

Supplementary Materials of CrowdVerse: Boosting Crowd Understanding with a Realistic and Comprehensive Multitask Dataset

CVPR 2026 submission

¹Paper ID 4871



The supplementary materials complement the content of the paper in four parts:

1. We include experimental details, results, and analysis for multiple benchmark experiments referenced in the paper.
2. We provide a more detailed explanation of the dataset annotation process.
3. We present demo.mp4, which is a comprehensive explanation of our work, including the motivation of the experiment, visualization examples of the dataset, and video visualization of relevant data.

1. Details of Benchmark Experiments

CrowdVerse supports a variety of crowd-based tasks. We evaluate the performance of existing methods on CrowdVerse in over six tasks to demonstrate the broad applicability of CrowdVerse.

Dataset	Size	Density	Annotation Numbers			Continuous Samples Included	Tasks Supported		
			Human Location	Text Words	Groups		Crowd Counting	Interaction Detection	Image Captioning
NWPU-Crowd	5.1k	<i>L, M, H</i>	2,133,238	×	×	×	✓	×	×
JHU-CROWD++	4.4k	<i>L, M, H</i>	1.51M	×	×	×	✓	×	×
ShanghaiTech	1.2k	<i>L, M, H</i>	330,165	×	×	×	✓	×	×
UCF-QNRF	1.5k	<i>H</i>	1,251,642	×	×	×	✓	×	×
GTA Head	5.1k	<i>H</i>	1,732,505	×	×	✓	✓	×	×
Venice	167	<i>H</i>	188,282	×	×	✓	✓	×	×
UCSD	32.6k	<i>L</i>	49,885	×	×	✓	✓	×	×
MOT15	11.3k	<i>L</i>	101,345	×	×	✓	△	×	×
TinyPerson	12k	<i>L</i>	72,651	×	×	×	✓	×	×
PathTrack	278k	<i>L</i>	15,000	×	×	✓	△	×	×
HollywoodHeads	224.7k	<i>L</i>	369,846	×	×	✓	✓	×	×
Mall	2k	<i>L, M</i>	62,325	×	×	✓	✓	×	×
AHU-Crowd	2.2k	<i>L, M</i>	45,000	×	×	×	✓	×	×
CityUHK-X	3.2k	<i>L, M</i>	106,783	×	×	✓	✓	×	×
SmartCity	50	<i>L</i>	369	×	×	×	✓	×	×
CUHK-SYSU	18.2k	<i>L</i>	99,809	×	×	×	△	×	×
CUHK-PEDES	40.2k	<i>L</i>	13,003	2,130,918	×	×	×	×	✓
PETA	19k	<i>L</i>	8,705	1,235,000	×	×	×	×	✓
ICFG-PEDES	54.5k	<i>L</i>	4,102	2,028,218	×	×	×	×	✓
RSTPReid	4.1k	<i>L</i>	4,101	529,029	×	×	×	×	✓
SGD	588	<i>L</i>	5415	×	1,719	×	△	✓	×
SID	716	<i>L</i>	5453	×	1,509	×	△	✓	×
CrowdVerse	1.15M	<i>L, M, H</i>	22,998,142	2,131,129	11,858	✓	✓	✓	✓

Table 1. Comparisons among the representative public dataset of crowds. *L*: low density, *M*: medium density, *H*:high density. ✓: task supported. ×: task not supported. △: task supported, but not typically used.

Methods	ShanghaiTech A		UCF-QNRF		JHU-Crowd+		NWPU Crowd		UCF_CC_50		CrowdVerse	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
GauNet [13]	61.2	97.8	84.2	152.4	69.4	262.4	-	-	215.4	296.4	65.5	295.2
MAN [53]	56.8	90.3	77.3	131.5	53.4	209.9	76.5	323.0	-	-	49.9	288.3
CLTR [51]	56.9	95.2	85.8	141.3	59.5	240.6	74.3	333.8	-	-	54.2	298.6
PET [54]	49.3	78.7	80.1	136.6	58.5	238.0	74.4	328.5	159.9	223.7	43.3	223.3
APGCC [9]	48.8	76.7	80.1	136.6	54.3	225.9	71.7	284.4	154.8	205.5	48.7	250.3

Table 2. Performance comparison of methods on crowd counting tasks on CrowdVerse and other datasets.

1.1. Crowd Counting

As shown in Table 2, crowd counting methods do not exhibit a significant performance drop on CrowdVerse compared to other datasets, indicating that existing approaches have achieved promising effectiveness in tackling challenging counting tasks. However, CrowdVerse encompasses more diverse scenarios and richer density patterns, making it a valuable comprehensive benchmark in the crowd counting domain. This enables a more thorough evaluation of crowd counting models, further advancing the field.

1.2. Trajectory Prediction

Task Description The pedestrian trajectory prediction task aims to develop a model that can predict future trajectories in crowd scenes based on observed trajectories. This task primarily faces the following three challenges: (1) The behavior of pedestrians in crowded scenes is influenced by neighboring pedestrians, thus requiring precise modeling of social interactions. (2) Each pedestrian has a specific motion pattern, moving with

Methods	ETH		HOTEL		ZARA1		ZARA2		UNIV		CrowdVerse	
	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
S-LSTM [1]	0.73	1.48	0.38	0.80	0.51	1.19	0.39	0.89	0.58	1.28	13.94	25.02
SGAN [24]	0.87	1.62	0.72	1.61	0.34	0.69	0.42	0.84	0.60	1.26	19.9	35.21
STGAT [30]	0.68	1.29	0.35	0.66	0.34	0.69	0.29	0.60	0.52	1.10	10.82	18.84
STAR [103]	0.57	1.11	0.17	0.36	0.26	0.55	0.22	0.46	0.31	0.62	6.28	11.96
SingularTrajectory [4]	0.35	0.42	0.13	0.19	0.19	0.32	0.15	0.25	0.25	0.44	10.02	14.43

Table 3. Performance comparison of methods on trajectort prediction tasks on CrowdVerse and other datasets.

different gaits, speeds, accelerations, and directions, meaning that future trajectories will be affected by individual motion characteristics. (3) Existing datasets contain only a very limited set of scenes, and some challenging cases, such as extremely dense crowd scenes, have not been included in the research scope.

Input and Output The inputs of the model are the past trajectories of pedestrians in crowd scenes, and the outputs of the model is the future trajectories of pedestrian. Mathematically, we assume that there are N pedestrians involved in a scene. The position of pedestrian i at time-step t is denoted as $p_i^t = (x_i^t, y_i^t)$. The models receive as input p_i^t of pedestrian $i = 1, 2, \dots, N$ at time-steps $t = T_1, \dots, T_{obs}$. Then it generates the predicted future position p_i^t at time-step $t = T_{obs+1}, \dots, T_{pred}$.

