PEDRA: Evaluating the Realism of Pedestrian Dynamics in Video Generation

Supplementary Material

Appendix

This section includes additional background, details on the methodology, and extended results and analysis. We first provide a brief summary of **video diffusion models** (§A). The **Evaluation Metrics** (§B) section covers implementation details not provided in the main paper. We describe the **prompt suite** (§C) and **inference process** (§D) including computing hardware and hyperparameters. We provide the additional image-to-video and text-to-video **results plots** (§E). Finally, we show additional **qualitative examples** (§F) of the video generations and postprocessing results, including successes and common failure modes.

A. Background on Video Diffusion Models

Modern video diffusion models (VDMs) use the latent diffusion paradigm [45, 79], which performs denoising in a compressed variational autoencoder (VAE) latent space. Early approaches adapted 2D U-Net [80] backbones by either inserting temporal modules into frozen spatial layers [12, 22, 46, 95] or using unified space-time architectures [9]. A subsequent architectural shift replaced the U-Net with the more scalable Diffusion Transformer (DiT) backbone [25, 72], which now underpins many SOTA models [23, 30, 51, 70, 75, 93, 100]. Conditioning signals are typically integrated via cross-attention [23], adaptive layer normalization [72], or unified self-attention across modalities [50, 76]. In standard DiT training for video models, there is no explicit object or agent representation or any type of multi-agent inductive bias.

B. Evaluation Metrics

Our evaluation protocol distinguishes between the I2V and T2V tasks as described in the main text. We largely draw inspiration from two recent sources. From Bae et al. [8] we adapt Velocity, Acceleration, and Distance, and Population. The primary change is that we compute per-agent rather than per-frame averages to allow for multiple repetitions of the same simulation. From Minartz et al. [63] we adapt Nearest Neighbor Distance and Flow.

B.1. Trajectory Kinematics

We first employ a Kalman smoother, $\mathcal{K}$, to estimate smoothed position $\tilde{\mathbf{p}}_k^i$ and velocity $\mathbf{v}_k^i$ states. For each agent i, we compute $(\tilde{\mathbf{p}}_k^i, \mathbf{v}_k^i)_{k=k_{\text{start}}^i}^{k_{\text{end}}^i} = \mathcal{K}(\mathcal{T}^i)$. Using the estimated velocity, we compute the average speed $\bar{s}^i$ for each agent i over its trajectory:

$$\bar{s}^i = \frac{1}{L_i - 1} \sum_{k=k_{\text{start}}^i+1}^{k_{\text{end}}^i} \|\mathbf{v}_k^i\|_2$$

From the velocities $\mathbf{v}_k^i$, we compute the instantaneous acceleration using finite difference of velocity, then compute a per-agent average acceleration magnitude:

$$\bar{a}^i = \frac{1}{L_i - 2} \sum_{k=k_{\text{start}}^i+2}^{k_{\text{end}}^i} \frac{\|\mathbf{v}_k^i - \mathbf{v}_{k-1}^i\|_2}{t_k - t_{k-1}}$$

where t_k is the timestamp (in seconds) for frame k.

The total path length d^i is calculated by summing the Euclidean distance between consecutive points along the smoothed agent path:

$$d^i = \sum_{k=k_{\text{start}}^i+1}^{k_{\text{end}}^i} \|\tilde{\mathbf{p}}_k^i - \tilde{\mathbf{p}}_{k-1}^i\|_2$$

B.2. Social Interaction

Collision Rate. We define a collision for an agent as any instance where another agent is within a distance threshold $\delta = 0.1$ meters. Define the set of agent indices active at a specific time step k as $\mathcal{A}_k$. The indicator function $\mathbb{I}_{\text{coll}}(i,k)$ is 1 if agent i is in a collision at time k, and 0 otherwise:

$$\mathbb{I}_{\text{coll}}(i,k) = 1 \iff \exists j \in \mathcal{A}_k, j \neq i : \|\mathbf{p}_k^i - \mathbf{p}_k^j\|_2 < \delta$$

These indicators are summed over the full set of generated trajectories to compute $\mathcal{M}_{\text{coll}}$ and $\mathcal{M}_{\text{coll}}^{\text{E}}$ (see Section 4.2).

Stationary Agents. An agent is classified as stationary using an indicator function if the Euclidean distance between its start and end positions is less than a threshold $\delta_{\text{stat}} = 0.2$ meters:

$$\mathbb{I}_{\text{stat}}(i) = \begin{cases} 1 & \text{if } \|\mathbf{p}_{k_{\text{end}}^i}^i - \mathbf{p}_{k_{\text{start}}^i}^i\|_2 < \delta_{\text{stat}} \\ 0 & \text{otherwise} \end{cases}$$

Flow. To compute $\mathcal{M}_{\text{flow}}$ and $\mathcal{M}_{\text{flow}}^{\text{E}}$, we explicitly define the local density calculation and directional partitioning. For each agent i at time k, local density ρ_k^i (agents/m^2) is estimated using the Euclidean distance r_k to its $K = 4$th nearest neighbor: $\rho_k^i = \frac{K}{\pi r_k^2}$.

We partition all agent-timestep pairs (i,k) into two sets based on the primary direction of movement:

- $\mathcal{S}_x$: The set of pairs where movement is predominantly along the x-axis, $|v^i_{k,x}| > |v^i_{k,y}|$.
- $\mathcal{S}_y$: The set of pairs where movement is predominantly along the y-axis, $|v^i_{k,y}| > |v^i_{k,x}|$.

The instantaneous flow is $f^i_k = \rho^i_k \cdot \|\mathbf{v}^i_k\|_2$. The metric distributions $\mathcal{F}_x$ and $\mathcal{F}_y$ are constructed from the set of all f^i_k belonging to $\mathcal{S}_x$ and $\mathcal{S}_y$, respectively.

