This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

CrowdVerse: A Bidirectional Reality-Calibrated Benchmark for Crowd Understanding and Simulation

Pingrui Lai[1] Yanshan Zhou[1] Zihao Xie[1] Hua Yang[1] *
[1] School of Information Science and Electronical Engineering & School of Integrated Circuits, Shanghai Jiao Tong University, Shanghai, China
laipingrui@sjtu.edu.cn shanshan333@sjtu.edu.cn xzh031202@sjtu.edu.cn hyang@sjtu.edu.cn

Abstract

Existing crowd datasets have led to progress in crowd understanding tasks. However, datasets specifically designed for crowd simulation remain scarce, making it challenging to evaluate the realism of simulated crowd movements in a reference-free manner. To address this limitation, we introduce CrowdVerse, a bidirectional dataset designed to simultaneously advance both crowd understanding and crowd simulation research. CrowdVerse is the largest real-world crowd dataset to date, comprising over one million annotated images. These annotations cover a wide range of information, including spatial locations, trajectories, crowd density, interaction groups, and textual descriptions, thereby supporting a broad spectrum of research tasks. Based on CrowdVerse, we construct an evaluation framework for reference-free analysis of crowd movements, capable of effectively assessing the alignment between simulated crowd behaviors and real human motion patterns. The introduction of CrowdVerse represents a foundational step toward advancing research on collective intelligence. The code and dataset are available at: https://github.com/longziyu/CrowdVerse

Table 1. Comparisons among the representative dataset of crowds. L: low density, M: medium density, H:high density.

Dataset	Size	Density	Annotation Numbers		
			Human Locations	Text Words	Groups
NWPU-Crowd [52]	5.1k	$L\ M\ H$	2.13M	×	×
JHU-CROWD++ [46]	4.4k	$L\ M\ H$	1.51M	×	×
ShanghaiTech [66]	1.2k	$L\ M\ H$	330k	×	×
UCF-QNRF [26]	1.5k	H	1.25M	×	×
GTA Head [69]	5.1k	H	1.73M	×	×
Venice [40]	167	H	188k	×	×
UCSD [5]	32.6k	L	49.9k	×	×
MOT15 [11]	11.3k	L	101k	×	×
TinyPerson [62]	12k	L	72.7k	×	×
PathTrack [42]	278k	L	15k	×	×
HollywoodHeads [50]	224.7k	L	369.8k	×	×
Mall [7]	2k	$L\ M$	62.3k	×	×
AHU-Crowd [22]	2.2k	$L\ M$	45k	×	×
CityUHK-X [27]	3.2k	$L\ M$	106.8k	×	×
SmartCity [65]	50	L	369	×	×
CUHK-SYSU [56]	18.2k	L	99.8k	×	×
CUHK-PEDES [33]	40.2k	L	13k	2.1M	×
PETA [13]	19k	L	8.7k	1.2M	×
ICFG-PEDES [14]	54.5k	L	4.1k	2.0M	×
RSTPReid [71]	4.1k	L	4,101	529k	×
SGD [10]	588	L	5415	×	1.7k
SID [60]	716	L	5453	×	1.5k
CrowdVerse (Ours)	**1.15M**	$L\ M\ H$	**23.0M**	**2.13M**	**11.9k**

1. Introduction

The study of crowds is of great significance for analyzing human collective intelligence, and a real-world crowd dataset serves as the foundation for crowd research. Crowd research involves two complementary directions. The forward process, referred to as crowd understanding, involves analyzing various aspects of a given crowd, such as pedestrian positions, velocities, postures, and interaction characteristics. The reverse process, crowd simulation, focuses on generating crowds that exhibit realistic movement patterns in an environment.

However, in recent years, the progress of crowd research has been relatively slow. Although many methods have achieved satisfactory results on existing datasets, there remains a substantial gap between these achievements and a true understanding of the principles governing crowd dynamics. This phenomenon can be largely attributed to the incompleteness of current crowd datasets. As illustrated in Figure 1, most datasets for crowd understanding tasks suffer from data scarcity. They are typically small in scale, containing only a few thousand samples, and are often designed for specific tasks with limited scene diversity. Such limitations hinder the robustness of many crowd understanding approaches. With the rapid advancement of deep learning, training increasingly large models on such limited data inevitably leads to overfitting. For example, a trajectory pre-

*Corresponding author.

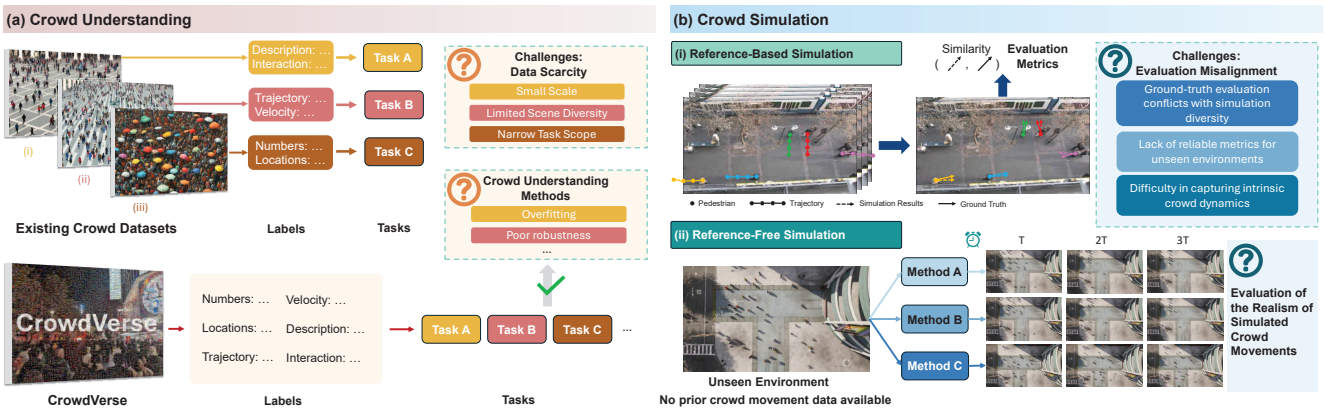


Figure 1. Overview of CrowdVerse and the research landscape in crowd analysis. (a) Existing datasets with limited diversity constrain understanding tasks and cause overfitting. (b) Reference-based simulations face evaluation misalignment, while reference-free simulations lack reliable metrics for assessing realism in unseen environments.