Benchmark Method We evaluated the performance of the following methods: first, Social LSTM [1], the first data-driven approach for the trajectory prediction task; second, Social GAN [24], the first generative model-based approach for pedestrian trajectory prediction; third, STGAT [30], one of the earliest models to adopt a graph structure in the trajectory prediction task; and finally, STAR [103], the first Transformer-based framework for trajectory prediction, which uses the attention mechanism to handle pedestrian motion patterns and social interactions. We also evaluate the current sota SingularTrajectory [4].

Evaluation Metrics Following prior works [1, 24, 30, 103], we use two error metrics: (1) Average Displacement Error (ADE): Average L2 distance between the predicted trajectories and the ground truth trajectories over all predicted time-steps. (2) Final Displacement Error (FDE): The L2 distance between the predicted destination and the ground truth destination at end of the prediction period.

Implementation Details We predict the future 12 timestep trajectories (4.8 sec) given the historical 8 timestep trajectories (3.2 sec). We train the models using Adam optimizer with a learning rate 0.001. And the implementation details of all models are consistent with the baselines.

Results and Analysis We show the performance of baseline models on CrowdVerse and two widely used public pedestrian trajectory prediction datasets ETH and UCY. The

experimental results of quantitative analysis are shown in Table 3. The overall experimental results show that the performances of baseline models on CrowdVerse are much worse than that on ETH and UCY. These quantitative experimental results indicate that CrowdVerse is a very large-scale dataset, containing a wide range of scenes and a large number of hard cases with complex interactions.

1.3. Social Interaction Detection

1.3.1. Task Description

Detecting Interaction Groups (IGs) [15, 14] has gained growing interest from the fields of computer vision, sociology and psychology. Gestalt psychology has identified several principles of perceptual groupings, such as proximity, similarity, and common fate [85]. However, most prior studies on visual group discovery are limited to interpersonal proximity and action similarity. It focuses on the consistency of group actions. Facing the same direction, queuing and other actions are considered as IGs. In this paper, we are motivated to identify Social Interaction Groups (SIGs) from single images, which is a relatively new task yet to be fully explored. As a psychological research [76] emphasised, the definition of social interaction is “*behavior that tries to influence or take into account another’s subjective experiences or intentions*”. The group formed by social interaction is called Social Interaction Group. Our work focuses more on the study of SIGs, that is, the mutual influence between intentions, such as communication, students listening to teachers in class, and the interaction between defensive players and offensive players on the court. The common pattern of these actions is that when individuals in the group produce an action or intention, others respond to it.

1.3.2. Related Work

Human Interaction Recognition (1) **Sociology-based methods:** Early research on visual perception of human interactions is commonly inspired by sociological studies. A particularly important notion is the F-formation, which are defined as the intrinsic spatial patterns that humans maintain during social interactions [34]. In practice, [17] exploited some typical arrangements of the predefined forms of social spaces to find interaction groups in static images with a Hough voting strategy [40]. [31] uses modularity cut method to estimation F-formation. [78] adopts a multi-scale method to adaptively discover the F-formation in the image. [77] proposed to detect F-formations using a computational model for clustering individuals, with the efficient graph-cut based optimisation [38]. [79] uses F-formation to find social groups in images or videos combined with estimating the person distribution in space. However, this kind of methods usually requires proxemic information such as head orientations and positions. In actual scenarios, it may not be easy to obtain directly. (2) **Action-based methods:** Recent works tend to detect interactions by action similarity. [37] fed an interaction model with the action features of individuals to identify the interactive relations between dyads. [98] combine CNN with LSTM [28] to extract the temporal and spatial features of each persons from video sequences. [91] combines action recognition with interaction inference in a unified network. The predicted interaction relations are taken as the result of interaction group discovery. introduced the action compatibility to constrain a graph network with a logic-

aware reasoning module. These methods commonly attend to a limited set of actions, which are sub-optimal for generic interaction recognition where an infinite variety of actions may take place. **(3) Other methods:** [102] identifies human interactions leveraging both geometric and social relations. However, facial information is indispensable for this model. In many crowded scenes, it may be difficult to detect faces due to occlusions. Different from these methods, inspired by the theory in [110], this paper uses the combination of position information and personal posture information to realize the detection of interaction groups in images.

Group Relation Analysis Understanding group relations is essential for interaction recognition and group discovery. Many studies of group activity recognition have explored to analyse the relations among individuals. Rather than directly recognising group activities, a common methodology is to introduce an intermediate representation referred to as structure groups [15], which models how people interact spatially. [32] proposed a hierarchical network by stacking multiple relational layers to represent interpersonal relations. [66] inferred the relations based on spatio-temporal attention and semantic graph. [95] also constructed the relation graphs to capture the underlying interactions between actors, and employed the graph convolutional network [35] with sparse temporal sampling strategy for the relational reasoning. A recent study [20] attempted to use the graph attention networks [82] for directly learning the potential interactions and meanwhile capturing the global activity context. The discovery of these relationships is more to help identify group activities, while this paper only focuses on the discovery of interaction, which is a kind of fundamental research.

1.3.3. Input and Ground Truth

Given an image and the associated b-boxes, our target is the interaction probability matrix $\mathbf{R} = (R_{ij})_{N \times N}$ (N represents the the number of people in an image), where $R_{ij} \in [0, 1]$ represents the interaction probability between the i -th and the j -th person. Naturally, \mathbf{R} should be a symmetric matrix with zeros on its diagonal. Subsequently, the social groups are determined via probability thresholding.

People belonging to the same group share the same group number. We aim to identify and analyze social interactions within this dataset using the provided social-group annotations (used for establishing the real interaction matrix) as our ground truth for training and testing.

1.3.4. Results and Analysis

We evaluate the methods using commonly employed metrics [16, 75], including precision (P), recall (R), F1-score (F1), accuracy (Acc), and the area under the curve (AUC) [25]. 4 presents the experimental results. Overall, models performed worse on CrowdVerse compared to SID, indicating that CrowdVerse presents a greater challenge. This can be attributed to its higher scene complexity, characterized by increased crowd density and greater perspective variations. The general decline in performance on methods suggests that improving model robustness and generalization in challenging environments remains a research direction.

Methods	SID					CrowdVerse				
	P	R	F1	Acc	AUC	P	R	F1	Acc	AUC
ARG [96]	41.87	54.14	47.22	69.83	0.534	38.81	38.54	38.67	89.33	0.847
JS [21]	40.02	61.62	48.53	67.41	0.596	27.73	28.40	28.06	87.29	0.772
CAGNet [92]	38.95	70.70	50.23	65.07	0.712	47.03	48.42	47.41	90.74	0.891
PEAN [104]	63.89	68.47	66.10	82.49	0.881	50.44	54.89	52.57	91.35	0.932

Table 4. Performance comparison of different methods in human interaction detection on CrowdVerse.