Nearest Neighbor Distribution. The nearest neighbor search includes a velocity threshold to exclude static agents. The metric is computed over all agent-timestep pairs (i, k) where the agent i is moving (speed $s^i_k > \epsilon = 0.1$m/s). For each moving agent, we find its nearest moving neighbor, j^*, within a 10m radius:

$$j^* = \underset{j \in \mathcal{A}_k, j \neq i, s^j_k > \epsilon}{\arg\min} \|\mathbf{p}^i_k - \mathbf{p}^j_k\|_2$$

The distance $\|\mathbf{p}^i_k - \mathbf{p}^{j^*}_k\|_2$ constitutes a sample in the distribution D_{nn}.

B.3. Video Fidelity

Track Disappearance. The central area of the frame described in Section 4.3 is defined with a proportion $\alpha = 0.10$ of the frame dimensions. For a frame of width W and height H (in pixels), the edge margins are $m_x = \alpha W$ and $m_y = \alpha H$. For each agent i, let $\mathbf{p}^i_{k^i_{\text{end}}} = (x^i_{k^i_{\text{end}}}, y^i_{k^i_{\text{end}}})$ be its last detected position. An agent is classified as having disappeared mid-scene using an indicator function:

$$\mathbb{I}_{\text{disp}}(i) = \begin{cases} 1 & \text{if } m_x < x^i_{k^i_{\text{end}}} < W - m_x \text{ and} \\ & \quad m_y < y^i_{k^i_{\text{end}}} < H - m_y \\ 0 & \text{otherwise} \end{cases}$$

C. T2V Prompt Suite

The *T2V Benchmark Method* section gave an overview of the method to systematically generate T2V prompts. Figure 5 visualizes the word cloud of the set of prompts showing the most common word is "people," which distinctly contrasts with the prompt suite in Vbench [48] where the most common word is "person" (singular). In GRADEO [67] the most common human-oriented words are "individual, person, boy, girl, man, woman." Figure 6 provides the full instruction script used to generate prompts by pasting in the desired density and interaction category. Other than the specific category strings, the instruction remains constant.

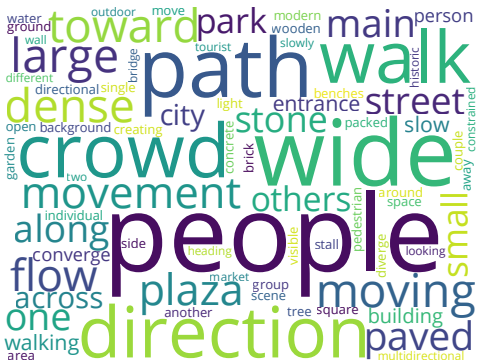


Figure 5. Word cloud visualization of prompt keywords in the T2V prompt suite.

D. Models and Inference Details

Model Selection. We selected five SOTA models that have both I2V and T2V variants: Wan2.1 (*WAN*) [93], CogVideoX1.5 (*CVX*) [100], HunyuanVideo (*HYV*) [51], LTX-Video (*LTX*) [36], and Open-Sora 2.0 (*OS*) [74]. The number of frames, video duration, video resolution, and frames per second (fps) vary by model due to the specific characteristics of the model architecture. Details of the model versions and generated video characteristics are provided in Table 5. While some of these models support a variable number of generated frames, Open-Sora 2.0 and HunyuanVideo are capped at 129 frames as of the time of writing. Therefore, we choose to generate all clips at (approx.) 5 second durations, with the resolution as close as possible to the original resolution of the input image for I2V, or as close as possible to 720p for T2V.

Hyperparameters. We keep the default or suggested hyperparameters for each of the models in order to match the best performance as reported by the authors. We use 50 inference steps for all models that take this as an input argument and otherwise leave the default. We specify `guidance_scale=6.0` for CogVideoX,

Table 5. Video Generation Specifications.

Model	FPS	Frames	Resolution	Duration (s)
Image-to-Video (I2V)				
CogVideoX1.5-5B-I2V	16	81	640×480*	5.0
hunyuan-video-i2v-720p	25	129	960×540	5.12
ltxv-13b-0.9.7-dev	30	153	Same as input	5.07
Open-Sora 2.0 768px	24	129	880×656	5.33
Wan2.1-I2V-14B-480P	16	81	832×480	5.0
Text-to-Video (T2V)				
CogVideoX1.5-5B	16	81	1360×768	5.0
hunyuan-video-t2v-720p	25	129	1280×720	5.12
ltxv-13b-0.9.7-dev	30	153	1216×704	5.07
Open-Sora 2.0 768px	24	129	1024×576	5.33
Wan2.1-T2V-14B	16	81	1280×720	5.0

*CogVideoX I2V resolution varies by dataset:
ETH (640×480), UCY (720×576)

