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

Supplementary Material

A. Implementation Details

We employ Swin Transformer v2-B [9] as the visual encoder, and Llama 3.2-1B [5] as the dialogue encoder. For the model training, following [8, 17], we employ a two-stage procedure. DiaNA is first pretrained on the synthetic text-image paired dataset, MALS [17]. To adapt to the dialogue-formatted input, each text in MALS is converted into a single-round dialogue by prepending a randomly sampled question (detailed in Appendix B.3), which serves as a language instruction to request an overall description of the target person. Then, we continue to train DiaNA on the established ChatPedes dataset. During training, we apply data augmentation strategies, including random horizontal flipping, random cropping with padding, and random erasing. Additionally, we introduce random masking to the dialogue data at a probability of 15%. The number of attribute queries in the adaptive attribute refiners is set to $K = 16$. More settings are detailed in Tab. 1.

Configuration	Pretraining	Finetuning
epoch	3	10
total batch size	512	128
image resolution	384 × 192	384 × 192
LLM sequence length	100	400
learning rate schedule	cosine decay	cosine decay
optimizer	AdamW [10]	AdamW [10]
optimizer hyper-parameters	β_1, β_2 = 0.9, 0.98	β_1, β_2 = 0.9, 0.999
weight decay	0.05	0.05
learning rate	1e-4	1e-5
warmup steps	1000	50
numerical precision	DeepSpeed bf16 [12]	DeepSpeed bf16 [12]
GPUs for training	8 × RTX 3090	4 × RTX 3090

Table 1. Training settings of DiaNA.

B. Instruction Designing

B.1. The Instruction for Dialogue Categorization

Considering that the generated dialogue data through LLMs inevitably introduces noise, we employ LLMs as specialized evaluators to classify each dialogue round into four categories for the subsequent data cleaning: Matched, Contradictory, Hallucinatory and Repeated. To fully exploit the advanced capability of LLMs, we meticulously design the instruction, as shown in Fig. 1a. The instruction is structured into three parts: task definition, workflow and a few manually annotated examples. These components are designed to respectively activate LLMs' exceptional instruction-following, chain-of-thought and context-learning capabilities for more accurate categorization.

Instruction for Dialogue Categorization

Task Definition

You are provided with two captions describing person and a dialogue with multi-round question-answer interactions. Classify each round of question and answer into the following four categories based on its semantic alignment with the captions:

- **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 introduces fabricated information that is not present in the captions.
- **Repeated**: The question or answer reiterates information that has been mentioned in earlier dialogue rounds.

Workflow

1. Understand the two sentences in the captions.
2. Read the questions and the provided answers of each round dialogue.
3. The Assistant asks the User about the person.
4. The User answers the question based on the captions.

…

8. Make sure that each round of dialogue is checked.

Examples

- **Captions**: ["A pedestrian with short black hair is wearing a purple shirt, red and white shoes ...", "The person is wearing denim crop shorts, red sneakers ..."]
- **Dialogue**:
  Assistant: What is the pedestrian wearing?
  User : The pedestrian is wearing a purple shirt.
  Assistant: Can you describe the pedestrian's shoes?
  User: The pedestrian is wearing a pair of blue shoes.
  …
- **Output**:
  Matched, Contradictory, ...

(a) The instruction for dialogue categorization, which is design to activate the LLMs to categorize each dialogue round for data cleaning.

System Message

You are an intelligent system designed to retrieve pedestrian images from dialogues. Extract key attributes such as clothing, accessories, gender, and distinctive features from the conversation. Use the aggregated information from all dialogue turns to perform accurate person image retrieval.

(b) The system message used to build an instruction-formatted input sequence, guiding the dialogue encoder to focus on the various detailed nuances within the dialogue data.

Instructions for Image Description

1. Please describe the person you saw.
2. What is the person wearing in the picture?
3. Can you detail the clothing and accessories of the person?
4. Describe the appearance of the individual in the image.
5. How would you describe the attire of the person in the image?
6. Please provide details about the person's outfit.
7. Describe the attire and notable features of the individual in the picture.

(c) The list of instructions for image description used to transform the single-shot text into a single-round dialogue.

Figure 1. The designed instructions.