diction model may perform well on one type of dataset but fail to generalize to other environments or datasets with different crowd densities. In crowd simulation tasks, existing methods commonly suffer from an evaluation misalignment problem. The goal of crowd simulation is to generate diverse yet realistic crowd behaviors, while the key challenge lies in how to effectively evaluate the realism of simulated crowd movements. Current approaches typically adopt reference-based evaluation methods, in which simulated trajectories are compared with real trajectories under the same initial conditions to measure similarity. However, this evaluation paradigm is misaligned with the goal of diversity in simulation, as real crowd movements are inherently non-unique. Moreover, when generating simulations in previously unseen environments, it becomes difficult to assess the quality of the results through such ground-truth-based evaluation.

Evaluating the realism of simulated crowd behaviors under reference-free conditions is a fundamental challenge, as it inherently relies on a thorough understanding of the underlying principles of human collective intelligence. Similarly to how ImageNet [12] established benchmarks for reference-free image realism through extensive real-world data, evaluating the realism of simulated crowds requires a large-scale real-world crowd dataset that closely aligns with real-world crowd movement patterns. To address these challenges, we present a bidirectional crowd dataset: CrowdVerse, the largest crowd dataset to date, contains over 1.15 million annotated crowd images with multiple labels, including positions, trajectories, density levels, interaction groups, and textual descriptions. For crowd understanding tasks, the diversity of these labels enables CrowdVerse to support a broader range of tasks, and we report performance indicators across various crowd understanding benchmarks. For crowd simulation, CrowdVerse is a large-scale dataset, establishing a foundational benchmark for assessing crowd realism. Building on this inspiration, we propose two reference-free evaluation metrics: Crowd Inception Score (CIS) and Crowd Fréchet Inception Distance (CFID), specifically designed to evaluate the authenticity of simulated crowds. Experiments demonstrate that this framework effectively quantifies realism, establishing a foundation for research in crowd simulation. Our contributions can be summarized as follows:

- We propose a bidirectional dataset CrowdVerse, the largest crowd dataset to support the training and evaluation of various crowd-related tasks, establishing a reference-free realism benchmark for crowd simulation.
- We design a reference-free evaluation framework for quantifying crowd simulation, validated on CrowdVerse to demonstrate its effectiveness in measuring crowd dynamic behavior realism.
- In CrowdVerse, over 1.15 million crowd images are manually annotated with diverse labels, including crowd locations, trajectories, interaction groups, and textual descriptions. We also report benchmark results on various crowd-understanding tasks.

2. Related Work

2.1. Human Crowd Dataset

Existing crowd datasets primarily focus on task-specific objectives. As summarized in Table 1, we mainly investigate related works in three domains:

Localization and Counting Datasets can be classified by capture perspectives: (1) Surveillance view: Early works like UCSD [5], Mall [7], and SmartCity [65] focused on fixed scenes. Later data sets enhanced diversity (WorldExpo 10 [64]), temporal dynamics (Venice [40], Fudan-ShanghaiTech [17]), and environmental variations

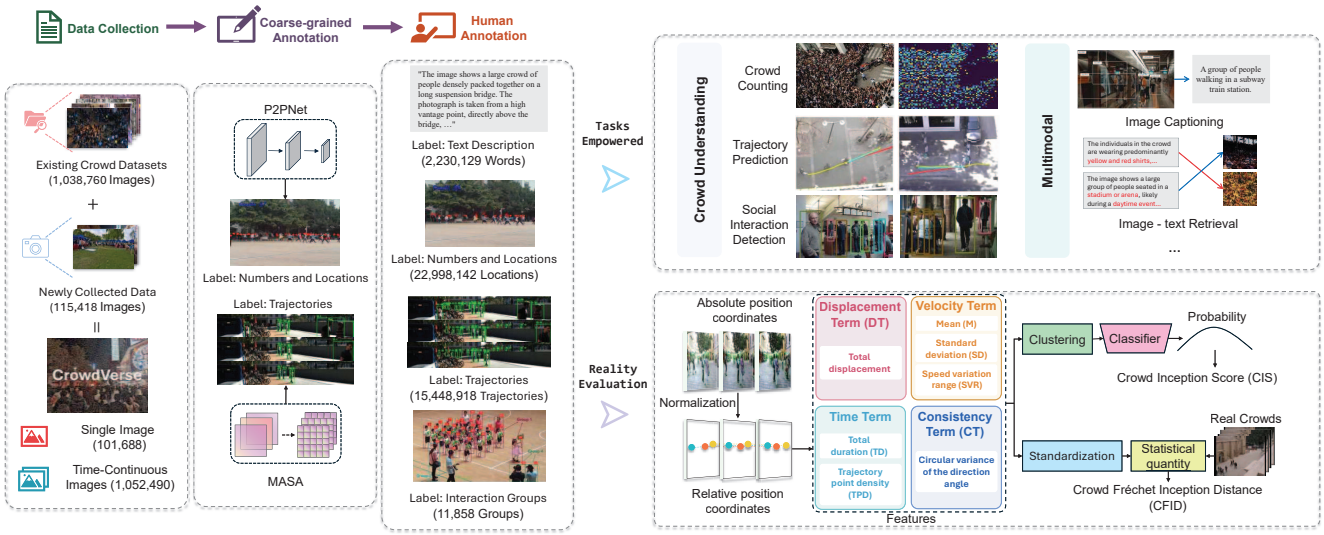


Figure 2. Overall architecture of CrowdVerse. The construction of CrowdVerse involves integrating existing datasets and expanding with newly collected data. The annotation process includes coarse-grained labeling by models, followed by fine-grained human annotations, resulting in rich labels that support various crowd-related tasks. Furthermore, CrowdVerse serves as a benchmark for evaluating crowd realism, and the simulation assessment framework built upon it effectively measures the authenticity of crowd movements.