BLIP 2	CrowdVerse		Flickr30K	COCO
	Test	Validation	Test Set	Test Set
Image \rightarrow Text	$R@1$	0.11	0.12	0.98
	$R@5$	0.68	0.62	1.00
Text \rightarrow Image	$R@1$	0	0.12	0.90
	$R@5$	0.58	0.62	0.98

Table 5. Performance comparison of BLIP2 on image-text retrieval tasks on CrowdVerse and other datasets.

1.4. Image-text Retrieval

1.4.1. Task Description

Image-text retrieval involves two primary tasks: (1) Image-to-Text Retrieval, where the objective is to retrieve the correct textual description given an image query, and (2) Text-to-Image Retrieval, where the model retrieves the most relevant image for a given text query. These tasks enable various applications, including search engines, recommendation systems, and content moderation in media platforms.

The main challenge in image-text retrieval is the inherent heterogeneity between visual and textual modalities. This disparity requires methods to map both images and text into a joint embedding space where similarities are computed. Moreover, ambiguous or noisy annotations, as well as fine-grained cross-modal correspondences, make it challenging to align specific regions of images with specific words or phrases in descriptions.

1.4.2. Related Work

A general framework of image-text retrieval consists of three components, i.e., feature representation, cross-modal fusion, and alignment matching. Each component plays a specific role and contributes to the overall framework of image-text retrieval, which is discussed detailedly in this section. (1) **Feature representation.** Early studies mainly contributed to the capacity of feature representation by utilizing different feature extractors. Various feature extractors of visual and textual modality have been proposed [36, 55, 108, 39, 87, 63]. Among them, Kiros et al.[36, 109] introduced Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to image-text retrieval. Later, backbones such as BERT [18] and Faster R-CNN [73] are leveraged to introduce fine-grained representation with local features, i.e., word vectors and image regions. Lo-

cal feature backbones effectively filter out redundant edges within each image, laying a foundation for the subsequent matching process. However, this approach relies on the capacity of the backbone in detecting or segmenting instances for the image galleries. Therefore, for the potential image-text pairs with nearly identical annotations but different identities, fine-grained representation cannot tell them apart effectively, and the retrieval performance is highly correlated with the selected feature backbone.

(2) **Cross-modal fusion and alignment matching.** Cross-modal fusion and alignment matching are indispensable during retrieval since the goal is to measure the similarity of image-text pairs. Great efforts have been devoted to the fusion paradigms [48, 67, 62, 93, 61, 86, 8, 56, 87, 47, 12, 63, 94, 46]. Among them, Wang et al.[93] proposed to employ early fusion, followed by attention networks to make further interaction. However, early fusion triggers difficulty in converging, thus semantic inconsistency cannot be fully exploited. Therefore, late fusion [61, 63, 67, 62, 56] is proposed subsequently, where fusing operations are employed during the matching process to gradually narrow the heterogeneity. Lee et al.[39] initiated region-level alignment by proposing a stacked cross-attention framework. Li et al.[48] proposed a two-stage model to explore identity-aware representation, which is tightly coupled with a CNN-LSTM structure to match remarkable fragments. Apart from this, graph-structured methods utilize external knowledge such as commonsense [56, 87, 88, 47, 12, 46, 23, 65] and special attribute [56, 90, 99, 41, 86], further exploiting the instance relationships. Specifically, Wang et al. [88] proposed a multi-modal scene graph to jointly characterize objects and relationships. Liu et al. [56] unfold a graph-based viewpoint by presenting node-level matching and structure-level matching strategies.

1.4.3. Experimental Settings

Metrics For evaluation, we adopt Recall@K ($R@K$, $K=1, 5, 10$) as the evaluation metric, which is commonly used in image-text retrieval. Specifically, Recall@K refers to the average percentage of images where at least one corresponding identity is retrieved correctly according to ground truth among all top-K results.

1.4.4. Benchmark Methodology.

There are various image-text retrieval baselines, ranging from different backbones to vision-language pretraining methods. First we introduce some benchmark methods.

- **SCAN:** SCAN utilizes cross-attention layers to align words with image regions, capturing detailed region-word relationships. By calculating image-text similarity at a finer level, SCAN provides more accurate retrieval results compared to global feature approaches.
- **VSE++:** This baseline model maps images and text descriptions to a shared embedding space using CNNs for visual features and RNNs for text features. It trains with hard negatives within mini-batches to improve alignment precision between paired modalities.
- **UNITER:** This transformer-based model is pre-trained on large-scale image-text pairs with multiple objectives, such as Masked Language Modeling (MLM) and

Masked Region Modeling (MRM). These objectives allow UNITER to learn fine-grained interactions between image regions and text tokens, improving retrieval accuracy.

With the rapid development of vision-language pretraining (VLP), current image-text retrieval gradually shifts to larger size, which can be attributed to the following methods:

- **CLIP**: CLIP is a pioneered large-scale contrastive learning between image and text pairs, which employs dual encoders to project images and text into a shared embedding space, trained with a contrastive loss on 400M image-text pairs. The architecture demonstrates remarkable zero-shot transfer capabilities across various vision tasks.
- **ALBEF**: ALBEF leverages contrastive learning with momentum distillation to align visual and textual representations, introducing an image-text matching mechanism that enhances cross-modal understanding.
- **BLIP**: BLIP introduces a unified approach for vision-language understanding and generation, which utilizes a novel MultiModal Mixture of Encoder-Decoder architecture, incorporating bootstrap captioning and image-text feature alignment. The model synergistically combines three modules: image encoder, text encoder, and image-grounded text encoder/decoder.
- **BLIP-2**: BLIP-2 establishes an efficient bridge between vision encoders and large language models through a two-stage framework with a lightweight Q-Former serving as an interface between the vision encoder and language model, significantly reducing computational costs while maintaining performance.
- **Florence**: Florence proposes a unified vision foundation model with strong generalization capabilities, which implements a unified pretraining framework that enables effective transfer across various vision tasks, including image-text retrieval and zero-shot learning.

For the reason that the inference of our proposed dataset poses a high demand on the length of textual prompt, which quite a few off-the-shelf VLP methods do not support. Therefore, we use BLIP-2 as the benchmark method for the meta-evaluation of image-text retrieval on our proposed dataset.

R@1 (i-t)	11%
R@1 (t-i)	0%
R@3 (i-t)	22%
R@3 (t-i)	44%
R@5 (i-t)	67%
R@5 (t-i)	56%

Table 6. Performance comparison of BLIP2 on Image-text Retrieval Tasks across CrowdVerse in Scene 1.

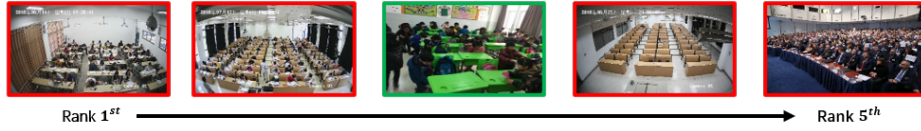
1.4.5. Results and Analysis

In the quantitative evaluation shown in Figure 6 and Figure 7, the BLIP2 model performs well on established datasets such as Flickr30K and COCO, achieving R@1 scores of 98%

R@1 (i-t)	12%
R@1 (t-i)	12%
R@3 (i-t)	38%
R@3 (t-i)	38%
R@5 (i-t)	62%
R@5 (t-i)	62%

Table 7. Performance comparison of BLIP2 on Image-text Retrieval Tasks across CrowdVerse in Scene 2.

