adaptive attribute refiner modules to bottleneck the conversational and visual information for fine-grained cross-modal alignment. Moreover, we propose a dialogue-specific data augmentation strategy, random round retaining, to further enhance the model's generalization ability across varying dialogue lengths. Extensive experiments demonstrate that DiaNA significantly outperforms existing TPR approaches, highlighting the effectiveness of conversational interactions for person retrieval.*

1. Introduction

Given the paramount importance of ensuring public safety through intelligent video surveillance, there is a substantial need to develop systems capable of accurately retrieving a person of interest from a large image gallery. In this context, text-based person retrieval (TPR) [23, 39, 53] employs free-form natural language as the query to perform the retrieval, effectively addressing the limitation that visual queries, such as person images (referred to as person re-identification [34, 51, 54, 56]), are not always readily available in real-world scenarios. Consequently, TPR has



(a) Illustration of text-based person retrieval (TPR).



(b) Illustration of chat-based person retrieval (ChatPR).

Figure 1. Comparison of TPR and ChatPR. (a) TPR relies on a single-shot text as the query, which fails to completely capture users' search intent due to the absence of several details. (b) ChatPR, in contrast, engages users in interactive conversations, prompting them to offer additional clues and clarify the negative intents. This interaction allows a progressive refinement of the query, thereby enhancing the retrieval accuracy. The correct image is marked with green rectangles.

*Corresponding Author: Mang Ye

gained ever-rising attention in recent years [24, 42, 50, 58].

Existing methods in TPR typically rely on a single-shot text query, assuming that the query comprehensively captures the user's search intent. However, in practical applications, it can be challenging to ensure the completeness of information in a single-shot text query for person retrieval. As illustrated in Fig. 1a, the user provides an incomplete single-shot text query with the absence of several details (*e.g.*, "bag"), leading to an unsatisfactory retrieval. Alternatively, engaging in multiple rounds of iterative interaction using a question-answer format allows for the conveyance of more detailed information. Recognizing this limitation of single-shot text query in TPR, we introduce a new retrieval paradigm: chat-based person retrieval (**ChatPR**), which takes interactive dialogues as the query to facilitate a more nuanced and comprehensive interpretation of users' search intent for accurate person retrieval. As shown in Fig. 1b, ChatPR engages users in conversational interaction, allowing the retrieval system to progressively refine the query by actively soliciting additional information, thereby achieving more accurate retrieval.

ChatPR presents three advantages over traditional TPR. Firstly, it aligns more closely with real-world applications, in which multiple rounds of iterative interaction through a question-answer format provide a more natural way to query the target person. Secondly, the question-answer format actively prompts users for additional information through conversational interactions, allowing the query to encompass more nuanced details. Lastly, it can explicitly capture the user's negative retrieval intents, which are often overlooked in single-shot text queries but are crucial for filtering out irrelevant targets and narrowing the search scope. These unique features highlight the necessity of ChatPR to leverage interactive dialogues for accurate person retrieval.

However, when addressing the ChatPR task, we are faced with two main challenges: ❶ **Lack of dialogue-image paired data**: While the research community has witnessed an increase in the availability of text-image paired datasets for TPR [12, 30, 61, 62], there remains an absence of dialogue-image paired dataset. This data deficiency poses a significant barrier to the research and development of ChatPR. ❷ **Complexity of cross-modal alignment**: In contrast to the single-shot text, dialogue data presents a more complex structure and longer context. This poses a significant challenge for the traditional TPR models in understanding conversational dialogues, since their text encoder, such as BERT [11] or CLIP [44], is predominantly pretrained on single-text corpora, thereby posing an obstacle to the subsequent cross-modal alignment. Additionally, dialogues tend to have sparser information density, necessitating multiple rounds of interaction to identify specific attributes, which further impedes the cross-modal alignment between dialogue and images.

In response to the first challenge, considering that collecting such data via human crowd-sourcing is not only time-consuming but also lacks standardized and well-defined guidelines [35], we propose leveraging the advanced instruction-following capabilities of large language models (LLMs) [8, 14] to convert existing image-text pairs into structured dialogue-image formats. While LLM-based dialogue generation is highly efficient, the resulting data inevitably introduces noise. To address this, we introduce a diverse set of specialized evaluators, composed of LLMs with meticulously crafted instructions, to assess the relevance and quality of each dialogue round. These individual evaluation results are then integrated to facilitate data cleaning. As a result, we construct **ChatPedes**, the first dataset specifically tailored for ChatPR. This dataset offers a robust benchmark for advancing person retrieval within real-world conversational contexts.

To address the second challenge, we propose a dialogue-refined cross-modal alignment (**DiaNA**) framework, which consists of dual encoders to comprehensively understand dialogues and images, respectively. Given the natural affinity of LLMs in handling dialogue, we adopt the open-source Llama 3 [14] as the dialogue encoder to effectively comprehend the multi-round interactions. Additionally, DiaNA incorporates two adaptive attribute refiners on the dialogue and image branches, which leverage attribute queries to respectively bottleneck the conversational and visual information. This design establishes a fine-grained cross-modal alignment, thereby enhancing the effectiveness of DiaNA. Furthermore, we propose a dialogue-specific data augmentation strategy, random round retaining (R3), to simulate the variability in dialogue lengths in real-world interactions, where complete information is not always available. This strategy further enhances DiaNA's generalization ability across varying lengths of dialogue context.