(ShanghaiTech Part B [66], NWPU-Crowd [52]). (2) Drone view: DLR-ACD [4] and DroneCrowd [54] capture crowds with aerial perspectives. (3) General view: The internet-sourced datasets evolved from small-scale (UCF_CC_50 [25], ShanghaiTech Part A [66]) to large-scale (UCF-QNRF [26], Crowd Surveillance [58]), with enriched annotations (JHU-CROWD [46]).

Human Interaction Volleyball Dataset [24] and Collective Activity Dataset [9] analyze collective actions, SALSA [2] captures multimodal social interactions, while SGD [10] and SID [60] propose structured group representations. These are prone to small scales and limited environmental diversity.

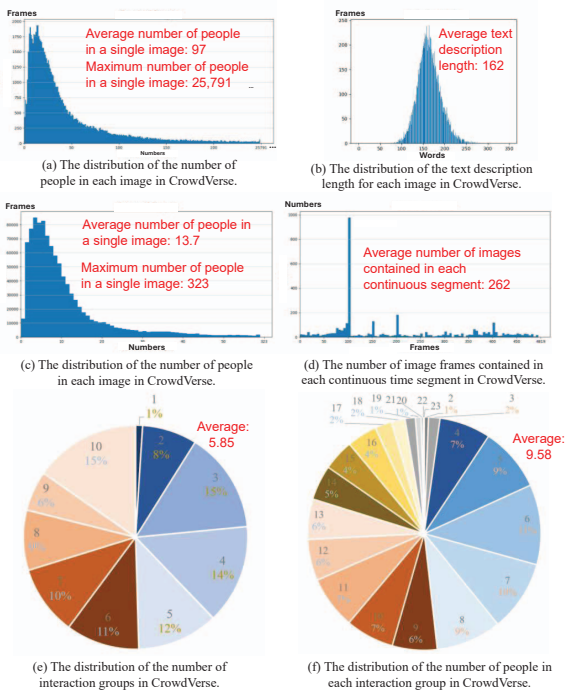
Multimodal Text-based person retrieval datasets (CUHK-PEDES [33], ICFG-PEDES [14], RSTPReid [71]) and attribute-focused datasets (Market-1501 [68], DeepFashion [41], PETA [13]) prioritize individual characteristics over crowd behaviors. Synthetic datasets such as GCC [51] lack behavioral authenticity. CrowdVerse introduces diverse crowd densities, environmental contexts, multimodal annotations, addressing limitations in existing resources.

2.2. Evaluation of Crowd Simulation

Currently, most methods for quantifying the realism of crowd simulations are **reference-based**. When real-world crowd trajectories are available, many studies [30] rely on trajectory-based metrics, such as Average Displacement Error (ADE), to compare simulated trajectories with real human movement data. However, such methods are difficult to apply in scenarios where real crowd trajectories are unavailable. Another approach focuses on the physical constraints of crowd movement, evaluating the adherence to physical laws by computing features such as each pedestrian's travel distance, velocity changes, angular variations, and collision rates [20]. Group dynamic behaviors can also be analyzed through clustering patterns [67]. Nevertheless, many modern physics-based rendering platforms can already simulate group behaviors that exhibit realistic physical movement characteristics, limiting the effectiveness of such metrics in further assessing crowd realism. Some psychological approaches have also been explored, such as comparing simulation outputs with human visual attention patterns [28] or using subjective human evaluations [18]. In terms of **reference-free** evaluation, a widely accepted strategy is to measure the "generalized distance" between the distributions of generated and real samples, similar to methods used in image generation assessment. In the field of crowd simulation, several preliminary efforts have been made. For instance, [57] evaluates realism by computing the Structural Similarity Index (SSIM) between simulated and real heatmaps, while [70] employs Inception Score (IS) and Fréchet Inception Distance (FID) on heatmaps. However, these methods often require environment-specific training, which makes it difficult to transfer them to new environments. The introduction of CrowdVerse opens new avenues for addressing this challenge by integrating rich and diverse crowd data.

3. CrowdVerse Dataset

Figure 2 illustrates the overall architecture of CrowdVerse.



(a) The distribution of the number of people in each image in CrowdVerse.

(b) The distribution of the text description length for each image in CrowdVerse.

(c) The distribution of the number of people in each image in CrowdVerse.

(d) The number of image frames contained in each continuous time segment in CrowdVerse.

(e) The distribution of the number of interaction groups in CrowdVerse.

(f) The distribution of the number of people in each interaction group in CrowdVerse.

Figure 3. Statistic data analysis of CrowdVerse.