The image depicts a classroom filled with young children, probably elementary school students, engaged in various activities. The classroom features bright green desks arranged in rows, with some desks empty and others occupied by children. Many students are seen standing near their desks, while a few are sitting and looking downward, seemingly focused on their tasks. The children are dressed in colorful winter clothing, including jackets and hats, indicating a cold weather setting. Several backpacks can be seen hanging from the backs of the chairs or on the floor, predominantly in shades of blue and pink. The walls of the classroom are decorated with educational posters and charts, likely showcasing letters, numbers, or other learning materials. Natural light streams in through large windows on one side of the room, giving it a bright and cheerful atmosphere. Overall, the scene conveys a lively and dynamic classroom environment filled with the activity of young learners.



The image depicts a large conference or seminar hall filled with a substantial audience. The focus is on the attendees, who are seated, facing forward towards a speaker or presentation. In the foreground, several people can be seen from behind, including a man in a light-colored shirt and a woman with long hair. The rows are filled with individuals of varying hair colors and styles, all seemingly engaged in the event. The backdrop features a presentation on a screen mounted on the wall, though details of the content are not visible. The venue is brightly lit, with modern lighting fixtures and a professional setup that includes multiple screens. The atmosphere appears focused, likely indicating a serious discussion or insightful presentation is taking place. Overall, the image conveys a sense of collaboration and learning among a diverse group of people.



Figure 1. Qualitative Results of Image-text Retrieval in Scene 1.

and 85% in the Image-to-Text task, and 90% and 68% in the Text-to-Image task, respectively. This indicates the model’s strong generalization capabilities on commonly used datasets. However, the model’s retrieval performance is relatively lower on CrowdVerse, with R@1 scores of 11% for Image-to-Text and 0% for Text-to-Image. This suggests that the diversity and complexity of CrowdVerse present unique challenges.

The qualitative results shown in Figure 1 and Figure 2 visually demonstrate these challenges, as seen in the retrieval outcomes. BLIP2 encounters difficulties in accurately identifying prominent elements within specific contexts, such as finish lines, traffic cones, and individuals wearing particular attire. This implies that, while BLIP2 performs well on traditional datasets, additional adaptation may be necessary to enhance its effectiveness on CrowdVerse.

1.5. Image Captioning

1.5.1. Task Description

Image captioning [84, 33, 74], as one of the fundamental cross-modal tasks, aims to automatically describe the visual content of a given image with fluent and coherent sentences.

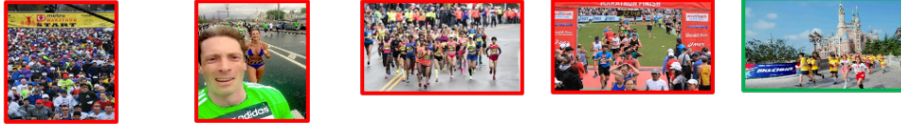
The image captures an exciting moment during a marathon event. In the foreground, a male runner wearing a black and red tank top with the number "15" prominently displayed is celebrating as he crosses the finish line. He has his arms raised high in triumph, showcasing his victory. Surrounding him are numerous other runners, some of whom are still in the process of completing the race. They exhibit a range of expressions, from determination to exhaustion, and they are dressed in various athletic gear, primarily featuring bright colors like red, blue, and white. The road is lined with orange traffic cones, possibly to guide the runners along the course. In the background, it appears to be a large crowd of spectators and additional participants, creating a vibrant and energetic atmosphere. The weather seems to be slightly overcast, suggesting damp conditions from recent rain. A time display reads "2:09:02," indicating a noteworthy finish time. Additionally, a staff member in a black jacket, possibly a race official, is positioned near the finish line, possibly holding a banner or flag that marks the conclusion of the race. The scene captures the excitement and competitive spirit often associated with marathon events.



Rank 1st

Rank 5th

The image captures an exciting moment during a marathon event. In the foreground, a male runner wearing a black and red tank top with the number "15" prominently displayed is celebrating as he crosses the finish line. He has his arms raised high in triumph, showcasing his victory. Surrounding him are numerous other runners, some of whom are still in the process of completing the race. They exhibit a range of expressions, from determination to exhaustion, and they are dressed in various athletic gear, primarily featuring bright colors like red, blue, and white. The road is lined with orange traffic cones, possibly to guide the runners along the course. In the background, it appears to be a large crowd of spectators and additional participants, creating a vibrant and energetic atmosphere. The weather seems to be slightly overcast, suggesting damp conditions from recent rain. A time display reads "2:09:02," indicating a noteworthy finish time. Additionally, a staff member in a black jacket, possibly a race official, is positioned near the finish line, possibly holding a banner or flag that marks the conclusion of the race. The scene captures the excitement and competitive spirit often associated with marathon events.



Rank 1st

Rank 5th

Figure 2. Qualitative Results of Image-text Retrieval in Scene 2.

Methods	CLIPscore \uparrow	RefCLIPscore \uparrow	B@1 \uparrow	B@4 \uparrow	METEOR \uparrow	ROUGE \uparrow	CIDER \uparrow	SPICE \uparrow	Human Score \uparrow
BLIP [45]	0.7168	0.7349	0.6752	0.0730	0.1149	0.2765	0.1020	0.0577	2.020
BLIP2 [43]	0.7547	0.7851	0.5977	0.0727	0.1129	0.2884	0.1068	0.0516	2.088
SmallCap [49]	0.6911	0.7483	0.5036	0.0493	0.0925	0.2661	0.0470	0.0440	1.804
ConZIC [106]	0.9817	0.8267	0.4016	0	0.0890	0.2152	0.0314	0.0404	1.216
LLaVa [58]	0.7464	0.7782	0.7081	0.1906	0.2015	0.4013	0.1612	0.1665	4.510
GT	0.8241	-	-	-	-	-	-	-	4.931

Table 8. Performance comparison of different methods in image captioning on CrowdVerse.

As such, it requires an algorithm to understand and model the relationships between visual and textual elements, and to generate a sequence of output words.

The primary challenges for current image captioning tasks are as follows:

- **Flexibility:** Image captioning tasks require models to comprehend complex visual contexts, including ever-changing diverse objects and detailed relationships between them, ensuring that the model possesses a strong ability to update knowledge for zero-shot inference.
- **Relevance:** The primary challenge in vision-language tasks lies in matching both coarse-grained and fine-grained details across the two modalities, as the relationship between images and text is not a simple pixel-to-word mapping but a multi-faceted correspondence involving both local and global alignments.
- **Efficiency:** To ensure generalizability, multimodal models typically rely on large parameters and huge-scale training datasets [5, 89, 11], which require extensive demands on computational resources. Therefore, ensuring efficiency is a critical but difficult challenge.