Generator	Evaluator	M.	C.	H.	R.	Total	Retained
Llama 3	Qwen 2.5	244,805 (75.2%)	38,188 (11.7%)	27,339 (8.4%)	15,344 (4.7%)	325,676	288,423 (88.6%)
	Llama 3	273,544 (84.0%)	19,318 (5.9%)	20,280 (6.2%)	12,534 (3.9%)		
	InternVL	297,097 (91.2%)	3,541 (1.1%)	24,523 (7.5%)	515 (0.2%)		
InternVL	Qwen 2.5	213,688 (70.9%)	28,975 (9.6%)	45,103 (15.0%)	13,525 (4.5%)	301,291	238,913 (79.3%)
	Llama 3	236,705 (78.6%)	15,752 (5.2%)	38,783 (12.9%)	10,051 (3.3%)		
	InternVL	237,201 (78.7%)	5,373 (1.8%)	53,760 (17.8%)	4,957 (1.7%)		

Table 2. Detailed statistics of data cleaning. We use a diverse set of evaluators to categorize each dialogue round into four classes: Matched (M.), Contradictory (C.), Hallucinatory (H.) and Repeated (R.). The percentage in () denotes the proportion to the total dialogue rounds.

B.2. System Message

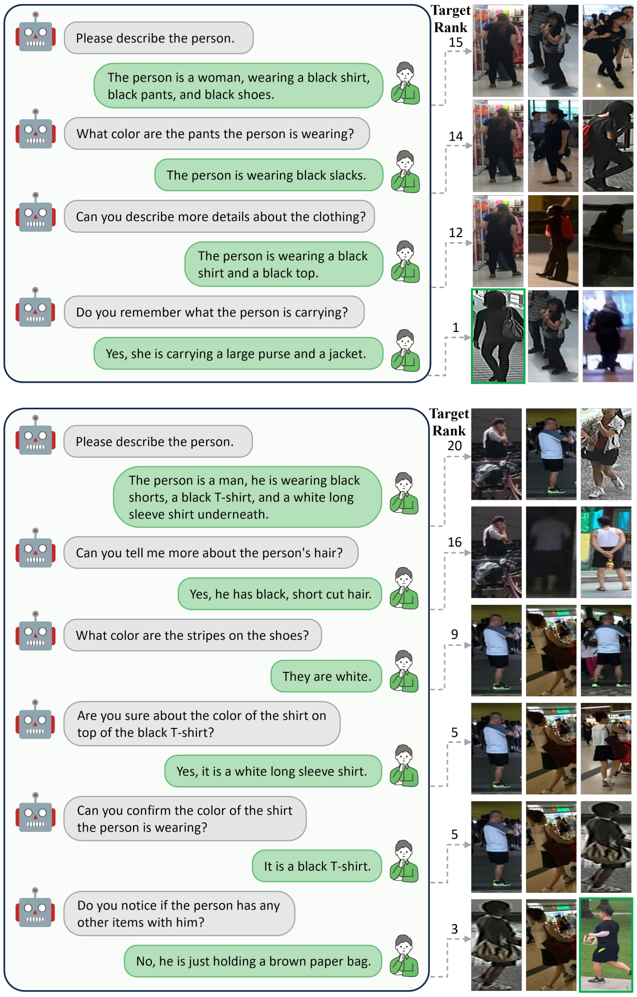
Following Vicuna [4], we prepend a system message shown in Fig. 1b to each dialogue input to construct an instruction-format sequence. The system message serves as an instruction to guide the dialogue encoder in DiaNA, Llama 3 [5], to comprehensively understand the semantics of the dialogue. By incorporating this instruction, we aim to encourage the dialogue encoder to focus on the various nuances of conversational interactions (*e.g.*, clothing, accessories, gender, *etc*. highlighted in the instruction), thereby enriching the feature extraction and the semantic understanding.

B.3. The Instructions for Image Description

To bridge the modality gap between dialogues and images, following APTM [17] and AUL [8], we pretrain DiaNA on the synthetic text-image paired dataset, MALS [17], which is originally collected to facilitate the alignment of textual and visual data in TPR. During the pretraining, to unify the dialogue-formatted input for the dialogue encoder, we convert each single-shot text into a single-round dialogue by prepending a randomly sampled question. The question is drawn from the instruction pool shown in Fig. 1c, which requests a comprehensive description of the target person, while the text is treated as the user's answer, forming a single-round dialogue. By constructing single-round dialogues in this manner, we simulate the interactive dialogue queries in real-world applications and enhance the cross-modal alignment between dialogues and images.