3.1. Acquisition and Filtering

Acquisition The data collection process is divided into two phases: data integration and data sampling. In the data integration phase, we survey images featuring crowds and incorporate selected images into CrowdVerse based on their applicability. A total of 1,038,760 existing crowd images are selected. In the data sampling phase, we consider the diversity of images in the existing dataset, including factors such as crowd density, shooting angles, and background environments. To address the shortcomings identified in the dataset, we capture images that encompass a variety of environments in real-world settings using RGB cameras. This collection includes shooting angles ranging from 0 to 90 degrees and crowd scenes ranging from low to high density. A total of 115,418 new images have been added.

Filtering We filter the dataset to construct an image set suitable for specific annotation tasks. Specifically, we annotate four categories: human numbers, human locations, human interaction groups, and text descriptions. Each image undergoes a usability check before labeling to determine whether the crowd patterns in the image meet the requirements of the respective label. Since the crowd patterns in certain images may not be appropriate for specific tasks, we manually select image sets tailored to each task. For example, images with extremely dense crowds are excluded from human interaction group annotations. For text description labels, when dealing with a sequence of frames over time, we select only frames with significant or distinctive features to avoid redundancy and to ensure that the descriptions could meaningfully differentiate between frames. Some datasets for specific tasks are constrained by environmental factors, for instance, trajectory prediction datasets typically use a top-down perspective. To enhance the generalization ability of the algorithm, we retain images with smaller tilt angles or near-eye-level perspectives, as these also contain valuable information for trajectory prediction. This inclusion further increases the diversity and challenge in CrowdVerse.

3.2. Annotation

Coarse-grained Annotation To handle the large-scale annotation of pedestrian localization and trajectory data, we initially employ two classic models—P2PNet [47] and Matching Anything by Segmenting Anything (MASA) [34]—for coarse-grained annotation. For scenarios with varying crowd densities, we apply two distinct annotation approaches for pedestrian position data: in high-density crowds, primarily for counting tasks, we use point annotations to mark head locations, while in low-density crowds, primarily for tracking tasks, we adopted bounding boxes for finer-grained annotation.

Manual Labeling At this stage, we conduct a more refined annotation process for the dataset. We recruit a large number of professional annotators to examine and modify the labels, manually revising any non-standard labels. In addition, annotators are required to manually label the interaction groups and text description information in the dataset. They use as much detailed language as possible to describe each image. Since all images in the dataset are related to crowds, lengthy and detailed text descriptions are essential for forming specific image labels. We ensure the accuracy of the annotations, allowing humans to correctly retrieve corresponding images based on the features described in the text.

3.3. Data Analysis

CrowdVerse comprises 1,154,178 images, with 1,052,490 images forming time-continuous frames. The dataset contains annotations for 14,572,428 pedestrian bounding boxes and 8,425,714 pedestrian location points, achieving annotation volume 10 times that of mainstream crowd-counting datasets. Additionally, 13,104 images include text descriptions totaling 2,130,129 words, with the average text description length per image surpassing that of existing multimodal datasets with crowds by a factor of 10. Furthermore, 2,027 images feature 11,858 interaction group tags, amounting to nearly 3 times the data volume of current interaction-labeled datasets. Figure 3 provides a detailed distribution of various labels of CrowdVerse. Our temporally continuous frames span over 58 unique scenes, addressing the limited scene diversity of traditional surveil-

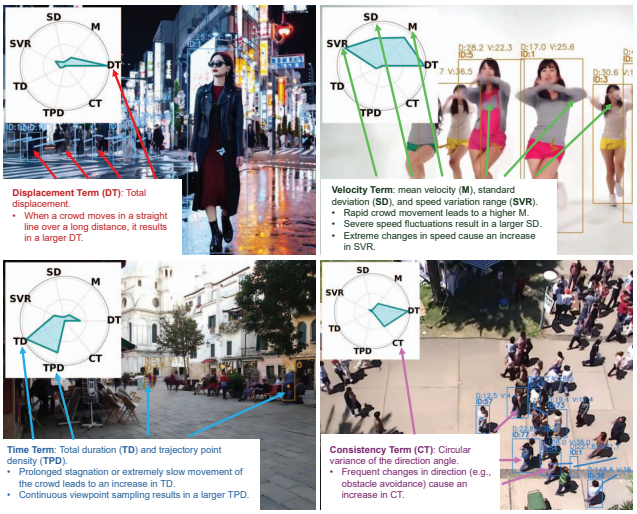


Figure 4. The visual representation of four physical terms for crowd analysis. The lines in the bounding box show the pedestrian's movement trajectory.

lance datasets while enhancing temporal correlation in general-scene datasets.

Data Split Data in continuous time frames is divided into train/test/validation sets with 841,992/105,249/105,249 frames; data in discrete event frames is divided into train/test/validation sets with 69,246/8,655/8,656 frames; social interaction data is split into train/test sets with 1,448/579 samples; and multimodal data is split into train/test sets with 11,793/1,311 samples.

3.4. Evaluation Framework

We propose a reference-free evaluation framework to assess the realism of simulated crowd movements. Previous studies [20, 67] have primarily focused on analyzing the realism of crowd dynamics in simulations by computing physical features such as pedestrian travel distance, speed change, and angle change. Building upon this foundation, we introduce four physical terms as metrics for evaluating crowd movement: Displacement Term, Velocity Term, Time Term, and Consistency Term. As shown in Figure 4, these four physical terms effectively capture various motion characteristics of the crowd. We further introduce two metrics, CIS and CFID, to evaluate the realism of simulated movements.

Crowd images are captured from varying perspectives and at different scales. To ensure translation invariance, we normalize coordinates by converting them into relative coordinates with respect to the starting point. Given a trajectory represented as a sequence of points $\mathbf{T} = \{(f_i, x_i, y_i)\}$, where f_i denotes the frame index, and (x_i, y_i) represents the spatial coordinates, the transformation is defined as:

$$\tilde{x}_i = x_i - x_0, \quad \tilde{y}_i = y_i - y_0 \tag{1}$$

where (x_0, y_0) is the position of the first recorded point in the trajectory. Given $\mathbf{T}$, we extract four physical terms.