Overall, the primary contributions of this work can be summarized as follows:

- Recognizing the limitations of single-shot text query in text-based person retrieval, we introduce a new retrieval paradigm: chat-based person retrieval (ChatPR), which engages users in interactive conversations to progressively refine their queries for accurate person retrieval.
- To fill in the data lack, we establish the first dialogue-image paired dataset, ChatPedes, offering the community a robust benchmark for ChatPR.
- To address the challenges in cross-modal alignment, we propose a dialogue-refined cross-modal alignment (DiaNA) framework to enable fine-grained dialogue-image alignment, and introduce a random round retaining strategy to further boost the model's generalization ability.
- Extensive experiments demonstrate the effectiveness of DiaNA in handling multi-round dialogues to improve person retrieval performance over conversational contexts.

2. Related Work

2.1. Text-based Person Retrieval

Text-based person retrieval [39] aims to find the specific person image based on a textual description. Since Li et al. [30] introduced this task, the community has witnessed its growing flourish in recent years. The studies initially depend on simple global alignment [17, 60], and then gradually evolve to multi-granularity correspondences [7, 19, 45]. More recently, a growing number of works [1, 2, 5, 18, 24, 42] have pivoted towards integrating visual-language pretraining models [28, 29, 44] to leverage general cross-modal knowledge for this fine-grained retrieval. Among them, IRRA [24] incorporates a multi-modal interaction encoder on CLIP [44] to facilitate implicit cross-modal relation reasoning. Furthermore, TBPS-CLIP [5] provides a comprehensive empirical study on the application of CLIP in this field from the perspectives of data augmentation, loss function and training tricks.

Despite these advancements, existing methods implicitly assume that the single text query fully encapsulates the user's search intent. However, in real-world scenarios, user-provided queries often lack sufficient details and complete information, necessitating iterative refinement for satisfactory retrieval. To address this limitation, this paper introduces a novel chat-based person retrieval paradigm, which engages users in interactive dialogues to progressively refine their queries for more accurate and effective retrieval.

2.2. Interactive Cross-Modal Retrieval

There has been a significant progress in interactive cross-modal retrieval systems, drawing inspiration from visual dialogue [10, 25]. These systems engage users in a dynamic feedback loop towards the desired retrieval target. Typically, user feedback can be categorized into two types: relevance feedback and difference feedback. Relevance feedback [48] involves users assigning scores to the retrieved results based on their relevance to the query, allowing the system to re-rank its retrieval list accordingly. In contrast, difference feedback [4, 33] requires users to specify the distinctions between the target image and the reference image through pre-defined attribute tags [22, 49] or open-form descriptions [21]. Moreover, recent methods [26, 27] bridge the feedback with the query through conversational questioning, enabling the users to be pro-actively asked according to the dialogue history in each search attempt.

Different from the aforementioned works, this paper focuses on the specific field of ChatPR, a more fine-grained retrieval task, where users need to provide more comprehensive and discriminative queries to retrieve the target pedestrian images from a vast image gallery.

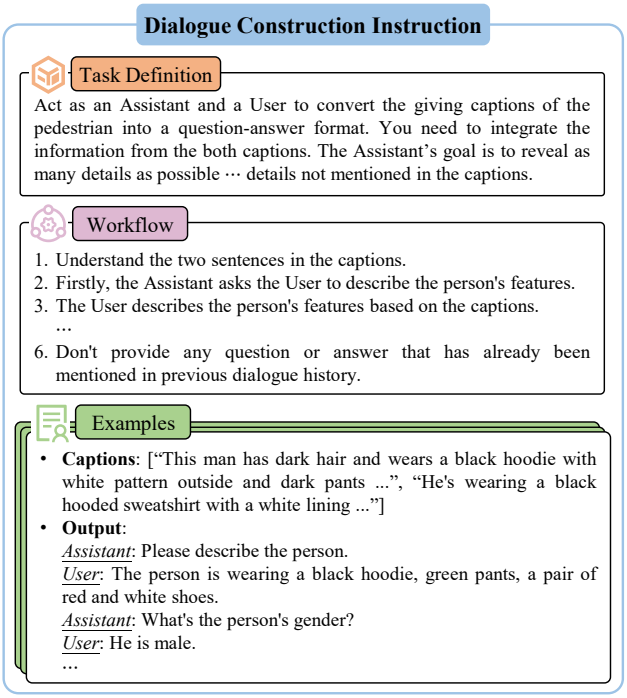


Figure 2. The instruction for dialogue construction, which consists of three parts: task definition, workflow and examples. This design aims to activate LLMs' instruction-following, chain-of-thought and in-context capabilities, respectively.