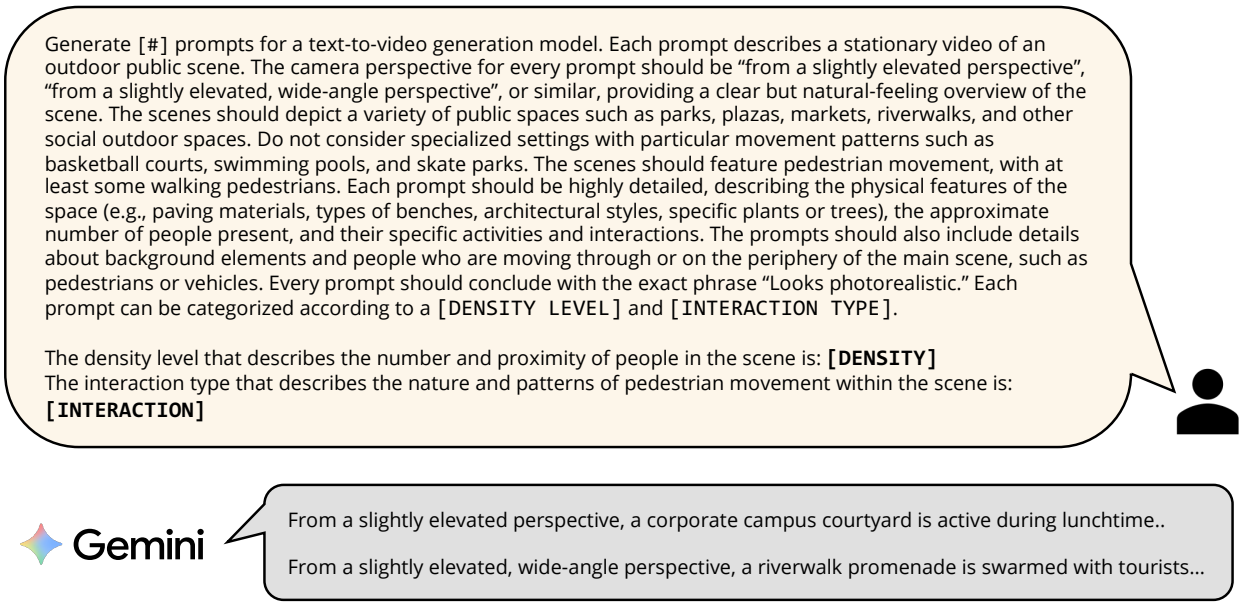


Figure 6. Script of the LLM instructions for generating text-to-video prompts. The instructions request a specific scene type, density, and interaction type, providing a standardized and compositional way to generate prompts. We used Gemini 2.5 Pro to generate the 180 prompts included in the supplementary material.

`guidance=7.5` and `guidance_img=3.0` for OpenSora 2.0, `embedded-cfg-scale=6.0` for HunyuanVideo, and `guidance_scale=5.0` for Wan2.1, which were all chosen by referencing examples in the model `README` or default values in provided sample generation scripts. For HunyuanVideo, we manually adjusted `cfg-scale=1.2` up from 1.0 as the default value of 1.0 causes the inference script to disregard a negative prompt.

We generate random seeds for the video model in order to vary the generations using the same prompt. We store the seed in the filename for future reproducibility.

Compute. In all generations, parallel inference was performed across four NVIDIA H200 GPUs, which resulted in generation times varying between 2 and 8 minutes per video depending on the model. Each GPU has 141GB memory, and we use a Linux machine with 16 CPUs and 128GB RAM. We use CUDA 12.4 and install all video generation models in conda environments according to the `README`.

Synthetic Datasets. Tables 6 and 7 document the statistics of the number of detected agents across the T2V and I2V benchmark, respectively. The total number of detections (N.D.) counts the total number of bounding boxes across all frames and all video clips. The number of unique agents (N.U.) counts the total number of unique identifiers assigned by the MOT model, where each track ID corresponds to a unique person tracked across multiple frames. The number of detections per frame (D/F) counts the average number of bounding boxes detected per frame.

As discussed in the *Method* section, the T2V prompt suite generated 5 repetitions for each of 20 prompts in each of the nine density/interaction categories (combinations of density {Cr., Mo., Sp.} with interaction {Di., Mu., Co.}). The videos are discarded if there is not sufficient agreement between the depth maps estimated by VGGT and Depth Pro. We require at least 100 pixels per frame for scale estimation, out of which at least 30% must be determined inliers, where the residual depth error after scaling is less than a threshold of 10% of the median metric depth. The discard rates per model for the T2V benchmark were: WAN: 16/900 (1.78%), HYV: 21/900 (2.33%), CVX: 25/900 (2.78 %), LTX: 45/900 (5.0%), OS: 29/900 (3.22%).

For I2V, as mentioned in the *I2V Benchmark* section of the paper, we extract non-overlapping start frames at 5-second intervals. We generate a single video for each start frame for each model. The goal is to develop a synthetic video dataset that is the same length and with the *same start distribution* of pedestrians as the ground truth dataset. For example, if we generated multiple videos for certain start frames, the resulting distributions would be biased towards that moment in time compared to the true reference. In the event that a model fails to produce a stationary and/or trackable video generation, we retry for that start image up to 5 times and retain the first video generation that contains any tracked agents.

Table 6. T2V Dataset Statistics: Number of Detections (N.D.), Number of Unique Agents (N.U.), and Average Number of Detections per Frame (D/F).

Model	Count	Total	Density			Interaction		
			Cr.	Mo.	Sp.	Di.	Mu.	Co.
WAN	N.D.	3.88e6	3.19e6	5.75e5	1.08e5	1.12e6	1.58e6	1.18e6
	N.U.	92431	79112	11367	1952	27148	39674	25609
	D/F	56.83	136.69	25.51	4.86	49.65	67.92	52.52
HYV	N.D.	3.76e6	2.58e6	1.02e6	1.65e5	1.22e6	1.36e6	1.19e6
	N.U.	68594	48871	17168	2555	22143	26297	20154
	D/F	35.36	72.21	27.58	4.89	35.17	37.99	32.95
OS	N.D.	2.47e6	1.55e6	7.69e5	1.53e5	6.93e5	1.05e6	7.24e5
	N.U.	36571	23605	11164	1802	10317	15608	10646
	D/F	22.79	42.13	20.73	4.43	19.40	28.52	20.27
LTX	N.D.	4.20e6	2.91e6	1.06e6	2.34e5	1.28e6	1.53e6	1.40e6
	N.U.	54816	36992	14646	3178	16753	20497	17566
	D/F	33.76	66.86	24.83	6.13	31.22	36.68	33.32
CVX	N.D.	1.90e6	1.30e6	5.30e5	74789	5.92e5	7.73e5	5.37e5
	N.U.	59584	44315	13563	1706	18197	25386	16001
	D/F	28.51	55.86	22.97	3.66	27.38	34.18	23.90

Table 7. I2V Dataset Statistics: Number of Detections (N.D.), Number of Unique Agents (N.U.), and Average Number of Detections per Frame (D/F). Note that the Ground Truth (GT) is shown as the top three rows.