Displacement Term (DT) We compute the total displacement between the starting and ending points:

$$DT = \|(x_N, y_N) - (x_0, y_0)\|_2 \tag{2}$$

where (x_0, y_0) and (x_N, y_N) are the coordinates of the starting and ending points.

Velocity Term We compute the mean, standard deviation, and range of speed across trajectory points. The instantaneous velocity at each frame is given by:

$$v_i = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{f_i - f_{i-1}}, \quad \text{for } i \geq 1. \tag{3}$$

From this, we derive the mean velocity (M), standard deviation (SD), and speed variation range (SVR):

$$M = \mathbb{E}[v], \quad SD = \text{std}(v), \quad SVR = \max(v) - \min(v) \tag{4}$$

Time Term This term contains total duration and trajectory point density.The total duration (TD) is: $TD = f_N - f_0$. The sampling density (TPD) is computed as: $TPD = N/T$, where N is the number of recorded points in $\mathbf{T}$.

Consistency Term (CT) The direction angle between consecutive points is given by: $\theta_i = \arctan 2(y_i - y_{i-1}, x_i - x_{i-1})$ The circular variance of these angles is defined as:

$$R = \frac{\sqrt{(\sum_{i=1}^{N-1} \sin\theta_i)^2 + (\sum_{i=1}^{N-1} \cos\theta_i)^2}}{N-1}, CT = 1 - R \tag{5}$$

Crowd Inception Score Given a set of real crowd features $X_r \in \mathbb{R}^{N\times d}$ and generated crowd features $X_g \in \mathbb{R}^{M\times d}$, we standardize the features and train a KMeans clustering model with K clusters on the real data:

$$X_r' = K\text{Means}(X_r), \quad X_g' = K\text{Means}(X_g). \tag{6}$$

The trained KMeans model assigns labels to real samples, which are then used to train a multinomial logistic regression classifier. The classifier predicts the probability distribution $P(y|X_g')$ for generated samples. The marginal probability $P(y)$ is computed as:

$$P(y) = \frac{1}{M}\sum_{j=1}^{M} P(y|X_{g,j}'). \tag{7}$$

The KL-divergence between $P(y|X_g')$ and $P(y)$ is computed for each sample, and the CIS is given by:

$$\text{CIS} = \exp\left(\frac{1}{M}\sum_{j=1}^{M}\sum_{y} P(y|X_{g,j}')\log\frac{P(y|X_{g,j}')}{P(y)}\right). \tag{8}$$

Crowd Fréchet Inception Distance Given a set of real features $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ and generated features $\boldsymbol{Y} \in \mathbb{R}^{M\times d}$, we apply feature standardization using a fitted scaler:

$$\boldsymbol{X}_{\text{norm}} = \frac{\boldsymbol{X} - \boldsymbol{\mu}_X}{\boldsymbol{\sigma}_X}, \quad \boldsymbol{Y}_{\text{norm}} = \frac{\boldsymbol{Y} - \boldsymbol{\mu}_X}{\boldsymbol{\sigma}_X} \tag{9}$$

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

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 [8]	61.2	97.8	84.2	152.4	69.4	262.4	-	-	215.4	296.4	65.5	295.2
MAN [36]	56.8	90.3	77.3	131.5	53.4	209.9	76.5	323.0	-	-	49.9	288.3
CLTR [35]	56.9	95.2	85.8	141.3	59.5	240.6	74.3	333.8	-	-	54.2	298.6
PET [38]	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 [6]	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 3. Performance comparison of methods on trajectory prediction tasks on CrowdVerse and other datasets.

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 [19]	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 [23]	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 [59]	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 [3]	0.35	0.42	0.13	0.19	0.19	0.32	0.15	0.25	0.25	0.44	10.02	14.43

where $\boldsymbol{\mu}_X$ and $\boldsymbol{\sigma}_X$ are the mean and standard deviation computed from the real dataset. The mean vectors for real and generated features are computed as:

$$\boldsymbol{\mu}_X = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i, \quad \boldsymbol{\mu}_Y = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{y}_i \tag{10}$$

The covariance matrices for features are:

$$\boldsymbol{\Sigma}_X = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}_X)(\boldsymbol{x}_i - \boldsymbol{\mu}_X)^T \tag{11}$$

$$\boldsymbol{\Sigma}_Y = \frac{1}{M-1}\sum_{i=1}^{M}(\boldsymbol{y}_i - \boldsymbol{\mu}_Y)(\boldsymbol{y}_i - \boldsymbol{\mu}_Y)^T \tag{12}$$

The Crowd Fréchet Inception Distance is computed as:

$$\text{CFID} = \|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|^2 + \text{Tr}(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y - 2(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_Y)^{\frac{1}{2}}) \tag{13}$$

where $\text{Tr}(\cdot)$ denotes the trace of a matrix and $(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_Y)^{\frac{1}{2}}$ is the square root of the product of covariance matrices.

4. Experiments

4.1. Crowd Understanding Benchmark

CrowdVerse supports a variety of crowd understanding tasks. We evaluate the performance of existing methods on five tasks to demonstrate the broad applicability of CrowdVerse. More details can be found in the appendix.

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 the tasks. However, CrowdVerse encompasses more diverse scenarios and richer density patterns, making it a valuable benchmark in the crowd counting domain.

Trajectory Prediction The results are shown in Table 3. The findings reveal that baseline methods perform worse on CrowdVerse compared to other datasets. This suggests that CrowdVerse encompasses a broader range of scenes and a higher proportion of challenging cases with intricate movements. CrowdVerse addresses the limitation of the lack of large-scale datasets with complex interactions, which can facilitate the development of more advanced methods.