2.3. Large Language Models

In recent years, the field of artificial intelligence has witnessed a significant leap forward with the advent of large language models (LLMs) [6, 32, 55, 59]. Among the most notable contributions to this domain are the generative pre-trained transformer (GPT) series [3, 40, 41, 43]. These models are trained on massive corpora, allowing them to capture a wide range of linguistic patterns and world knowledge. Moreover, the large-scale parameters endow LLMs with emergent abilities [47] that are not present in smaller models [16], such as instruction-following [38] and in-context [13] learning. These capabilities allow LLMs to accurately interpret and execute complex human instructions, as well as understand and utilize context from a few examples to perform tasks effectively.

In this work, we utilize LLMs to automate question generation and simulate user responses for the construction of ChatPedes. By providing carefully crafted instructions and a few examples, we fully exploit their instruction-following and in-context learning capabilities to create the first benchmark for ChatPR.

3. Benchmark

Recognizing the limitations of single-shot text queries with incomplete descriptions in TPR, we propose the ChatPR task, which enables the progressive query refine-

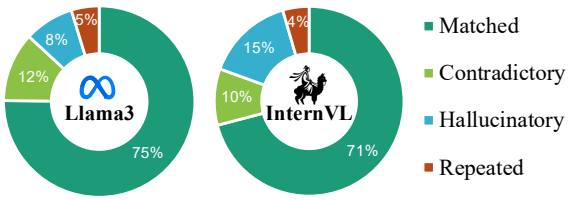


Figure 3. Category distribution of dialogue data from Llama 3 [14] and InternVL [8], evaluated by Qwen 2.5 [46]. Over one-quarter of the data is not matched, highlighting the necessity of data cleaning.

ment through conversational interactions for accurate person retrieval. However, the primary challenge lies in the lack of dialogue-image paired data. To address this, we construct the ChatPedes dataset through two key phases: dialogue construction and data cleaning.

3.1. Dialogue Construction

In recent years, the research community has witnessed an increase in the availability of text-image paired datasets for TPR [12, 30, 61, 62]. Typically, to create high-quality datasets, a large workforce is employed to manually annotate the images with textual descriptions. However, when it comes to annotating with conversational dialogues via human crowd-sourcing for ChatPR, two major challenges arise: ❶ Increased difficulty of dialogue annotation: Compared to single-shot text, dialogue data presents a more complex structure and longer context, significantly increasing the manual annotation workload. ❷ Less well-defined guidelines: Unlike textual annotation tasks with clear instructions, formulating appropriate conversations for dialogue annotation lacks well-defined guidelines [35], complicating the annotation process. Consequently, these challenges make it difficult to employ human crowd-sourcing for creating dialogue-image paired datasets.

Inspired by the success of LLMs in text annotation tasks [20, 50], we leverage their advanced instruction-following capabilities to automate dialogue annotations, based on the well-known text-image paired CUHK-PEDES dataset [30] in TPR. As illustrated in Fig. 4, dialogue construction aims to construct on-going conversations through the interaction between the LLM-based AI agent and the user. During the interaction, the AI agent formulates follow-up questions based on previous dialogue history. The user, also simulated by the LLM, responds with corresponding answers based on the characteristics of the target person, drawing from the existing annotated captions.

In general, it is expected that the dialogue data produced by the LLM-based AI agent contains comprehensive and valuable information, which is advantageous for subsequent model training. Consequently, we empirically design the instructions shown in Fig. 2. Structurally, the instruction consists of three parts: task definition, workflow, and a few manually designed examples, aiming to activate LLMs' instruction-following, chain-of-thought and in-context capabilities to construct high-quality dialogue data. In the dialogue content, the first question provided by the AI agent typically inquires about a holistic description of the target person. Furthermore, to enhance data diversity, we leverage two publicly available LLMs, Llama 3 [14] and InternVL [8], to generate two distinct dialogues for each person image. The experiments detailed in Sec. 5.3.2 substantiate the benefits of this data diversification strategy.

Datasets	IDs	Images	Texts	Dlgs.	Rounds			Words		
					max	min	avg	max	min	avg
CUHK-PEDES [30]	13,003	40,206	80,440	-	-	-	-	96	15	23.5
ICFG-PEDES [12]	4,102	54,522	54,522	-	-	-	-	83	9	37.2
RSTPReid [61]	4,101	20,505	41,010	-	-	-	-	70	11	26.5
UFine6926 [62]	6,926	26,206	52,412	-	-	-	-	218	30	80.8
ChatPedes (Ours)	12,003	37,128	-	74,256	17	1	7.7	388	12	114.4

Table 1. Comparison between our established ChatPedes and the widely-used TPR benchmarks. "Dlgs." refers to dialogues. ChatPedes uses the multi-round dialogues as queries, demonstrating a longer context with conversational interactions.

3.2. Data Cleaning

Although LLMs facilitate the dialogue data collection, the generated data inevitably contains noise, which would significantly impair the optimization of the retrieval process [57]. To perform the data cleaning, we first define the following four classes for each dialogue round:

- **Matched**: The answer either aligns with the information present in the annotated captions or accurately reflects the absence of such details not mentioned in the captions.
- **Contradictory**: The answer contradicts the information present in the captions.
- **Hallucinatory**: The answer contains fabricated information that is not present in the captions.
- **Repeated**: The question or answer contains information that has been mentioned in earlier dialogue rounds.