Model	Count	ETH	UNIV	HOTEL	ZARA1	ZARA2
GT	N.D.	1861	101471	30505	8970	31624
	N.U.	224	2488	1142	353	788
	D/F	1.43	19.08	2.08	1.77	3.74
WAN	N.D.	4957	21984	5945	4593	8558
	N.U.	311	905	504	220	368
	D/F	1.74	8.73	1.49	1.78	2.21
HYV	N.D.	3738	59513	10099	10278	14987
	N.U.	204	1209	435	315	499
	D/F	1.60	11.84	1.51	1.99	2.25
OS	N.D.	3329	28684	1950	4238	8760
	N.U.	246	602	184	159	317
	D/F	1.25	5.47	1.04	1.28	1.79
LTX	N.D.	15305	73268	13769	16181	36580
	N.U.	547	1460	629	345	651
	D/F	2.12	12.86	1.60	2.76	3.87
CVX	N.D.	2446	22066	2002	1970	2002
	N.U.	236	944	227	159	193
	D/F	1.51	6.79	1.18	1.39	1.45

E. Additional Quantitative Results

E.1. Image-to-Video

Fig. 8 plots a heatmap of the agent positions on top of a background image of the UNIV scene. The plots show that all of the models roughly capture the shape of the ground truth (GT) spatial distribution. However, the relative densities vary. LTX appears to capture the true distribution best. CVX and OS show sparser overall distributions due to the lower trackability of the agents, resulting in fewer detected pedestrians over the same number of generated video clips; we note that the MOT Confidence metric ($\mathcal{M}_{\text{mot}}$) captures this in Table 9 as the lowest two scores in the UNIV scene.

Figure 9 shows the polar histograms for the same scene, which illustrate the relative location of the nearest neighbor to each agent. Prior research has noted that this position typically follows a bimodal distribution with peaks at distances around 0.5-0.75 meters [63]. The GT distribution indeed follows this pattern. Intuitively, this results from the fact that people walk side-by-side with some personal space in between themselves and the nearest other person. To some degree it reflects collision avoidance behavior, as two colliding walkers would lead to a nearest neighbor distance near zero, which would result in a dense cluster near the origin. Models HYV and WAN capture the GT NN distribution well, including the distance and angle of the two modes. LTX captures the distance of the NN modes well but displays a different relative orientation angle. CVX roughly captures the GT pattern but less clearly due to the sparser nature of detections resulting from lower agent trackability. OS displays the largest visual dissimilarity against the GT distribution, which is accurately reflected in Table 9 by the worst score in the UNIV scene for the $\mathcal{M}^{\text{E}}_{\text{nn}}$ metric.

E.2. Text-To-Video

Details on Real-World Reference Datasets. To establish reference ranges for T2V evaluation, we process each of the ten public pedestrian benchmarks using dataset-specific loaders from the OpenTraj toolkit [4]. For datasets with known camera parameters (ETH, UCY, Town Center, PETS-2009, WildTrack), we apply pre-computed homography matrices or calibration files to project pixel-space detections into metric bird's-eye view coordinates. For datasets lacking explicit camera models (Edinburgh, Grand Central, KITTI, HERMES), we use the native world coordinates provided by the original annotations. We apply the same preprocessing pipeline used for tracking data produced from the video generation models, including 5-second temporal windowing, matching the duration of the synthetic video. We compute the full suite of trajectory kinematics, social interaction, and video fidelity metrics (Section 4) on each processed dataset, aggregating results across all scenes to establish the reference distribution ranges reported in our evaluation.

Details on Flow. As discussed in the paper, an inverse relationship is expected where increasing crowd density results in decreasing average walking speeds [85]. The Fruin

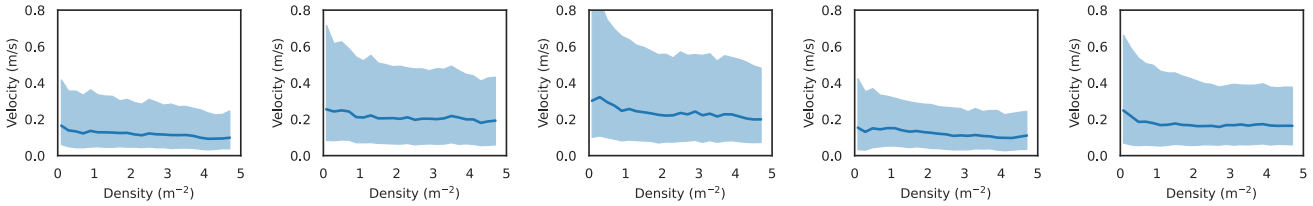


Figure 7. Fundamental diagram plots for Crowded (**Cr.**) pedestrian density in the T2V benchmark, showing the relationship between pedestrian flow and density simulated by different models. The center line and error bounds represent the median and Q1/Q3 quartiles. Plotting code adapted from Minartz et al. [63].

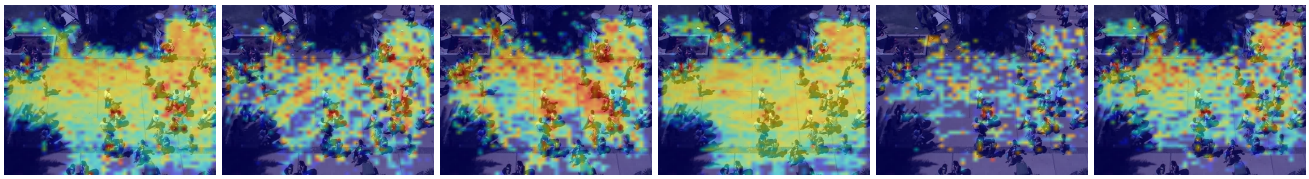


Figure 8. 2D histograms (heatmaps) of pedestrian positions for the UNIV scene from the I2V benchmark. Each subfigure shows the spatial distribution of pedestrian locations for ground truth and different models.