1.5.2. Related Work [42]

Image captioning Image captioning aims to describe the contents of a given image. It can be roughly divided into two approaches: non-LLMs-based methods and LLMs-based ones. The former approaches [97, 2] typically employ a visual encoder and a language decoder in an end-to-end fashion to generate captions. However, they are incapable of describing open-world objects. The latter one leverages pre-trained large-scale vision models (CLIP [69], ViT [19]) and LLMs (GPTs [68, 7], T5 [70], LLaMA [80]) by bridging the gap between two modalities using either pre-training with large-scale data or the learned mapper or prompt techniques. LLMs-based models [60, 44, 11, 10] demonstrate advancements in image captioning challenges, allowing the capacity to describe anything as long as pre-trained vision models can recognize it.

Novel object captioning Novel object captioning is a branch of image captioning that describes images containing objects that were not seen during training. Non-LLMs-based methods explore more objects by learning from unpaired image-sentence sources (DCC [26], NOC [83]) or rely on novel object detectors to recognize novel concepts (NBT [59], OSCAR [50] and VinVL [107]). LLMs-based methods such as ViECap [22] leverage the pre-trained CLIP [69] to obtain object entities. Nevertheless, the cut-off in training time of the pre-trained object detector or CLIP prevents it from detecting novel objects that arise quickly in reality.

1.5.3. Results and Analysis

Retrieval-augmented image captioning Retrieval-augmented image captioning is a recently popular approach that augments the captioning model with retrieved information for better open-world understanding. AoANet [29] uses a memory bank of image-sentence pairs and target words. SmallCap [72] employs image-to-text retrieval to obtain sampled captions from a captions datastore. RA-CM3 [101] retrieves documents from an external memory of a mixture of text and image via a dense multimodal retriever. EXTRA and Re-ViLM [100] exploit the similarity of the input image and vision candidates to retrieve captions.

1.5.4. Experimental Settings

Metrics For evaluation, we utilize two methods to both qualitatively and quantitatively verify the quality of the captions: (1) GPT-4o and human as the judge, to rate each generated caption according to designed prompt and instruction, (2) CLIPscore, RefCLIPscore and other reference-based metrics: matching score between the generated caption and ground-truth image/text.

RefCLIPScore [27] measures the semantic similarity between references and predictions. Besides, we also use an unsupervised metric, CLIPScore [27]. CLIPScore is a reference-free metric measuring the similarity between an image and the corresponding caption, which is the most critical metric for zero-shot IC.



BLIP : people walking around a plaza with a clock on the side of it. (GPT Score: 2, Human Score: 1)

BLIP2 : a view of a courtyard from above. (GPT Score: 3, Human Score: 2.5)

SMALLCAP : a group of people walking down a sidewalk. (GPT Score: 2, Human Score: 2)

ConZIC : image of a walk connecting campus building from movie about two bikes. (GPT Score: 1, Human Score: 1)

LLaVa : The image depicts a large, open area with a cement ground, surrounded by trees and benches. There are several people walking around and enjoying the space, with some of them carrying handbags. The area appears to be a park or a public space, where people can relax and socialize. In addition to the people, there are a few birds scattered throughout the scene, adding to the lively atmosphere. The overall scene is bustling with activity, as people move around and interact with one another in this outdoor space. (GPT Score: 3, Human Score: 4.5)

GT : The image depicts an outdoor area, likely a plaza or courtyard, characterized by a wide, tiled ground surface. There are several people walking across the space, with some gathered in small groups. In the foreground, a lone figure is walking, casting a long shadow. Trees with bare branches are visible in planters, suggesting it may be early spring or late fall. To the left, a rocky structure or garden feature is evident, indicating landscaping efforts. The greenery, including shrubs and patches of grass, surrounds the area, providing a natural feel. Some individuals are standing near the building on the right side of the image, which appears modern with large glass windows. The overall atmosphere seems peaceful, with a balance of urban design and natural elements. Sunlight illuminates the scene, enhancing the clarity and brightness of the surroundings. (GPT Score: 5, Human Score: 5)



BLIP : people walking down a street with a lot of people wearing masks. (GPT Score: 2, Human Score: 2.5)

BLIP2 : a group of people wearing masks and black clothing walking down the street. (GPT Score: 3, Human Score: 2.5)

SMALLCAP : a group of people walking down a street. (GPT Score: 2, Human Score: 2.5)

ConZIC : image of a male combat party near athens olympic village reuters website. (GPT Score: 1, Human Score: 1)

LLaVa : The image depicts a large group of people gathered on a street, with many of them wearing masks. The crowd appears to be protesting or participating in some kind of event, as they are all dressed in black and some are carrying backpacks. There are at least 14 people visible in the scene, with some standing closer to the foreground and others further in the background. The masks worn by the individuals vary in design and style, adding to the diverse nature of the crowd. The backpacks can be seen on the backs of several individuals, possibly containing supplies or belongings for the event. (GPT Score: 4, Human Score: 5)

GT : The image depicts a group of individuals participating in a protest or riot. Many of them are wearing masks or bandanas covering their faces, with a notable number wearing gas masks, suggesting they are preparing for the use of tear gas or other crowd control measures. The people are dressed in casual clothing, with some wearing black shirts, cargo pants, and shorts. The scene appears chaotic, as the group is moving down a street amidst debris, including scattered plastic bottles and other refuse. Some individuals appear to be holding objects that could be used as weapons, such as sticks or poles. The background features urban elements, including public transportation kiosks and street signs, which indicate that the incident is taking place in a city environment. Overall, the atmosphere conveys tension and unrest. (GPT Score: 5, Human Score: 5)

Figure 3. Qualitative Analysis of Image Captioning.

For the purpose of measuring the Fluency, Precision and Recall of generated captions from different perspectives, we use the full set of captioning metrics entailing BLEU [64], METEOR [6], ROUGE [52], CIDEr [81] and SPICE [3]. Among them, CIDEr is more emphasized in image captioning because it can capture key information.