Human Interaction Detection Table 5 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.

Image Captioning Table 4 presents the results. Human ratings favor the ground-truth (GT) descriptions in CrowdVerse, reflecting the gap between model-generated content and human understanding. While models like BLIP and LLaVA can generate relevant descriptions, they still struggle to capture complex scene details and subtle contextual cues. This suggests that the annotations in CrowdVerse pose challenges to the multimodal field, highlighting the need for improvements in generating context-rich descriptions.

Image-text Retrieval The results are shown in Table 6, the BLIP2 model performs well on established datasets. 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 Crowd-

Table 4. Performance comparison of different methods in image captioning on CrowdVerse.

Methods	CLIPscore↑	RefCLIPscore↑	B@1↑	B@4↑	METEOR↑	ROUGE↑	CIDER↑	SPICE↑	Human Score↑
BLIP [31]	0.7168	0.7349	0.6752	0.0730	0.1149	0.2765	0.1020	0.0577	2.020
BLIP2 [32]	0.7547	0.7851	0.5977	0.0727	0.1129	0.2884	0.1068	0.0516	2.088
SmallCap [33]	0.6911	0.7483	0.5036	0.0493	0.0925	0.2661	0.0470	0.0440	1.804
ConZIC [63]	0.9817	0.8267	0.4016	0	0.0890	0.2152	0.0314	0.0404	1.216
LLaVa [39]	0.7464	0.7782	0.7081	0.1906	0.2015	0.4013	0.1612	0.1665	4.510
GT	0.8241	-	-	-	-	-	-	-	4.931

Table 5. Performance comparison of different methods in human interaction detection on CrowdVerse and SID.

Methods	SID					CrowdVerse				
	P	R	F1	Acc	AUC	P	R	F1	Acc	AUC
ARG [55]	41.87	54.14	47.22	69.83	0.534	38.81	38.54	38.67	89.33	0.847
JS [16]	40.02	61.62	48.53	67.41	0.596	27.73	28.40	28.06	87.29	0.772
CAGNet [53]	38.95	70.70	50.23	65.07	0.712	47.03	48.42	47.41	90.74	0.891
PEAN [61]	63.89	68.47	66.10	82.49	0.881	50.44	54.89	52.57	91.35	0.932

Table 6. Performance comparison of BLIP2 on image-text retrieval tasks on CrowdVerse and other datasets.

BLIP 2		**CrowdVerse**		**Flickr30K** [43]	**COCO** [37]
		Test	Validation	Test Set	Test Set
Image → Text	$R@1$	0.11	0.12	0.98	0.85
	$R@5$	0.68	0.62	1.00	0.97
Text → Image	$R@1$	0	0.12	0.90	0.68
	$R@5$	0.58	0.62	0.98	0.88

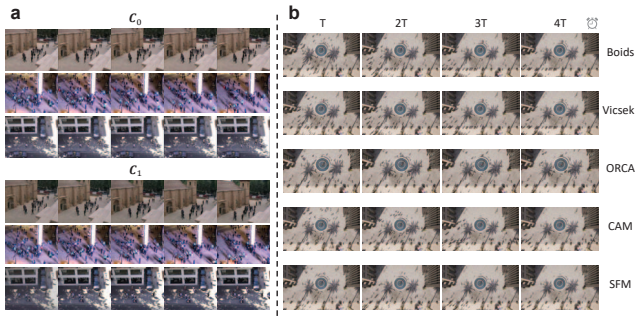


Figure 5. Visualization of the simulation. **a**: Samples from $\mathbf{C}_0$ and $\mathbf{C}_1$, which share the same environment but exhibit slight differences in crowd motion patterns. **b**: Simulation results of reference-free crowd simulation methods on the CARLA platform.

Verse present unique challenges.

4.2. Validation of CIS and CFID

We first design experiments to validate the effectiveness of the evaluation metrics proposed in Section 3.4. However, this validation is nontrivial since the metrics are specifically designed for scenarios without ground-truth crowd motion data. To address this, we introduce the following definition: *for a metric intended to assess crowd motion realism, under the same environment, its score for real crowd motions should be greater than or equal to that for any simulated crowd motions, considering that an ideal simulation should reproduce realistic motion patterns.*

Accordingly, we design the following experiments. We first sample 200 five-second video clips of real-world crowd motions that are not included in CrowdVerse, denoted as $\mathbf{C}_0$. For each clip in $\mathbf{C}_0$, we extract its initial frame and employ an image-to-video generation model (Sora Turbo) to synthesize a video of the same duration, constructing a new dataset denoted as $\mathbf{C}_1$. Figure 5.a presents a comparison of samples from $\mathbf{C}_0$ and $\mathbf{C}_1$. The datasets $\mathbf{C}_0$ and $\mathbf{C}_1$ share identical numbers of individuals and visual characteristics, differing only slightly in their crowd motion patterns, which poses a challenge for distinguishing truly realistic behaviors. In our experiments, we use datasets excluding $\mathbf{C}_0$ as benchmarks and evaluate motion realism using the CIS and CFID metrics. According to the metric definitions, a crowd motion is considered more realistic if it achieves a higher CIS score and a lower CFID score.

Implementation Details The videos in both $\mathbf{C}_0$ and $\mathbf{C}_1$ are standardized to a resolution of 720×1280 and a frame rate of 30 fps. When employing Sora Turbo as the generation model, the text prompts contain only descriptions of the environment and crowd scenes, without any explicit motion instructions, ensuring that the dynamic characteristics of the generated results arise primarily from the model's intrinsic inference. For CFID computation, the dimensionality of the latent feature vectors is set to 512.