Subsequently, we employ LLMs, including Qwen 2.5 [46], Llama 3 [14] and InternVL [8], as specialized evaluators through the crafted instructions (detailed in Appendix B.1) to classify each dialogue round into the four predefined categories. As illustrated in Fig. 3, the category distribution estimated by Qwen 2.5 reveals that over one-quarter of the dialogue data is classified as unmatched, highlighting the necessity for data cleaning. To address this, a vote-based ensemble strategy is adopted to perform the data cleaning. Specifically, we calculate the matching score s for each dialogue round, which represents the proportion of the answers labeled as "Matched" by the LLMs:

$$s = \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbb{1}(a_i = \text{Matched}), \tag{1}$$

where $\mathbb{1}(\cdot)$ is identity function, N_q denotes the number of

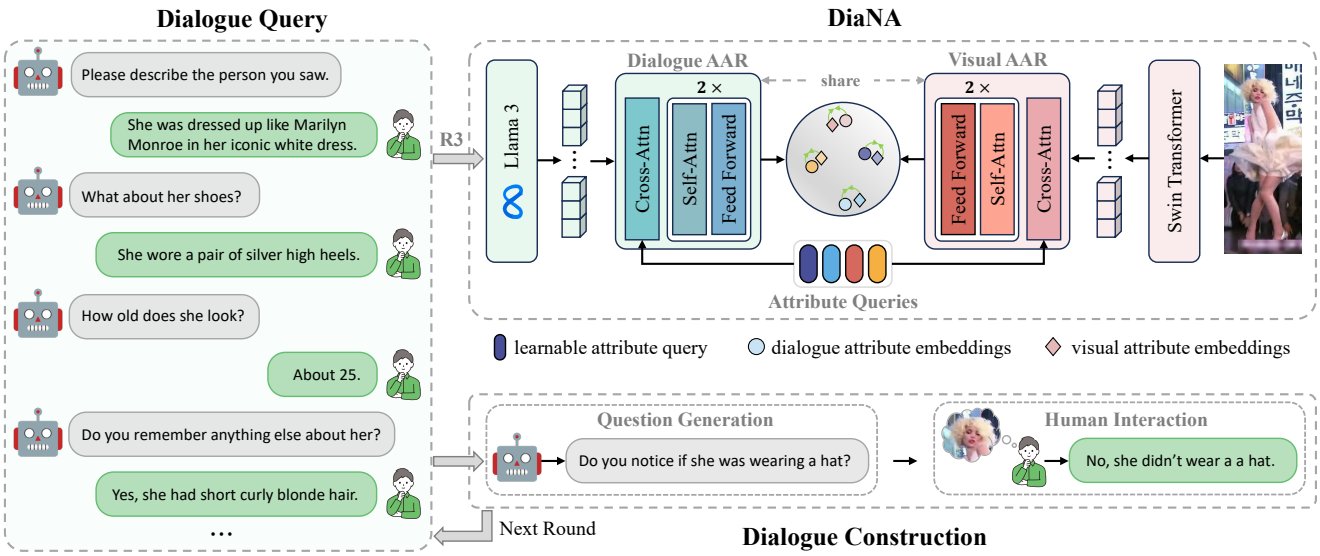


Figure 4. The overview of ChatPR. **Dialogue Construction** aims to construct an on-going conversation, in which an AI agent formulates follow-up questions based on previous dialogue history, prompting the user to provide additional information about the target person to be searched. This iterative process forms a comprehensive **Dialogue Query** for the target person. **DiaNA** aims to bridge the modality gap between the dialogue and the image, which incorporates two adaptive attribute refiner (AAR) modules by leveraging learnable attribute queries to extract key information for fine-grained cross-modal alignment. During training, random round retaining (R3) strategy simulates the incomplete query in reality to enhance the model's generalization ability.

the adopted evaluators, and a_i represents the i-th evaluator's answer. If the score s is less than half of the number of evaluators N_q, this round of conversation is regarded as data noise and then discarded from the dialogue. Detailed statistics of data cleaning are provided in Appendix C.

Finally, we establish ChatPedes, a dialogue-image paired dataset designed for ChatPR. As shown in Tab. 1, ChatPedes comprises $37,128$ images from $12,003$ identities, with each image accompanied by two generated dialogues. Consistent with the data split in CUHK-PEDES [30], we divide $34,054$ image from $11,003$ identities into the training set, and $3,074$ image from $1,000$ identities into the test set. In addition, ChatPedes is characterized by multi-round dialogues and longer context, exhibiting the longest average length of 114.4 words per dialogue among all benchmarks. Visualization examples are presented in Appendix E.

4. Method

4.1. Dialogue-Refined Cross-Modal Alignment

ChatPR engages users in interactive dialogues to progressively refine their queries, where they are pro-actively questioned and prompted to provide more detailed information for more accurate person retrieval. To bridge the modality gap between dialogues and images, we propose a dialogue-refined cross-modal alignment (**DiaNA**) framework, as illustrated in Fig. 4, which aims to learn a dialogue-image shared latent space, pulling the pairs with consistent semantics closer together, while pushing those with different semantics further apart. DiaNA is composed of dual encoders to comprehend dialogue queries and person images, respectively, along with two adaptive attribute refiners to bottleneck the conversational and visual information for fine-grained cross-modal alignment.