1.5.5. Benchmark Methodology

We compare with several SOTAs. According to the trainable parameters size, they can be divided into 1) Vision-Language Pre-train Model based on the Transformer: BLIP [45], BLIP-2 [44]. Through pretraining, they extensively learn feature representations and cross-modal interactions between vision and language, and are then fine-tuned for a variety of downstream tasks; 2) Retrieval-Augmented Lightweight Model: SmallCap [71], which generates a caption conditioned on an input image and related captions retrieved from a datastore to enhance generalization ability, demonstrating high utility in image captioning; 3) Training-free Controllable Model: ConZIC [105], with a novel sampling-based non-autoregressive language model named Gibbs-BERT, which can generate and continuously polish every word; and also 4) Multi-modal Large Language Model: LLaVA [57], an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language, using language-only GPT-4 to generate

multimodal language-image instruction-following data. Table 8 and Figure 3 present the

The image features a young woman posed against a bright yellow background. She has long, slightly wavy brown hair that frames her face, with lighter highlights adding dimension. Her shirt is a floral pattern, predominantly in soft colors such as blue, pink, and white, with a collar and three-quarter length sleeves. The shirt has a relaxed fit and a small bow tied at the waist, accentuating her silhouette. She is wearing fitted blue jeans that are paired with the shirt. Her expression is confident with a slight smile, and she stands with one hand on her hip and the other held slightly out, creating a casual yet poised demeanor. The vibrant yellow wall serves as a striking backdrop, contrasting nicely with her attire and enhancing the overall cheerful and lively vibe of the image.



DALLE



GT: 5843

The image depicts a lively street scene during twilight, likely in a bustling urban area. The cobblestone pavement is laid out in intricate patterns, adding a decorative element to the environment. On either side of the street, there are numerous outdoor dining spaces with tables covered in orange tablecloths, where people are enjoying their meals. A diverse crowd can be seen at the tables and walking along the street; some individuals appear to be engaged in conversation while others enjoy their food. The atmosphere is vibrant, illuminated by warm, golden street lamps that enhance the evening ambiance. Buildings line the street, showcasing a mix of architectural styles, and most have large windows with glowing lights, indicating that they are occupied. One establishment is marked as "Cafeteria S. Nicolau," and there are more food-related windows displaying treats and delicacies. Additional pedestrians are visible, with some forming a line outside of a shop, suggesting popularity. The overall scene conveys a festive and welcoming vibe, typical of a lively dining district in a city, rich in culture and social interaction.



DALLE



GT: 11519

Figure 4. Qualitative results of Image-text generation.

results of different models on the CrowdVerse dataset. From the quantitative analysis, it is evident that the Conzic model achieves higher scores on CLIP due to its continual polishing towards higher CLIP scores during the generation process. However, ConZIC’s performance on traditional evaluation metrics, such as BLEU, METEOR, and CIDEr, is significantly lower than that of other models, reflecting its limitations in capturing common natural language patterns. In terms of human ratings, the ground truth (GT) scores are significantly higher, and qualitative analysis further elucidates this difference. Models like BLIP and LLaVA can generate content that aligns with human descriptions to some extent, but they often fail to fully capture scene details or subtle contextual cues, such as crowd dynamics and specific activity descriptions. For instance, when describing scenes containing crowds, descriptions generated by models like ConZIC and SmallCap miss critical information such as “black shirts, masks, and sticks,” which are effectively captured in the GT descriptions. This indicates that while current models can produce generally relevant content, they still face challenges in capturing complex scenes rich in context. There remains a gap between the descriptions generated by models and those at human levels.

1.6. Image Text Generation

1.6.1. Implementation Details

We randomly selected a subset of samples from the dataset and generated images based on the textual descriptions. The generated images were quantitatively compared with the real images, and the quality of the generated images was evaluated by image-text matching scores.

1.6.2. Results and Analysis

The results of the text-to-image generation demonstrate that the generated images are highly similar to the real images in terms of layout, color, objects, actions, and spatial relationships. This indicates that our annotated caption effectively aids the model in reconstructing the images, capturing nearly all global and local visual information.

Scenes	Backgrounds of images in CrowdVerse
Indoor	Shopping mall, station, stairs, escalator, platform, classroom, canteen, runway, indoor stadium, stage, gymnasium, store, supermarket, training room, house, laboratory ...
Outdoor	Street, square, sidewalk, crossing, alley, courtyard, park, overpass, amusement park, bridge, playground, track, seaside, beach, stadium, forest, football field, basketball court, snow field ...

Table 9. The display of diverse backgrounds in CrowdVerse.

2. Annotation Details



Figure 5. The labeling interface of LabelImg.

We further elaborate on the details of the dataset annotation. During the data collection phase, we reviewed most of the current crowd-related datasets and annotated their label details. We compiled detailed dataset labels for those containing images. It was observed that there are numerous inconsistencies in the labeling across existing datasets. Therefore, we first integrated the dataset formats.

We downloaded almost all available crowd datasets and performed a selection process. For image-format data, we loaded the corresponding labels into LabelImg for manual fine-tuning. Figure 5 shows the interface, and upon inspection, we found significant errors in the labels of many datasets. For example, the labels regarding the heads of crowds in Figure 5 contained many omissions and mistakes. To address this, we invited professional annotators to manually correct these labels to ensure they more accurately re-

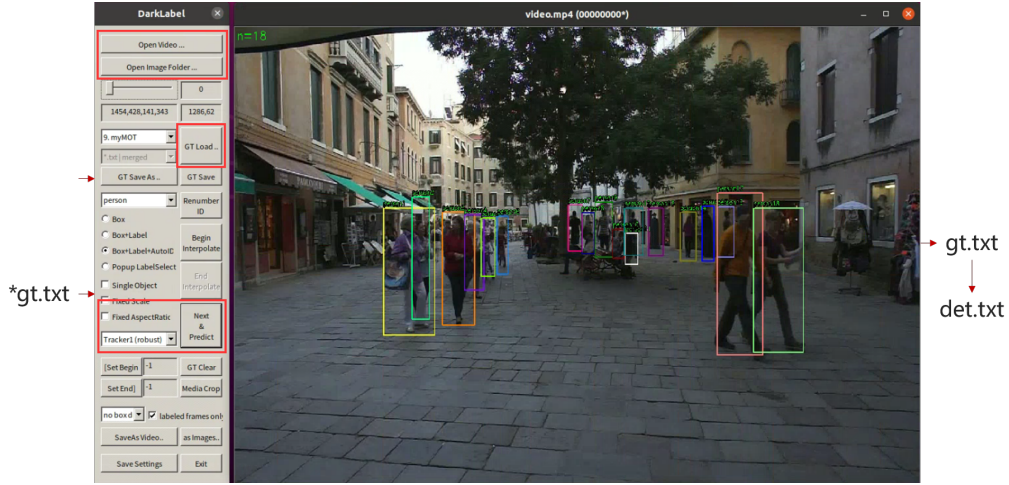


Figure 6. The labeling interface of DarkLabel.

flected reality. For video or images from continuous time frames, we used the DarkLabel tool for annotation, which also corrected many irregular labels.

Additionally, we collected new datasets in two ways: using a 48-megapixel sensor camera that supports 4K60P HDR video and 4K100P slow-motion shooting, specifically for gathering new crowd data. For images captured from normal perspectives, we used manual shooting, considering various environmental factors. For top-down perspective images, drones were used, with the shooting equipment supporting a horizontal 90° and vertical 72° field of view, a maximum horizontal flight speed of 16 meters per second, and a maximum take-off altitude of 4,000 meters, ensuring the collection of multi-scale crowd images.