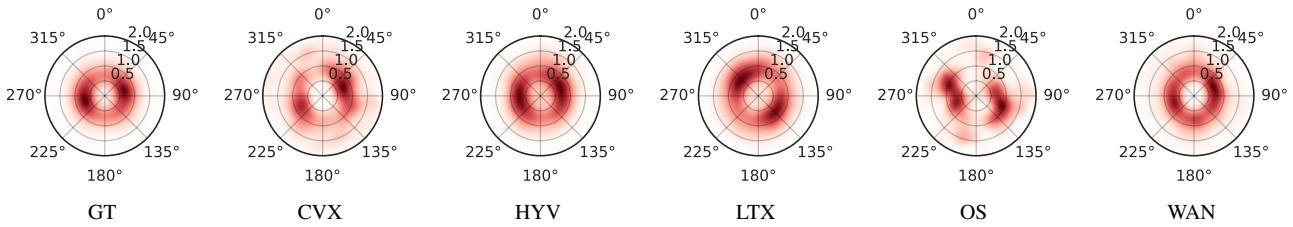


Figure 9. Polar histograms of nearest neighbor (NN) relative positions for UNIV scene from the I2V benchmark. Each subfigure shows the angular distribution of the nearest neighbor with respect to the focused agent for different models and ground truth. Plotting code adapted form Minartz et al. [63].

level of service (LOS) [29] provides an intuitive understanding of different crowd densities:

- LOS A, >13 ft^2/ped (<0.83 ped/m^2)
- LOS B, 10-13 ft^2/ped (.83-1.08 ped/m^2)
- LOS C, 6-10 ft^2/ped (1.08-1.79 ped/m^2)
- LOS D, 3-6 ft^2/ped (1.79 - 3.59 ped/m^2)
- LOS E, 2-3 ft^2/ped (3.59 - 5.38 ped/m^2)
- LOS F, <2 ft^2/ped (> 5.38 ped/m^2)

LOS A corresponds to free standing and circulation without disturbing others. LOS C corresponds to restricted circulation but still within the range of comfort. LOS E corresponds to serious discomfort where physical contact with others is unavoidable.

Figure 7 shows the fundamental diagrams for the Croweded (Cr.) density of the T2V benchmark. The maximum density on these plots of 5 people per sq. m results in shoulder-to-shoulder spacing with highly restricted movement. All five T2V models roughly capture the expected decreasing trend. However, the decrease in walking speed tends to plateau for all models above 2-3 ped/m^2, which does not reflect the expected behavior.

Real-World Interpretation of Velocity.

The $\mathcal{M}_{\text{vel}}$ and $\mathcal{M}^{\text{E}}_{\text{vel}}$ metrics reported in the main paper include all pedestrians in the scene. Since each scene contains some percentage of stationary pedestrians (given by $\mathcal{M}_{\text{stat}}$), this decreases the average walking speed. Here we analyze the walking speeds of only agents that have a non-

Table 8. Mean agent speed (in m/s) $\pm$ standard deviation for non-stationary agents (displacement > 0.2m).

Benchmark	Scene/Category	GT	WAN	HYV	CVX	LTX	OS
I2V	**ETH**	1.29 ± 0.49	1.09 ± 0.61	1.93 ± 1.73	2.43 ± 1.55	1.78 ± 1.19	1.82 ± 1.13
	HOTEL	1.38 ± 0.51	1.46 ± 0.78	1.05 ± 0.60	1.84 ± 1.16	1.82 ± 0.98	1.83 ± 0.86
	UNIV	1.30 ± 0.65	1.03 ± 0.64	0.89 ± 0.57	1.01 ± 0.64	1.47 ± 0.72	0.92 ± 0.64
	ZARA1	1.51 ± 0.49	1.27 ± 0.82	1.59 ± 0.95	2.04 ± 1.21	1.80 ± 0.95	1.95 ± 0.89
	ZARA2	1.57 ± 0.58	1.04 ± 0.67	1.41 ± 0.53	1.87 ± 1.14	1.50 ± 0.65	1.55 ± 0.66
T2V	**Sp.Di.**	—	0.89 ± 1.07	1.40 ± 1.32	0.64 ± 0.64	1.08 ± 1.16	0.73 ± 0.82
	Cr.	—	0.53 ± 0.76	0.71 ± 0.97	0.42 ± 0.59	0.80 ± 0.99	0.40 ± 0.54
	Mo.	—	0.61 ± 0.90	0.76 ± 1.04	0.47 ± 0.60	0.93 ± 1.07	0.46 ± 0.62
	Sp.	—	0.71 ± 0.91	1.19 ± 1.40	0.61 ± 0.72	1.02 ± 1.18	0.62 ± 0.74
	Di.	—	0.76 ± 0.89	0.76 ± 0.94	0.48 ± 0.60	1.03 ± 1.14	0.53 ± 0.64
	Mu.	—	0.61 ± 0.90	0.76 ± 1.04	0.47 ± 0.60	0.93 ± 1.07	0.46 ± 0.62
	Co.	—	0.53 ± 0.76	0.71 ± 0.97	0.42 ± 0.59	0.80 ± 0.99	0.40 ± 0.54

Table 9. Complete I2V evaluation metrics.