4.1.1. Dual Encoders

In contrast to the single-shot text query in TPR, dialogue data in ChatPR presents a more complex structure and longer context, rendering traditional text processing models, such as BERT [11] and CLIP [44] (text encoder), less suitable for this task. Given the proficiency of LLMs in handling dialogue data, we adopt the publicly accessible Llama 3 [14] to comprehend the dialogues within our framework.

Formally, we denote a dialogue as $\mathbf{D} = \{(Q_i, A_i)\}_{i=1}^{N}$, where Q_i and A_i represent the i-th round of question and answer in the dialogue with a total of N rounds. We construct the instruction-format input, as depicted in Fig. 5. Following Vicuna [9], we prepend a system message $\boldsymbol{X}_{\text{system-message}}$ (detailed in Appendix B.2) to activate the instruction-following capability of the LLM, enabling a more comprehensive understanding of the dialogue data. Finally, the entire input is fed into Llama 3, yielding an embedding sequence $\mathcal{W} = \{w_{bos}, w_1, \cdots, w_{eos}\}$, where w_{eos} captures the full semantics of the dialogue due to the decoder-only architecture of Llama 3. Subsequently, the global embedding w_{eos} is mapped into the shared space:

$$w = \Phi_D(w_{eos}) \in \mathbb{R}^d, \tag{2}$$

where Φ_D is a projector composed of a linear layer, d is the dimension of the space.

Figure 5. The instruction-formatted input sequence. A system message is prepended to prompt Llama 3 in capturing the semantics of the dialogue context.

For the visual branch, given an image $\mathbf{I}$, we employ Swin Transformer [36, 37] to encode the image into a visual sequence $\mathcal{V} = \{v_1, v_2, \cdots, v_{h\times w}\}$, where h and w represent the height and width of the feature map at the last stage, respectively. Subsequently, a global average pooling layer is adopted to integrate the overall semantics, yielding a global visual embedding $v_g = \frac{1}{h\times w}\sum_{i=1}^{h\times w} v_i$. Similarly, we introduce a projector to map v_g into the shared space:

$$v = \Phi_I(v_g) \in \mathbb{R}^d. \tag{3}$$

4.1.2. Adaptive Attribute Refiner

Unlike the single-shot text where content is generally concentrated, dialogue data typically exhibits sparser information density, requiring multiple rounds of interaction to ascertain specific attributes. This sparsity poses a significant challenge in cross-modal alignment, necessitating information refinement to effectively bridge the modality gap between dialogues and images. To address this, we propose two adaptive attribute refiner (AAR) modules on the dialogue and the image encoders to learn the attribute representations for fine-grained cross-modal alignment.

Specifically, we first introduce a dialogue AAR upon the dialogue encoder. Considering that different persons have varying attributes, we introduce a set of learnable attribute queries $\mathcal{Q} = \{q_1, q_2, \cdots, q_K\}$ to focus on the distinct attributes of each individual, where K denotes the number of attribute queries. By interacting with the dialogue embeddings $\mathcal{W}$, these attribute queries adaptively extract the person-specific attribute representations from the sparse dialogue content. As illustrated in Fig. 4, the dialogue AAR applies a computationally efficient structure with a single multi-head cross attention (MCA) followed by 2-layer transformer blocks. The interaction between the attribute queries and the dialogue embeddings is performed through the cross-attention mechanism:

$$\hat{\mathcal{W}} = \text{Transformer}(\text{MCA}(\mathcal{Q}, \mathcal{W}, \mathcal{W})), \tag{4}$$

where $\hat{\mathcal{W}} = \{\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_K\}$ indicates the adaptively refined dialogue attribute embeddings.

In parallel, we symmetrically introduce a visual AAR to extract the visual attributes from the image, mitigating the impacts of the inherent redundancy and background noise in visual information:

$$\hat{\mathcal{V}} = \text{Transformer}(\text{MCA}(\mathcal{Q}, \mathcal{V}, \mathcal{V})), \tag{5}$$

where $\hat{\mathcal{V}} = \{\hat{v}_1, \hat{v}_2, \cdots, \hat{v}_K\}$ denotes the visual attribute embeddings. Furthermore, we leverage a parameter-sharing mechanism between the two AAR modules, which not only reduces the parameter amount but also mitigates their modality gap [15] between dialogues and images.