For unlabeled images, we followed the steps mentioned in the main text, first performing coarse-grained annotation, then fine-grained annotation. CrowdVerse ensured the consistency of annotation formats. Additionally, we categorized the tasks and provided dataset versions with a single label format for each specific task.

References

- [1] Alexandre Alahi et al. “Social lstm: Human trajectory prediction in crowded spaces”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 961–971.
- [2] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [3] Peter Anderson et al. “SPICE: Semantic Propositional Image Caption Evaluation”. In: *ECCV (5)*. Vol. 9909. Lecture Notes in Computer Science. Springer, 2016, pp. 382–398.
- [4] Inhwon Bae, Young-Jae Park, and Hae-Gon Jeon. “Singulartrajectory: Universal trajectory predictor using diffusion model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17890–17901.
- [5] Jinze Bai et al. “Qwen-vl: A frontier large vision-language model with versatile abilities”. In: *arXiv preprint arXiv:2308.12966* (2023).
- [6] Satanjeev Banerjee and Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. In: *IEEevaluation@ACL*. Association for Computational Linguistics, 2005, pp. 65–72.
- [7] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.
- [8] Hui Chen et al. “Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 12655–12663.
- [9] I-Hsiang Chen et al. “Improving point-based crowd counting and localization based on auxiliary point guidance”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 428–444.
- [10] Xi Chen et al. “PaLI-X: On Scaling up a Multilingual Vision and Language Model”. In: *arXiv preprint arXiv:2305.18565* (2023).
- [11] Xi Chen et al. “PaLI: A Jointly-Scaled Multilingual Language-Image Model”. In: *International Conference on Learning Representations (ICLR)*. 2023.
- [12] Yuhao Cheng et al. “Cross-modal graph matching network for image-text retrieval”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18.4 (2022), pp. 1–23.

- [13] Zhi-Qi Cheng et al. “Rethinking spatial invariance of convolutional networks for object counting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 19638–19648.
- [14] Wongun Choi and Silvio Savarese. “A unified framework for multi-target tracking and collective activity recognition”. In: *Proceedings of the European conference on computer vision (ECCV)*. Springer. 2012, pp. 215–230.
- [15] Wongun Choi et al. “Discovering groups of people in images”. In: *Proceedings of the European conference on computer vision (ECCV)*. Springer. 2014, pp. 417–433.
- [16] Cyril Cleverdon, Jack Mills, and Michael Keen. “Factors determining the performance of indexing systems”. In: *(No Title)* (1966).
- [17] Marco Cristani et al. “Social interaction discovery by statistical analysis of Formations.” In: *BMVC*. Vol. 2. Citeseer. 2011, p. 4.
- [18] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [19] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *ICLR*. OpenReview.net, 2021.
- [20] Mahsa Ehsanpour et al. “Joint learning of social groups, individuals action and sub-group activities in videos”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2020, pp. 177–195.
- [21] Mahsa Ehsanpour et al. “Joint learning of social groups, individuals action and sub-group activities in videos”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 177–195.
- [22] Junjie Fei et al. “Transferable decoding with visual entities for zero-shot image captioning”. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2023.
- [23] Zheren Fu et al. “Learning semantic relationship among instances for image-text matching”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15159–15168.
- [24] Agrim Gupta et al. “Social gan: Socially acceptable trajectories with generative adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2255–2264.

- [25] James A Hanley and Barbara J McNeil. “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” In: *Radiology* 143.1 (1982), pp. 29–36.
- [26] Lisa Anne Hendricks et al. “Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
- [27] Jack Hessel et al. “CLIPScore: A Reference-free Evaluation Metric for Image Captioning”. In: *EMNLP (1)*. Association for Computational Linguistics, 2021, pp. 7514–7528.
- [28] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [29] Lun Huang et al. “Attention on Attention for Image Captioning”. In: *ICCV*. IEEE, 2019, pp. 4633–4642.
- [30] Yingfan Huang et al. “Stgat: Modeling spatial-temporal interactions for human trajectory prediction”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6272–6281.
- [31] Hayley Hung and Ben Kröse. “Detecting f-formations as dominant sets”. In: *Proceedings of the 13th international conference on multimodal interfaces*. 2011, pp. 231–238.
- [32] Mostafa S Ibrahim and Greg Mori. “Hierarchical relational networks for group activity recognition and retrieval”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 721–736.
- [33] Andrej Karpathy and Li Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *CVPR*. IEEE Computer Society, 2015, pp. 3128–3137.
- [34] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive, 1990.
- [35] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *International Conference on Learning Representations (ICLR)*. 2017.
- [36] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. “Unifying visual-semantic embeddings with multimodal neural language models”. In: *arXiv preprint arXiv:1411.2539* (2014).

- [37] Yu Kong, Yunde Jia, and Yun Fu. “Interactive phrases: Semantic descriptions for human interaction recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.9 (2014), pp. 1775–1788.
- [38] L’ubor Ladický et al. “Inference methods for crfs with co-occurrence statistics”. In: *International journal of computer vision* 103.2 (2013), pp. 213–225.
- [39] Kuang-Huei Lee et al. “Stacked cross attention for image-text matching”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 201–216.
- [40] Bastian Leibe, Ales Leonardis, and Bernt Schiele. “Combined object categorization and segmentation with an implicit shape model”. In: *Workshop on statistical learning in computer vision, ECCV*. 2004, pp. 17–32.
- [41] Jiangtong Li, Li Niu, and Liqing Zhang. “Action-aware embedding enhancement for image-text retrieval”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1323–1331.
- [42] Jiaxuan Li et al. “EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension”. In: *CoRR* abs/2311.15879 (2023).
- [43] Junnan Li et al. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.
- [44] Junnan Li et al. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *Proc. International conference on machine learning (ICML)*. 2023.
- [45] Junnan Li et al. “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12888–12900.
- [46] Kunpeng Li et al. “Image-text embedding learning via visual and textual semantic reasoning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022), pp. 641–656.
- [47] Kunpeng Li et al. “Visual semantic reasoning for image-text matching”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4654–4662.
- [48] Shuang Li et al. “Identity-aware textual-visual matching with latent co-attention”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1890–1899.