C. Statistics of Data Cleaning

To avoid model bias, we employ a diverse set of LLMs, including Qwen 2.5 [14], Llama 3 [5] and InternVL [3] as specialized evaluators for the dialogue categorization. These evaluators are instructed to classify each generated dialogue round into the four predefined categories: Matched, Contradictory, Hallucinatory and Repeated. As shown in Tab. 2, a large number of interactions in dialogues are identified to be unmatched. In particular, the dialogue data generated by InternVL tends to introduce more noise compared to the data produced by Llama 3, which We attribute to the superior capability of Llama 3 in generating high-quality dialogues. Finally, we propose a vote-based ensemble strategy to clean the dialogue data, retaining 88.6% of the dialogue interactions generated from Llama 3 and 79.3% from InternVL to form the ChatPedes dataset.



Please describe the person.

The person is a woman, wearing a black shirt, black pants, and black shoes.

What color are the pants the person is wearing?

The person is wearing black slacks.

Can you describe more details about the clothing?

The person is wearing a black shirt and a black top.

Do you remember what the person is carrying?

Yes, she is carrying a large purse and a jacket.

Please describe the person.

The person is a man, he is wearing black shorts, a black T-shirt, and a white long sleeve shirt underneath.

Can you tell me more about the person's hair?

Yes, he has black, short cut hair.

What color are the stripes on the shoes?

They are white.

Are you sure about the color of the shirt on top of the black T-shirt?

Yes, it is a white long sleeve shirt.

Can you confirm the color of the shirt the person is wearing?

It is a black T-shirt.

Do you notice if the person has any other items with him?

No, he is just holding a brown paper bag.

Figure 2. More retrieval examples over dialogue round. Through conversational interactions, the users are prompted to refine their queries for more accurate person retrieval. Correct retrieval is marked by green rectangle.

Method	Ref	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Joint Encoder													
RaSa [1]	IJCAI'23	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	66.90	86.50	91.35	52.31
APTM [17]	MM'23	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
AUL [8]	AAAI'24	77.23	90.43	94.41	-	69.16	83.32	88.37	-	71.65	87.55	92.05	-
Dual Encoder													
ViTAA [15]	ECCV'20	55.97	75.84	83.52	-	-	-	-	-	-	-	-	-
LapsCore [16]	ICCV'21	63.40	-	87.8	-	-	-	-	-	-	-	-	-
SRCF [13]	ECCV'22	64.04	82.99	88.81	-	57.18	75.01	81.49	-	-	-	-	-
IRRA [7]	CVPR'23	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
BiLMa [6]	ICCV'23	74.03	89.59	93.62	66.57	63.83	80.15	85.74	38.26	61.20	81.50	88.80	48.51
TBPS-CLIP [2]	AAAI'24	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	61.95	83.55	88.75	48.26
RDE [11]	CVPR'24	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
DiaNA (Ours)	-	73.26	88.15	93.78	65.72	63.78	80.85	85.66	38.86	61.15	83.35	89.65	47.59

Table 3. Comparison with state-of-the-art methods on TPR benchmarks. Existing TPR approaches can be categorized into two types according to whether performing cross-modal interaction during inference: Joint Encoder (applying the cross-modal interaction) and Dual Encoder (discarding the cross-modal interaction).

D. Experiments on TPR Benchmarks

Although DiaNA is specifically designed for ChatPR, it can also handle text-image inputs to perform TPR. In this section, we conduct extensive experiments on TPR benchmarks to verify the generalization ability of DiaNA.

Current TPR methods can be categorized into two types according to whether performing cross-modal interaction during inference: Joint Encoder and Dual Encoder. Joint Encoder typically leverages cross-attention mechanism to perform the cross-modal interaction between images and texts, resulting in overall performance superiority but suffering from burdensome computation complexity, while Dual Encoder offers higher efficiency yet relatively suboptimal retrieval performance due to the non-interactive architecture. As shown in Tab. 3, our proposed DiaNA employs a simple Dual Encoder architecture but still achieves promising results, demonstrating the strong generalization ability of DiaNA. We also observe that DiaNA falls short of achieving the notable advantage in TPR as in ChatPR. On one hand, DiaNA is tailored for ChatPR with specific designs to handle conversational dialogues in ChatPR rather than single-shot texts in TPR. On the other hand, existing TPR methods struggle to manage the complex structure and longer context of dialogue data, leading to suboptimal performance in ChatPR, which is more aligned with real-world scenarios. In our future work, we will delve into a unified framework that excels in both ChatPR and TPR tasks.