Dataset	Model	Trajectory Kinematics			Social Interaction					Fidelity	
		$\mathcal{M}^{\mathrm{E}}_{\mathrm{vel}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{acc}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{dist}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{coll}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{stat}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{pop}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{nn}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{flow}} \downarrow$	$\mathcal{M}^{\mathrm{E}}_{\mathrm{disp}} \downarrow$	$\mathcal{M}_{\mathrm{mot}} \uparrow$
ETH	WAN	**0.427**	**0.204**	**0.081**	0.001	0.192	0.360	0.185	1.331	0.218	0.476
	HYV	0.713	1.947	0.209	**0.000**	0.157	0.200	0.333	0.551	**0.061**	**0.485**
	OS	0.660	1.785	0.278	**0.000**	**0.047**	0.211	0.155	**0.162**	0.092	0.473
	LTX	0.815	2.379	0.841	0.000	0.318	0.800	**0.131**	3.104	0.367	0.478
	CVX	1.666	0.489	1.002	0.006	0.327	**0.085**	0.329	0.420	0.143	0.479
UNIV	WAN	**0.383**	1.027	0.554	0.081	0.189	1.934	0.004	0.004	0.110	0.500
	HYV	0.436	0.525	0.349	0.063	**0.021**	0.819	0.001	0.002	0.189	0.505
	OS	0.778	0.874	0.837	0.084	0.877	2.236	0.012	0.005	0.569	0.482
	LTX	0.402	**0.240**	**0.245**	**0.029**	0.223	**0.715**	**0.001**	**0.001**	**0.074**	**0.510**
	CVX	0.434	0.968	0.603	0.111	0.301	2.297	0.004	0.005	0.099	0.488
HOTEL	WAN	**0.214**	0.696	0.288	0.013	**0.016**	0.458	0.067	**0.052**	0.196	**0.517**
	HYV	0.402	0.212	0.369	0.023	0.175	0.441	0.048	0.183	**0.031**	0.499
	OS	0.459	**0.108**	0.487	0.023	0.247	0.815	0.397	0.209	0.504	0.498
	LTX	0.637	0.406	**0.048**	**0.011**	0.158	**0.376**	**0.040**	0.097	0.037	0.494
	CVX	0.540	0.632	0.471	0.023	0.328	0.703	0.201	0.209	0.272	0.499
ZARA1	WAN	0.603	0.947	0.318	**0.013**	0.231	**0.057**	0.138	0.177	0.569	0.500
	HYV	**0.336**	**0.221**	**0.179**	0.034	**0.134**	0.198	0.104	0.283	0.185	0.503
	OS	0.697	0.264	0.320	0.056	0.386	0.425	0.165	0.146	0.659	0.491
	LTX	0.569	0.557	0.571	0.086	0.239	0.866	0.079	0.514	**0.021**	**0.512**
	CVX	0.865	0.744	0.375	0.056	0.209	0.330	**0.076**	**0.104**	0.175	0.499
ZARA2	WAN	0.658	1.034	0.610	**0.037**	0.168	0.698	0.026	0.356	0.560	0.494
	HYV	0.205	0.291	0.336	0.044	0.189	0.679	0.016	0.120	0.325	0.506
	OS	0.150	0.485	0.386	0.072	**0.023**	0.891	0.048	0.377	0.745	0.486
	LTX	**0.128**	**0.154**	**0.250**	0.076	0.375	**0.085**	**0.004**	**0.007**	**0.150**	**0.520**
	CVX	0.534	0.695	0.655	0.075	0.233	1.046	0.076	0.400	0.156	0.492

zero overall displacement in order to give a more intuitive analysis of how realistic the pace is (Table 8).

A number of peer-reviewed studies report statistics on the distributions of human walking speeds in unobstructed environments, i.e., not restricted by the presence of other humans or obstacles [14, 65, 90]. They range from 0.8 m/s for elderly populations to as high as 1.6 m/s for healthy adult males, with an average walking speed reported around 1.3 m/s. For I2V, we can compute the ground truth walking speeds as a point of comparison. Table 8 reports the results on the GT datasets from the ETH/UCY scenes. The walking speeds range from 1.29 m/s (ETH) to 1.57 m/s (ZARA2), which strongly agree with the results expected from the literature. The Sparse/Directional subset (*Sp./Di.*) the T2V benchmark corresponds to the same type of scenery as ETH/UCY and serves as a good point of comparison.

In the I2V benchmark, WAN is the model that produces the closest match overall to GT walking speed distributions. The other models have variable performance, with some scenes very close to the GT distribution and others clearly too fast or too slow, although still within a range of physically plausible movement speeds. Speeds as high as 2.43 m/s (CVX, ETH scene) approach running speeds rather than walking, which appears to result from a scale mismatch where the model generates humans that are too large relative to the scene and therefore walk too quickly in the real-world coordinate system.

In the T2V benchmark, HYV is the model that closest approximates the expected speed distribution, averaging 1.40 m/s in the Sp./Di.category. All of the other models produce *walking speeds which are generally too slow* (especially CVX), although the standard deviation is high enough that many pedestrians fall within the normal range. This result is especially interesting given the feasible walking speeds produced in the I2V benchmark.

F. Additional Qualitative Results

F.1. T2V Scene Variety

Figure 10 shows additional examples of trajectory extraction and BEV coordinates from T2V generations by various models with all three interaction types. We note the high degree of success of the multi-object tracking and the realistic metric scales computed using the process described in the *Method* section. Figure (b) demonstrates that even with large degrees of camera motion, the use of frame-wise camera extrinsics from VGGT allows a consistent world coordinate system to be established such that 1) the walking trajectories remain aligned on a *straight line following the red path*, despite the pixel-coordinate paths taking on a curve due to the camera motion; and 2) the seated people in the bottom right corner retain stationary locations. Figure (d) demonstrates the significant scene and behavior variety that can be obtained through text prompts alone, especially in scenes that would be challenging or impossible to specify in conventional simulation software. Figure (e) demonstrates that crowded scenes with over 100 pedestrians remain successfully tracked, showing the power of this method to extract large numbers of trajectories in a single generation.

F.2. Failure Modes

Figures 11 and 12 illustrate examples of failure modes for the image-to-video and text-to-video benchmarks, respectively.

Common Failure Modes

- Disappearing Pedestrians (Figures 11b and 12b): One of the most prevalent issues is the spontaneous vanishing of pedestrians mid-trajectory .
- Merging/Colliding People (Figures 11d and 12d): Rather than exhibiting realistic collision avoidance behavior, pedestrians frequently merge together or occupy the same spatial location.
- Visual Distortions (Figures 11e and 12e): Degradation in pedestrian appearance may render individuals unrecognizable or untrackable by the multi-object tracker. Distorted objects that are neither pedestrian nor vehicle sometimes appear.

I2V-Specific Failure Modes

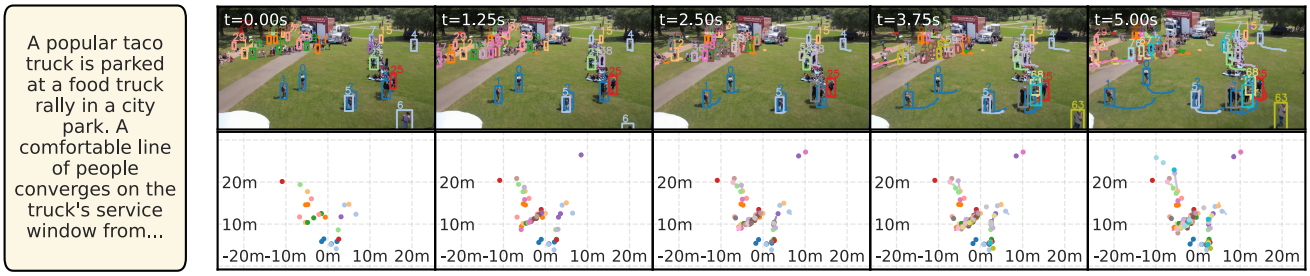
- Unwanted Camera Motion (Figure 11a): Models may introduce camera movement despite static camera prompts. Since we use ETH/UCY pre-computed homography matrices, this represents a failure mode for the I2V benchmark, although the T2V benchmark is designed to expect camera motion.
- Scene Changes (Figure 11c): Models may spontaneously change scene from the input image.
- Scene Understanding (Figure 11f): Models may inappropriately animate static objects, such as moving parked cars in pedestrian-only zones. This suggests limitations in the latent representation of the input condition image.