- [49] Shuang Li et al. “Person search with natural language description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1970–1979.
- [50] Xiujun Li et al. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *ECCV (30)*. Vol. 12375. Lecture Notes in Computer Science. Springer, 2020, pp. 121–137.
- [51] Dingkan Liang, Wei Xu, and Xiang Bai. “An end-to-end transformer model for crowd localization”. In: *European Conference on Computer Vision*. Springer, 2022, pp. 38–54.
- [52] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.
- [53] Hui Lin et al. “Boosting crowd counting via multifaceted attention”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 19628–19637.
- [54] Chengxin Liu et al. “Point-query quadtree for crowd counting, localization, and more”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 1676–1685.
- [55] Chunxiao Liu et al. “Focus your attention: A bidirectional focal attention network for image-text matching”. In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 3–11.
- [56] Chunxiao Liu et al. “Graph structured network for image-text matching”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10921–10930.
- [57] Haotian Liu et al. *Visual Instruction Tuning*. 2023. arXiv: 2304.08485 [cs.CV]. Available from: <https://arxiv.org/abs/2304.08485>.
- [58] Haotian Liu et al. “Visual instruction tuning”. In: *Advances in neural information processing systems* 36 (2024).
- [59] Jiasen Lu et al. “Neural baby talk”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [60] Ron Mokady, Amir Hertz, and Amit H. Bermano. “ClipCap: CLIP Prefix for Image Captioning”. In: *CoRR* abs/2111.09734 (2021).

- [61] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. “Dual attention networks for multimodal reasoning and matching”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 299–307.
- [62] Renjie Pan et al. “Joint Intra & Inter-Grained Reasoning: A New Look Into Semantic Consistency of Image-Text Retrieval”. In: *IEEE Transactions on Multimedia* (2023).
- [63] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. “Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19275–19284.
- [64] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *ACL*. ACL, 2002, pp. 311–318.
- [65] Khoi Pham et al. “Composing object relations and attributes for image-text matching”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14354–14363.
- [66] Mengshi Qi et al. “stagnet: An attentive semantic rnn for group activity recognition”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 104–120.
- [67] Leigang Qu et al. “Dynamic modality interaction modeling for image-text retrieval”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 1104–1113.
- [68] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [69] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.
- [70] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of Machine Learning Research (JMLR)* 21.1 (2020), pp. 5485–5551.
- [71] Rita Ramos et al. “SmallCap: Lightweight image captioning prompted with retrieval augmentation”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.
- [72] Rita Ramos et al. “Smallcap: Lightweight Image Captioning Prompted with Retrieval Augmentation”. In: *CVPR*. IEEE, 2023, pp. 2840–2849.

- [73] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).
- [74] Steven J. Rennie et al. “Self-Critical Sequence Training for Image Captioning”. In: *CVPR*. IEEE Computer Society, 2017, pp. 1179–1195.
- [75] C van Rijsbergen. “Information retrieval 2nd ed buttersworth”. In: *London [Google Scholar]* 115 (1979).
- [76] R.J. Rummel. *Understanding conflict and war*. Sage Publications, ISBN 0470745010., 1981.
- [77] Francesco Setti et al. “F-formation detection: Individuating free-standing conversational groups in images”. In: *PloS one* 10.5 (2015), e0123783.
- [78] Francesco Setti et al. “Multi-scale F-formation discovery for group detection”. In: *2013 IEEE International Conference on Image Processing*. IEEE. 2013, pp. 3547–3551.
- [79] Hyun Soo Park and Jianbo Shi. “Social saliency prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4777–4785.
- [80] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).
- [81] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. “CIDEr: Consensus-based image description evaluation”. In: *CVPR*. IEEE Computer Society, 2015, pp. 4566–4575.
- [82] Petar Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations*. 2018.
- [83] Subhashini Venugopalan et al. “Captioning images with diverse objects”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
- [84] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *CVPR*. IEEE Computer Society, 2015, pp. 3156–3164.
- [85] Johan Wagemans et al. “A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization.” In: *Psychological bulletin* 138.6 (2012), p. 1172.

- [86] Haoran Wang et al. “CODER: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 700–716.
- [87] Haoran Wang et al. “Consensus-aware visual-semantic embedding for image-text matching”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer. 2020, pp. 18–34.
- [88] Sijin Wang et al. “Cross-modal scene graph matching for relationship-aware image-text retrieval”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020, pp. 1508–1517.
- [89] Weihan Wang et al. “CogVLM: Visual Expert for Pretrained Language Models”. In: *arXiv preprint arXiv:2311.03079* (2023).
- [90] Zhe Wang et al. “Vitaa: Visual-textual attributes alignment in person search by natural language”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer. 2020, pp. 402–420.
- [91] Zhenhua Wang et al. “Consistency-Aware Graph Network for Human Interaction Understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13369–13378.
- [92] Zhenhua Wang et al. “Consistency-aware graph network for human interaction understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13369–13378.
- [93] Zihao Wang et al. “Camp: Cross-modal adaptive message passing for text-image retrieval”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5764–5773.
- [94] Xi Wei et al. “Multi-modality cross attention network for image and sentence matching”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10941–10950.
- [95] Jianchao Wu et al. “Learning actor relation graphs for group activity recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9964–9974.
- [96] Yiling Wu et al. “Learning fragment self-attention embeddings for image-text matching”. In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 2088–2096.

- [97] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *Proc. International conference on machine learning (ICML)*.
- [98] Yichao Yan, Bingbing Ni, and Xiaokang Yang. “Predicting Human Interaction via Relative Attention Model”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 2017, pp. 3245–3251.
- [99] Xinxin Yang, Renjie Pan, and Hua Yang. “Adaptive and Collaborative Multi-scale Alignment for Text-Based Person Search”. In: *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE. 2023, pp. 1–5.
- [100] Zhuolin Yang et al. “Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning”. In: *EMNLP (Findings)*. Association for Computational Linguistics, 2023, pp. 11844–11857.
- [101] Michihiro Yasunaga et al. “Retrieval-augmented multimodal language modeling”. In: *International Conference on Learning Representations (ICLR)*. 2023.
- [102] Haanju Yoo et al. “Detection of interacting groups based on geometric and social relations between individuals in an image”. In: *Pattern Recognition* 93 (2019), pp. 498–506.
- [103] Cunjun Yu et al. “Spatio-temporal graph transformer networks for pedestrian trajectory prediction”. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* 16. Springer. 2020, pp. 507–523.
- [104] Jiaqi Yu et al. “Psychology-Guided Environment Aware Network for Discovering Social Interaction Groups from Videos”. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.8 (2024), pp. 1–23.
- [105] Zequn Zeng et al. “ConZIC: Controllable Zero-shot Image Captioning by Sampling-Based Polishing”. In: *CVPR*. IEEE, 2023, pp. 23465–23476.
- [106] Zequn Zeng et al. “Conzic: Controllable zero-shot image captioning by sampling-based polishing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 23465–23476.
- [107] Pengchuan Zhang et al. “VinVL: Revisiting Visual Representations in Vision-Language Models”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

- [108] Ying Zhang and Huchuan Lu. “Deep cross-modal projection learning for image-text matching”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 686–701.
- [109] Zhedong Zheng et al. “Dual-path convolutional image-text embeddings with instance loss”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.2 (2020), pp. 1–23.
- [110] Chen Zhou et al. “A social interaction field model accurately identifies static and dynamic social groupings”. In: *Nature human behaviour* 3.8 (2019), pp. 847–855.