E. More Examples

E.1. Retrieval Examples over Dialogue Round

Fig. 2 presents more retrieval examples over dialogue round. It can be observed that a single-text query, provided as the first dialogue round, is insufficient to pinpoint the desired person image. As the dialogue progresses, the users are prompted to refine their queries and gradually achieve accurate retrieval. For instance, the user's responses, such as "*carrying a large purse*" in the first example and "*holding a brown paper bag* in the second example, effectively facilitate the filtering of irrelevant images.

E.2. More Examples in ChatPedes

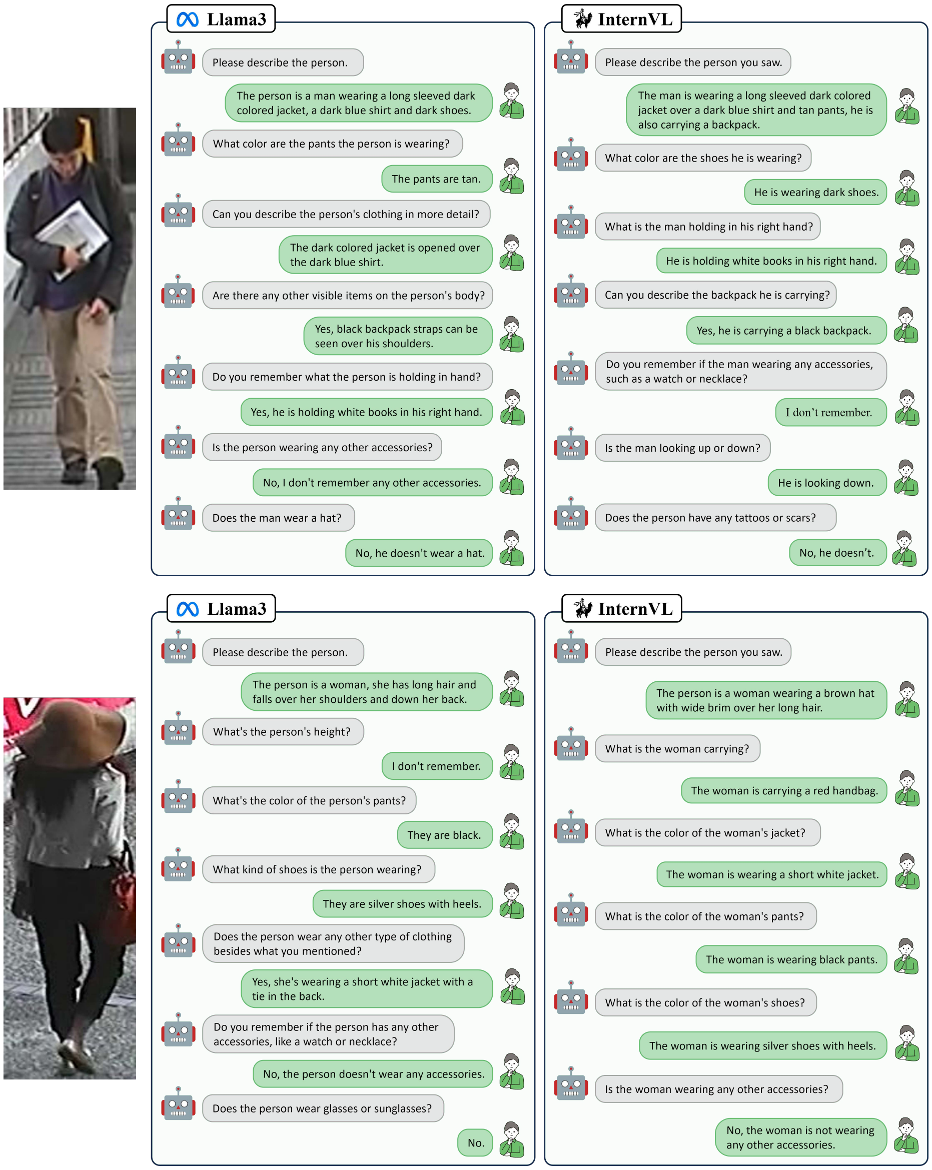
We provide more examples from our established ChatPedes dataset in Fig. 3. Each image is paired with two dialogues generated by Llama 3 [5] and InternVL [3], respectively. As distinct annotators, they produce dialogues with varying styles, expressions, and even different focuses, significantly enhancing the data diversity of ChatPedes. Additionally, Fig. 4 visualizes the word occurrences in questions and answers from ChatPedes, showing their different focuses and keywords in the dataset.

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] 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

[3] 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 *Pro-*



Llama3

Please describe the person.

The person is a man wearing a long sleeved dark colored jacket, a dark blue shirt and dark shoes.