4.2. Optimization Objectives

Incorporating the dual encoders and the AAR modules, we construct DiaNA, which is then optimized through both global (coarse-grained) and local (fine-grained) alignment. Specifically, we use the cosine similarity to measure the global alignment between the dialogue and the image:

$$s_g(\mathbf{D}, \mathbf{I}) = \frac{w^{\mathrm{T}} v}{\|w\|\|v\|}. \tag{6}$$

Furthermore, the average cosine distance among the refined attribute embeddings is utilized to measure the local semantics consistency between the dialogue and the image:

$$s_l(\mathbf{D}, \mathbf{I}) = \frac{1}{K}\sum_{i=1}^{K} \frac{\hat{w}_i^{\mathrm{T}} \hat{v}_i}{\|\hat{w}_i\|\|\hat{v}_i\|}. \tag{7}$$

According to the global and local similarities between the dialogue and the image, we construct the overall optimization object by employing the normalized image-text contrastive (N-ITC) loss [5] and the similarity distribution matching (SDM) loss [24]:

$$\mathcal{L} = \underbrace{\mathcal{L}_g^{nitc} + \mathcal{L}_g^{sdm}}_{\text{global alignment}} + \underbrace{\mathcal{L}_l^{nitc} + \mathcal{L}_l^{sdm}}_{\text{local alignment}}. \tag{8}$$

4.3. Dialogue-specific Data Augmentation

In real-world scenarios, complete dialogues that capture users' full search intent are not always available, since obtaining these complete dialogues typically requires numerous rounds of interaction, which can be not user-friendly. To enhance the model's generalization ability over dialogues of varying rounds, we propose a dialogue-specific data augmentation strategy, **random round retaining** (R3), which simulates the inherent variability of the dialogue in practice.

Specifically, at each training iteration, for a complete dialogue $\mathbf{D} = \{(Q_i, A_i)\}_{i=1}^{N}$, only the initial random i rounds are retained and fed into the model:

$$\hat{\mathbf{D}} = \{(Q_i, A_i)\}_{i=1}^{n}, \quad n \sim \text{randint}(1, N), \tag{9}$$

where $\text{randint}(1, N)$ denotes a discrete uniform distribution with equal probability for each integer in range $[1, N]$. The retained $\hat{\mathbf{D}}$ reflect the partial availability of information in practical scenarios, effectively enhancing the model's generalization ability over varying dialogue lengths.

Method	Ref	Pret. Data	w/o Data Cleaning				w/ Data Cleaning			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
IRRA [24]	CVPR'23	WIT	46.08	65.60	73.02	42.90	48.26	67.96	75.65	44.64
TBPS-CLIP [5]	AAAI'24	WIT	47.74	66.41	73.78	43.80	50.21	69.27	76.43	45.44
RDE [42]	CVPR'24	WIT	50.20	68.25	75.13	45.94	52.77	70.61	77.47	47.98
APTM [52]	MM'23	MALS	39.62	56.99	64.23	36.57	41.15	57.91	65.16	38.00
AUL [31]	AAAI'24	MALS	40.05	57.01	64.41	37.02	42.01	59.21	66.01	38.56
DiaNA (Ours)	-	MALS	**73.72**	**89.66**	**94.24**	**65.40**	**75.67**	**90.68**	**94.71**	**66.89**

Table 2. Comparison with state-of-the-art methods on ChatPedes. "Pret. Data" denotes pretraining data. All reported results of other methods are reproduced using their publicly available code.

5. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our method. The implementation details are provided in Appendix A.

5.1. Evaluation Protocol

We adopt the widely-used Rank-K (R-K for short, K=$1, 5, 10$) metric in our experiments, which denotes the percentage of successful retrieval that the ground-truth image is found within the top K ranked results. In addition, we also employ the mean average precision (mAP) as an auxiliary metric to evaluate the overall retrieval performance.

5.2. Comparison with SOTA Methods

We compare our method against five state-of-the-art (SOTA) baselines: IRRA [24], TBPS-CLIP [5], RDE [42], APTM [52] and AUL [31]. Among these, IRRA, TBPS-CLIP and RDE are built on the vision-language foundation model CLIP [44], which is pretrained on a large-scale dataset WIT with 400 million image-text paired data collected from the Internet. APTM and AUL are pretrained on the synthetic text-image paired dataset MALS [52]. It is worth noting that these baselines are originally developed for the TPR task. We reproduce their performance on ChatPedes dataset using their open-source codes, treating the multi-turn dialogue data as a single long-form text input. As shown in Tab. 7, regardless of whether data cleaning is applied, our method outperforms these SOTA baselines by a large margin, while these baselines demonstrate a suboptimal retrieval accuracy, indicating their limitations in effectively handling the complex structure and longer contextual dependencies of dialogue data.

In addition, we further conduct extensive comparative experiments over the dialogue round, simulating real-world scenarios where users are pro-actively questioned to progressively refine their queries through conversational interactions. This process is exemplified in Appendix E.1. As shown in Fig. 6, it can be observed that: ❶ As the dialogue progresses, the retrieval performance improves across all methods. However, the performance of the baselines becomes saturated after the initial few rounds of interaction,

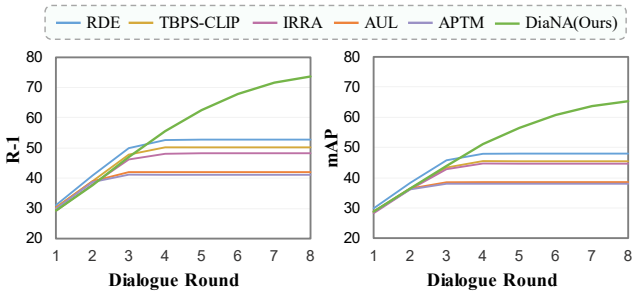


Figure 6. Performance comparison over dialogue round.