T2V-Specific Failure Modes

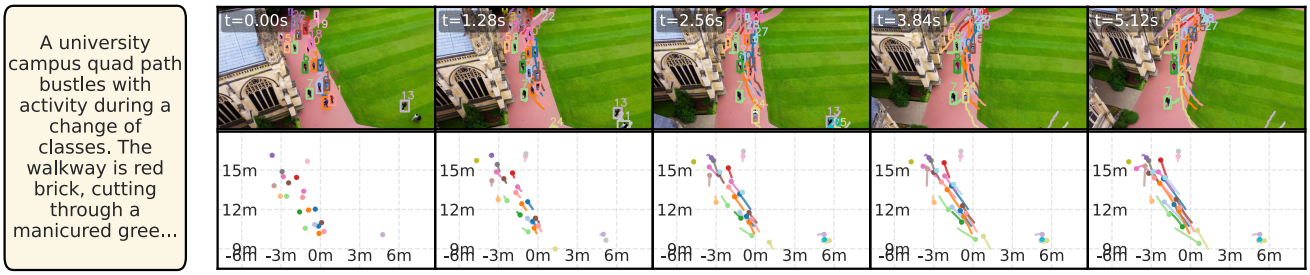
- Pixelated Masses (Figure 12a): In crowded scenarios, models often fail to render distinct individuals, instead producing untrackable, fluid-like pixelated masses.
- Sped Up/Time-Lapse Effects (Figure 12c): Models sometimes generate unwanted temporal acceleration, causing pedestrians to appear as motion blur streaks. This doesn't affect the velocity metrics ($\mathcal{M}_{\text{vel}}, \mathcal{M}^{\text{E}}_{\text{vel}}$) as the blurred people are not detected by the MOT model.
- Improbable Scene Generation (Figure 12f): T2V models may create impossible scenarios with inappropriate semantic context or 3D physicality.

Table 10. T2V evaluation metrics across trajectory kinematics, social interaction, and video fidelity categories. The highest values are indicated in bold in order to emphasize trends between density and interaction categories. The real-world reference (Ref.) includes the ($\mu \pm \sigma$) of each metric computed across 10 public pedestrian benchmark datasets, as described in Section 4.