What color are the pants the person is wearing?

The pants are tan.

Can you describe the person's clothing in more detail?

The dark colored jacket is opened over the dark blue shirt.

Are there any other visible items on the person's body?

Yes, black backpack straps can be seen over his shoulders.

Do you remember what the person is holding in hand?

Yes, he is holding white books in his right hand.

Is the person wearing any other accessories?

No, I don't remember any other accessories.

Does the man wear a hat?

No, he doesn't wear a hat.

InternVL

Please describe the person you saw.

The man is wearing a long sleeved dark colored jacket over a dark blue shirt and tan pants, he is also carrying a backpack.

What color are the shoes he is wearing?

He is wearing dark shoes.

What is the man holding in his right hand?

He is holding white books in his right hand.

Can you describe the backpack he is carrying?

Yes, he is carrying a black backpack.

Do you remember if the man wearing any accessories, such as a watch or necklace?

I don't remember.

Is the man looking up or down?

He is looking down.

Does the person have any tattoos or scars?

No, he doesn't.

Llama3

Please describe the person.

The person is a woman, she has long hair and falls over her shoulders and down her back.

What's the person's height?

I don't remember.

What's the color of the person's pants?

They are black.

What kind of shoes is the person wearing?

They are silver shoes with heels.

Does the person wear any other type of clothing besides what you mentioned?

Yes, she's wearing a short white jacket with a tie in the back.

Do you remember if the person has any other accessories, like a watch or necklace?

No, the person doesn't wear any accessories.

Does the person wear glasses or sunglasses?

No.

InternVL

Please describe the person you saw.

The person is a woman wearing a brown hat with wide brim over her long hair.

What is the woman carrying?

The woman is carrying a red handbag.

What is the color of the woman's jacket?

The woman is wearing a short white jacket.

What is the color of the woman's pants?

The woman is wearing black pants.

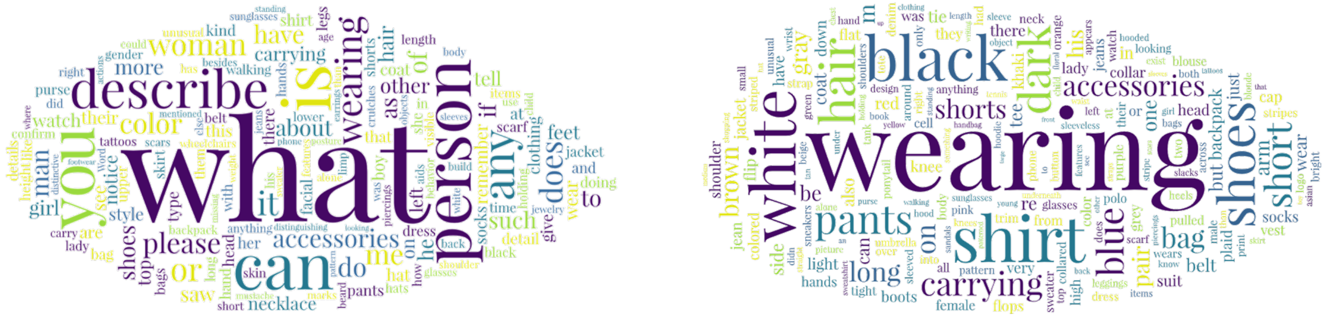
What is the color of the woman's shoes?

The woman is wearing silver shoes with heels.

Is the woman wearing any other accessories?

No, the woman is not wearing any other accessories.

Figure 3. More examples in our established ChatPedes dataset. Each image is annotated with two dialogues by Llama 3 [5] and InternVL [3]. The continuous interactions exhibit the characteristic of ChatPR that users are prompted to progressively refine their queries.



(a) Word cloud of questions in ChatPedes.

(b) Word cloud of answers in ChatPedes.

Figure 4. Top 1000 most frequently occurring words in questions and answers from ChatPedes (larger size indicates higher frequency).

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024. 2, 3, 4

[4] 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. 2

[5] 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. 1, 2, 3, 4

[6] 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

[7] 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. 3

[8] 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. 1, 2, 3

[9] 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. 1

[10] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] 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. 3

[12] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 1

[13] Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. A simple and robust correlation filtering method for text-based person search. In *Proceedings of the European Conference on Computer Vision*, pages 726–742, 2022. 3

[14] Qwen Team. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5, 2024. 2

[15] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Proceedings of the European Conference on Computer Vision*, pages 402–420, 2020. 3

[16] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021. 3

[17] 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. 1, 2, 3