which we attribute to their limitations in processing longer and more complex dialogues. ❷ In contrast, our method presents a significantly pronounced improvement as the dialogue evolves on account of our carefully crafted design for handling dialogue data. This comprehensive comparison demonstrates the efficacy of our method and showcases its superior capability in handling the intricacies of dialogue-image alignment. ❸ Additionally, we also notice that our method achieves competitive performance at the first round of dialogue. This round typically involves a holistic description of the target person, functioning similarly to the single-shot text in the traditional TPR task. The experimental comparison also indicates the promising generalization capabilities of our method in TPR. Further detailed experimental results on TPR benchmarks are provided in Appendix D.

5.3. Ablation Study

5.3.1. Effectiveness of Data Cleaning

During dialogue construction, we leverage LLMs to generate dialogue data. However, a major limitation of this automatic generation through LLMs lies in the inevitable introduction of noise, which can be classified into three categories: Contradictory, Hallucinatory and Repeated, as delineated in Sec. 3.2. To evaluate the impact of the noise, we conduct comparative experiments with and without data cleaning. Tab. 7 presents the performance comparison with the five SOTA baselines, where all approaches exhibit a significant improvement. In particular, RDE achieves a notable increase of 2.57% at R-1, demonstrating the effectiveness of data cleaning and the importance of high-quality data.

Data Diversity		R-1	R-5	R-10	mAP
Llama 3	InternVL				
✗	✓	71.01	88.40	93.18	63.08
✓	✗	71.62	88.58	93.59	63.73
✓	✓	**75.67**	**90.68**	**94.71**	**66.89**

Table 3. Ablation study on data diversity. We employ Llama 3 [14] and InternVL [8] to construct the ChatPedes benchmark.

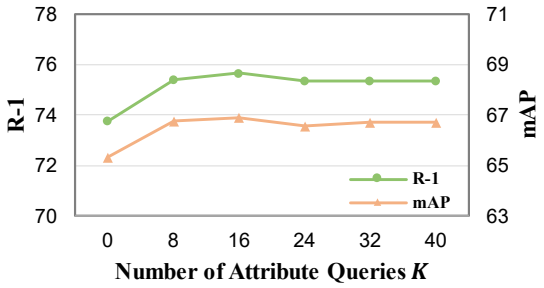


Figure 7. Performance comparison across a varying number of attribute queries K in AAR modules. Notably, $K = 0$ indicates that the AAR is not used.

5.3.2. Effectiveness of Data Diversity

We employ Llama 3 [14] and InternVL [8] as different annotators for labeling diverse dialogue data when constructing ChatPedes. Tab. 3 presents the contribution of data diversity to the retrieval performance. It can be observed that: ❶ Relying on a single LLM yields a suboptimal outcome, with 71.01% for InternVL and 71.62% for Llama 3 at R-1. Additionally, Llama 3 slightly outperforms InternVL due to its more advanced capability in dialogue generation. ❷ The combination of Llama 3 and InternVL leads to a substantial increase across all metrics, demonstrating the importance of data diversity. By combining annotations generated by different LLMs, the established ChatPedes dataset effectively captures a wider variety of dialogue contexts and nuances, which in turn bolsters the model's retrieval capabilities.

5.3.3. Effectiveness of Adaptive Attribute Refiner

Considering the sparse information density in dialogues and the inherent redundancy in images, we propose two adaptive attribute refiner (AAR) modules, which employ a set of learnable attribute-specific queries to adaptively extract the relevant attributes from the dialogue and the image. Fig. 7 illustrates the performance variation over the number of attribute queries K, where, notably, $K = 0$ denotes AAR is discarded. We observe that: ❶ Only relying on global alignment without AAR results in a suboptimal retrieval performance, underscoring the importance of key information extraction. ❷ When introducing AAR, the performance exhibits slight fluctuations, reaching the peak of 75.67% on R-1 and 66.89% on mAP at $K = 16$. Overall, the performance remains relatively stable, indicating the model's robustness when equipped with AAR.

Method	Metric	Dialogue Round					
		1	2	4	6	8	all
DiaNA(w/o R3)	R-1	26.50	34.89	53.07	66.40	73.16	75.18
DiaNA(w/ R3)	R-1	29.25	37.64	55.60	67.99	73.67	75.67
Δ		**+2.75**	**+2.75**	**+2.53**	**+1.59**	**+0.51**	**+0.49**
DiaNA(w/o R3)	mAP	26.33	33.67	49.06	59.67	64.64	66.19
DiaNA(w/ R3)	mAP	28.78	36.37	51.05	60.88	65.32	66.89
Δ		**+2.45**	**+2.70**	**+1.99**	**+1.21**	**+0.68**	**+0.70**

Table 4. Comparison of random round retaining (R3) over dialogue round. "all" refers to the complete dialogue.