Results and Analysis Table 7 presents the discriminative performance of various crowd datasets composed of sequential image frames when used as benchmarks. Experimental results demonstrate that the CrowdVerse-based metrics successfully identify the crowd motion in $\mathbf{C}_0$ as more realistic than that in $\mathbf{C}_1$, as evidenced by the higher CIS and lower CFID scores on $\mathbf{C}_0$. However, most datasets fail to be correctly identified by both metrics. Certain datasets, such

Table 7. CIS and CFID values of $\mathbf{C}_0$ and $\mathbf{C}_1$ across different benchmark datasets. If CIS is higher and CFID is lower on $\mathbf{C}_0$, it indicates that the discriminator based on this dataset can effectively distinguish between real and generated crowd movements.

Datasets	CIS ↑		CFID ↓	
	$\mathbf{C}_0$	$\mathbf{C}_1$	$\mathbf{C}_0$	$\mathbf{C}_1$
AHUCrowd	3.44	4.17	105.44	99.60
CUHK-Square	2.55	5.84	100.07	90.40
EWAP	3.44	4.74	49.02	46.68
FDST	3.40	4.53	64.18	61.26
HiEve	3.17	3.02	3.28	2.56
STEP	2.63	3.00	36.25	32.43
UCSD	4.13	5.17	295.21	291.34
VENICE	4.11	4.46	24.30	18.52
RGBT-CC	3.28	2.64	16.95	16.02
ShanghaiTech	3.47	1.68	7.48	6.12
Disco	2.51	3.14	50.28	49.53
FUDAN-UCC	3.38	2.57	57.72	42.20
CrowdVerse	5.28	1.79	3.83	13.17

Table 8. Ablation study of physical terms. Δ: the difference in the metric between $\mathbf{C}_0$ and $\mathbf{C}_1$.

	w/o M	w/o SD	w/o SVR	w/o TD	w/o TPD	w/o CT	All terms
Δ CIS	1.680	2.685	2.100	2.063	3.067	2.133	**3.489**
Δ CFID	-9.336	-8.686	-8.440	-8.690	-9.260	-6.451	**-9.340**

Table 9. CIS and CFID of different reference-free methods.

Methods	Boids	Vicsek	SFM	ORCA	CAM	$\mathbf{C}_0$
CIS↑	1.051	1.073	1.986	1.006	1.061	**5.28**
CFID↓	14.508	13.992	12.279	14.982	14.226	**3.83**

as HiEve and ShanghaiTech, produce correct results on only one of the two metrics, while showing contradictory results on the other. This inconsistency arises because, although these datasets exhibit some degree of scale in specific visual features, they lack diversity or coverage in others, making them unreliable as benchmarks for real-world crowd data.

Ablation Study To investigate the impact of the selected physical terms, we conduct an ablation study, with the results shown in Table 8. The experimental results show that the evaluation framework based on seven physical terms yields the largest absolute difference in CIS and CFID values in $\mathbf{C}_0$ and $\mathbf{C}_1$. The findings indicate that each feature of the crowd movement contributes to the improvement of the performance of the evaluation framework, confirming the effectiveness of the selected physical quantities.

4.3. Benchmarking Reference-free Simulation

Based on the above metrics, we conduct a study on the realism of crowd behaviors generated by existing reference-free crowd simulation methods. The number of such methods is relatively limited, as most recent simulation approaches typically require training in specific environments. Specifically, under identical initial conditions, we evaluate five representative methods: Boids [44], Vicsek [49], Social Force Model (SFM) [21], Reciprocal n-body Collision Avoidance (ORCA) [48], and Cellular Automata Model (CAM) [45].

Implementation Details The simulation experiments are conducted in ten representative environments [29] in CARLA platform [15], covering a variety of urban scenes and crowd distribution patterns. All environments maintain fixed weather and lighting conditions to ensure consistency. Individual trajectory information, including position, velocity, and orientation, is recorded using CARLA's built-in sensors. Each method is executed three times in every environment to mitigate the effects of randomness.

Results and Analysis Table 9 presents the experimental results, with $\mathbf{C}_0$ serving as the upper bound for motion realism. Figure 5.b provides visual comparisons across different simulation methods. From the results, it is evident that traditional reference-free approaches achieve significantly lower CIS scores and notably higher CFID scores than real crowd data, indicating a substantial gap between simulated and authentic crowd behaviors. Among these methods, SFM performs the best, as it explicitly models social interaction and repulsive forces between individuals, producing more natural group movement patterns. In contrast, ORCA imposes overly strict collision avoidance constraints, leading to rigid and unnatural trajectories. Boids and Vicsek are capable of simulating low-level behaviors such as alignment and aggregation but struggle to capture complex crowd interactions. Overall, a clear gap remains between current simulation methods and real-world crowd movements, highlighting the limitations of existing reference-free crowd simulation techniques in modeling realistic crowd dynamics.

5. Conclusion

We introduce CrowdVerse, a bidirectional dataset designed to advance both crowd understanding and crowd simulation. CrowdVerse comprises over 1.1 million images that span diverse crowd densities, environments, and viewpoints. It provides extensive annotations to support a wide range of crowd-related tasks. Compared to conventional datasets, the performance of classic methods across multiple tasks drops significantly on CrowdVerse, highlighting the substantial challenges posed by its diversity. We conduct extensive experiments, establishing baselines to facilitate future research. Furthermore, we propose a framework for evaluating the realism of reference-free crowd simulations based on CrowdVerse, effectively capturing the motion patterns of real-world crowds and distinguishing between real-world and simulated crowd movements.

6. Acknowledgments

This research was partly supported by grants of National Natural Science Foundation of China (NSFC, Grant No. 62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 25GA32001003).