5.3.4. Effectiveness of Random Round Retaining

The proposed random round retaining (R3) strategy selectively retains a random number of initial rounds of the dialogue during training, simulating the users' incomplete search intent in real-world interactions. As shown in Tab. 4, we observe that: ❶ R3 yields performance gains across all dialogue rounds, with gains from 2.75% in single-round interactions to 0.49% in complete conversations at R-1, demonstrating its robustness and versatility. ❷ In particular, R3 brings more substantial improvement at the initial few rounds of dialogues, suggesting that R3 is particularly effective when dialogue information is limited. This is especially crucial in practical scenarios where users expect to quickly retrieve the target person with minimal interaction. Overall, R3 empowers users to obtain accurate retrieval results even in the early stages of interaction, thereby addressing the real-world demands for both efficiency and precision.

6. Conclusion

Recognizing the inherent limitations of traditional TPR where single-shot text queries fail to fully encapsulate users' search intent, we introduce ChatPR, a pioneering paradigm for person retrieval that engages users in interactive dialogues to progressively refine their search queries. For this, we build the first benchmark ChatPedes through LLMs to automate question generation and simulate users' responses. To tackle this task, we propose the DiaNA framework that leverages two adaptive attribute refiners to bottleneck the conversational and visual information for fine-grained cross-modal alignment. Additionally, a novel data augmentation strategy of random round retaining is proposed to mimic the incomplete queries in real-world scenarios, remarkably enhancing the model's generalization ability across varying dialogue lengths. Extensive experiments significantly demonstrate the importance of conversational interaction in person retrieval and the superior effectiveness of our method in ChatPR.

Acknowledgement. This work is supported by the National Science Foundation of China under Grants (62176188, 62361166629, 62476188). The numerical calculations are supported by the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 555–563, 2023. 3

[2] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 757–767, 2023. 3

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[4] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: active detail enriching for cross-modal retrieval with partial query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1835–1844, 2021. 3

[5] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 465–473, 2024. 3, 6, 7

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. 3

[7] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 3

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 4, 8

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna, 2023. 5

[10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 3

[11] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[12] 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. 2, 4

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 3

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 4, 5, 8

[15] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024. 6

[16] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[17] Ammarah Farooq, Muhammad Awais, Fei Yan, Josef Kittler, Ali Akbari, and Syed Safwan Khalid. A convolutional baseline for person re-identification using vision and language descriptions. *arXiv preprint arXiv:2003.00808*, 2020. 3

[18] Takuro Fujii and Shuhei Tarashima. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2786–2790, 2023. 3

[19] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Fangzhou Lin, Xing Sun, and Xiang Bai. Conditional feature learning based transformer for text-based person search. *IEEE Transactions on Image Processing*, 31:6097–6108, 2022. 3

[20] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. 4

[21] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018. 3

[22] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12147–12157, 2021. 3

[23] S. Irene, A. John Prakash, and V. Rhymend Uthariaraj. Person search over security video surveillance systems using deep learning methods: A review. *Image and Vision Computing*, 143:104930, 2024. 1

[24] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 2, 3, 6, 7

[25] Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The dialog must go on: Improving visual dialog via generative self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6746–6756, 2023. 3

[26] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large

language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 791–809, 2024. 3

[27] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems*, pages 61437–61449, 2023. 3

[28] 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 *Advances in Neural Information Processing Systems*, pages 9694–9705, 2021. 3

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900, 2022. 3

[30] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 4, 5

[31] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3172–3180, 2024. 7

[32] Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[33] Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11091–11101, 2023. 3

[34] Fangyi Liu, Mang Ye, and Bo Du. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2(1):28, 2024. 1

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916, 2023. 2, 4

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 6

[38] Renze Lou, Kai Zhang, and Wenpeng Yin. A comprehensive survey on instruction following. *arXiv preprint arXiv:2303.10475*, 2023. 3

[39] Kai Niu, Yanyi Liu, Yuzhou Long, Yan Huang, Liang Wang, and Yanning Zhang. An overview of text-based person search: Recent advances and future directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 7803–7819, 2024. 1, 3

[40] OpenAI. Gpt-4 technical report, 2023. 3

[41] OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2023. 3

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 2, 3, 7

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 5, 7

[45] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5566–5574, 2022. 3

[46] Qwen Team. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5, 2024. 4

[47] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 3

[48] Hong Wu, Hanqing Lu, and Songde Ma. Willhunter: interactive image retrieval with multilevel relevance. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 1009–1012, 2004. 3

[49] Yingjia Xu, Mengxia Wu, Zixin Guo, Min Cao, Mang Ye, and Jorma Laaksonen. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. *Visual Intelligence*, 3(1):1–13, 2025. 3

[50] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024. 2, 4

[51] Jinxi Yang, He Li, Bo Du, and Mang Ye. Cheb-gr: Rethinking k-nearest neighbor search in re-ranking for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

[52] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4492–4501, 2023. 7

[53] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. 1

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

[55] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025. 3

[56] Mang Ye, Zesen Wu, and Bo Du. Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[57] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 4

[58] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 686–701, 2018. 2

[59] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 3

[60] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23, 2020. 3

[61] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 209–217, 2021. 2, 4

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024. 2, 4