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 6

[2] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1707–1720, 2015. 3

[3] Inhwan 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*, pages 17890–17901, 2024. 6

[4] Reza Bahmanyar, Elenora Vig, and Peter Reinartz. Mrcnet: Crowd counting and density map estimation in aerial and ground imagery. *arXiv preprint arXiv:1909.12743*, 2019. 3

[5] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2008. 1, 2

[6] I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, and Sy-Yen Kuo. Improving point-based crowd counting and localization based on auxiliary point guidance. In *European Conference on Computer Vision*, pages 428–444. Springer, 2024. 6

[7] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, page 3, 2012. 1, 2

[8] Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19638–19648, 2022. 6

[9] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009. 3

[10] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 417–433. Springer, 2014. 1, 3

[11] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6172–6181, 2019. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 1, 3

[14] 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. 1, 3

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 8

[16] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. 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*, pages 177–195. Springer, 2020. 7

[17] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-constrained spatial transformer network for video crowd counting. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 814–819. IEEE, 2019. 2

[18] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 3

[19] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 6

[20] Stephen J Guy, Jatin Chhugani, Sean Curtis, Pradeep Dubey, Ming C Lin, and Dinesh Manocha. Pledestrians: A least-effort approach to crowd simulation. In *Symposium on computer animation*, pages 119–128, 2010. 3, 5

[21] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 8

[22] Yaocong Hu, Huan Chang, Fudong Nian, Yan Wang, and Teng Li. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38:530–539, 2016. 1

[23] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019. 6

[24] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 3

[25] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 3

[26] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018. 1, 3

[27] Di Kang, Debarun Dhar, and Antoni Chan. Incorporating side information by adaptive convolution. *Advances in neural information processing systems*, 30, 2017. 1

[28] Sujeong Kim, Stephen J Guy, Dinesh Manocha, and Ming C Lin. Interactive simulation of dynamic crowd behaviors using general adaptation syndrome theory. In *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, pages 55–62, 2012. 3

[29] Pingrui Lai, Renjie Pan, Jiaqi Yu, and Hua Yang. Hydrodynamics regularization in reinforcement learning for navigating crowded scenarios. *IEEE Robotics and Automation Letters*, 2025. 8

[30] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, pages 655–664. Wiley Online Library, 2007. 3

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 7

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 7

[33] 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*, pages 1970–1979, 2017. 1, 3, 7

[34] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18973, 2024. 4

[35] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022. 6

[36] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19628–19637, 2022. 6

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[38] Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1676–1685, 2023. 6

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[40] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5099–5108, 2019. 1, 2

[41] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 3

[42] Santiago Manen, Michael Gygli, Dengxin Dai, and Luc Van Gool. Pathtrack: Fast trajectory annotation with path supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 290–299, 2017. 1

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7

[44] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 25–34. ACM, 1987. 8

[45] Andreas Schadschneider, Wolfram Klingsch, Hubert Klüpfel, Tobias Kretz, Christian Rogsch, and Armin Seyfried. Evacuation dynamics: Empirical results, modeling and applications. *Encyclopedia of complexity and systems science*, pages 3142–3176, 2009. 8

[46] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020. 1, 3

[47] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021. 4

[48] Jur van den Berg, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research*, pages 3–19. Springer, 2008. 8

[49] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a

system of self-driven particles. *Physical review letters*, 75 (6):1226, 1995. 8

[50] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. Context-aware cnns for person head detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2893–2901, 2015. 1

[51] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 3

[52] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 1, 3

[53] Zhenhua Wang, Jiajun Meng, Dongyan Guo, Jianhua Zhang, Javen Qinfeng Shi, and Shengyong Chen. Consistency-aware graph network for human interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13369–13378, 2021. 7

[54] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network. *arXiv preprint arXiv:1912.01811*, 2019. 3

[55] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9964–9974, 2019. 7

[56] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017. 1

[57] Dapeng Yan, Gangyi Ding, Kexiang Huang, and Tianyu Huang. Generating natural pedestrian crowds by learning real crowd trajectories through a transformer-based gan. *The Visual Computer*, 41(2):1079–1096, 2025. 3

[58] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 952–961, 2019. 3

[59] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 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*, pages 507–523. Springer, 2020. 6

[60] Jiaqi Yu, Jinhai Yang, Hua Yang, and Guangtao Zhai. Prn: Psychology-inspired relation network for detecting social interaction groups from single images. In *BMVC*, page 364, 2021. 1, 3

[61] Jiaqi Yu, Jinhai Yang, Hua Yang, Renjie Pan, Pingrui Lai, and Guangtao Zhai. Psychology-guided environment aware network for discovering social interaction groups from videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(8):1–23, 2024. 7

[62] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020. 1

[63] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476, 2023. 7

[64] Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016. 2

[65] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1113–1121. IEEE, 2018. 1, 2

[66] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 1, 3

[67] Mingbi Zhao, Wentong Cai, and Stephen John Turner. Clust: simulating realistic crowd behaviour by mining pattern from crowd videos. In *Computer graphics forum*, pages 184–201. Wiley Online Library, 2018. 3, 5

[68] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 3

[69] Xiaopin Zhong, Guankun Wang, Weixiang Liu, Zongze Wu, and Yuanlong Deng. Mask focal loss: a unifying framework for dense crowd counting with canonical object detection networks. *Multimedia Tools and Applications*, pages 1–23, 2024. 1

[70] Yanshan Zhou, Pingrui Lai, Jiaqi Yu, Yingjie Xiong, and Hua Yang. Hydrodynamics-informed neural network for simulating dense crowd motion patterns. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4553–4561, 2024. 3

[71] 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*, pages 209–217, 2021. 1, 3