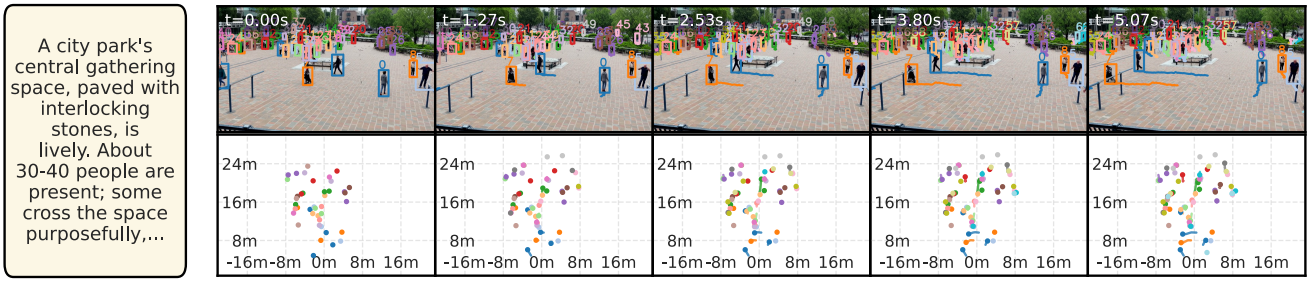
Model	Category	Trajectory Kinematics			Social Interaction					Video Fidelity		
		$\mathcal{M}_{\text{vel}}$ (m/s)	$\mathcal{M}_{\text{acc}}$ (m/s^2)	$\mathcal{M}_{\text{dist}}$ (m)	$\mathcal{M}_{\text{coll}}$ (%)	$\mathcal{M}_{\text{stat}}$ (%)	$\mathcal{M}_{\text{pop}}$ (Count)	$\mathcal{M}_{\text{flow}}$ (1/m/s)	$\mathcal{M}_{\text{nn}}$ (m)	$\mathcal{M}_{\text{disp}}$ (%)	$\mathcal{M}_{\text{mot}} \uparrow$	$\mathcal{M}_{\text{geo}} \uparrow$
	Ref.	*.91±.84*	*.65±.98*	*3.62±3.56*	*1.19±11.28*	*.19±.39*	*13.77±19.07*	*.54±.45*	*1.18±1.51*	-	-	-
WAN	Cr.	**0.564**	**0.858**	**2.304**	**6.077**	0.199	**136.693**	**1.541**	0.637	**30.496**	0.531	2.698
	Mo.	0.556	0.694	1.327	1.950	0.330	25.514	0.230	1.013	25.732	0.619	4.687
	Sp.	0.520	0.554	1.241	1.272	**0.338**	4.860	0.029	**1.806**	12.398	**0.676**	**6.181**
	Co.	0.452	0.689	1.576	**8.047**	**0.261**	52.518	1.053	0.600	21.863	**0.552**	3.009
	Di.	**0.714**	**1.035**	**2.746**	6.267	0.149	49.653	**2.861**	0.661	29.608	0.545	2.650
	Mu.	0.529	0.784	2.140	2.639	0.237	**67.923**	0.408	**0.868**	**33.995**	0.540	**3.271**
HYV	Cr.	0.553	0.823	1.517	**11.926**	0.253	**72.206**	**2.000**	0.582	27.910	0.526	2.234
	Mo.	0.897	**1.131**	**1.712**	5.836	0.256	27.582	1.015	0.786	**33.434**	0.584	1.807
	Sp.	**0.997**	1.055	1.365	2.749	**0.289**	4.889	1.912	**1.597**	30.685	**0.618**	**2.403**
	Co.	0.606	0.891	1.453	**12.446**	0.277	32.947	1.687	0.650	26.134	**0.556**	2.003
	Di.	**0.709**	**0.962**	**1.873**	11.763	0.187	35.169	**2.670**	0.580	26.135	0.545	1.650
	Mu.	0.649	0.876	1.378	5.945	**0.295**	**37.985**	0.884	**0.862**	**34.069**	0.534	**2.641**
OS	Cr.	0.310	0.466	1.512	**3.278**	0.269	**42.129**	**0.245**	0.941	**33.112**	0.535	3.091
	Mo.	**0.522**	**0.589**	**1.833**	1.909	0.288	20.731	0.183	1.156	32.659	0.585	3.700
	Sp.	0.477	0.431	1.738	0.535	**0.310**	4.428	0.051	**1.688**	12.875	**0.657**	**4.591**
	Co.	0.310	0.432	1.302	**4.124**	**0.329**	20.268	**0.283**	0.883	28.161	0.559	**4.008**
	Di.	**0.476**	**0.590**	**2.111**	2.324	0.193	19.401	0.239	0.984	30.445	**0.562**	2.378
	Mu.	0.371	0.491	1.515	1.927	0.297	**28.519**	0.143	**1.229**	**35.277**	0.550	3.546
LTX	Cr.	0.760	**1.207**	**2.207**	**9.827**	0.212	**66.856**	**1.489**	0.606	27.200	0.555	1.411
	Mo.	**0.904**	1.188	1.706	2.955	0.287	24.829	0.448	0.961	41.704	**0.574**	**1.563**
	Sp.	0.820	1.039	1.402	1.364	**0.317**	6.130	0.164	**1.458**	**42.259**	0.569	1.403
	Co.	0.648	1.004	1.597	9.011	**0.303**	33.325	1.006	0.718	30.121	0.563	1.433
	Di.	**0.966**	**1.370**	**2.518**	**9.094**	0.162	31.218	**1.819**	0.652	30.663	**0.569**	1.258
	Mu.	0.799	1.208	1.992	5.114	0.245	**36.675**	0.717	**0.837**	**35.020**	0.552	**1.625**
CVX	Cr.	0.370	0.629	**1.222**	**6.760**	0.301	**55.856**	**0.605**	0.702	**34.178**	0.494	**3.020**
	Mo.	0.468	**0.698**	1.163	3.803	0.319	22.971	0.314	0.926	34.152	0.553	2.244
	Sp.	**0.494**	0.644	1.036	2.642	**0.322**	3.664	0.088	**1.546**	20.340	**0.598**	2.053
	Co.	0.333	0.546	0.912	**7.299**	**0.382**	23.899	0.470	0.816	30.486	**0.519**	**3.033**
	Di.	**0.441**	**0.719**	**1.511**	6.227	0.234	27.381	**0.617**	0.746	30.829	0.513	2.541
	Mu.	0.402	0.655	1.166	4.367	0.309	**34.175**	0.425	**0.835**	**37.777**	0.503	2.876



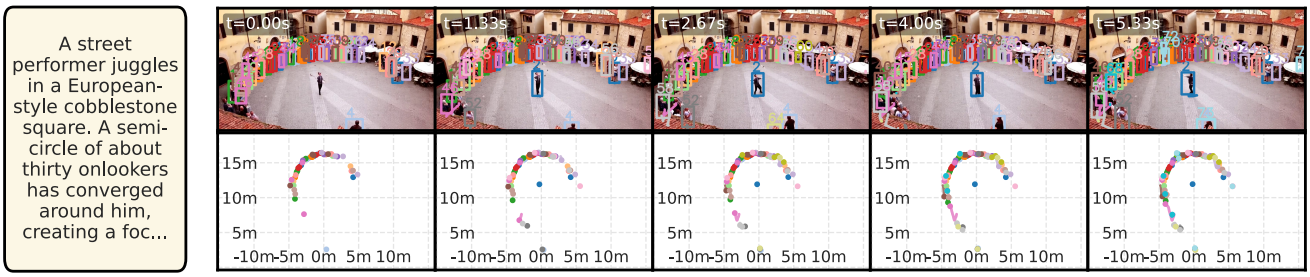
(a) CogVideoX1.5 (CVX), Co.



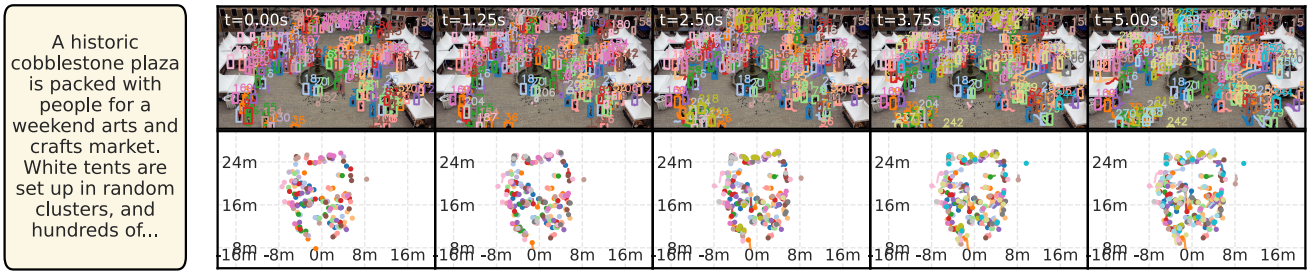
(b) HunyuanVideo (HYV), Di.



(c) LTX-Video (LTX), Mu.



(d) Open-Sora 2.0 (OS), Co.



(e) Wan2.1 (WAN), Mu.

Figure 10. Additional qualitative examples showing a variety of specified interaction types from the T2V prompt suite.

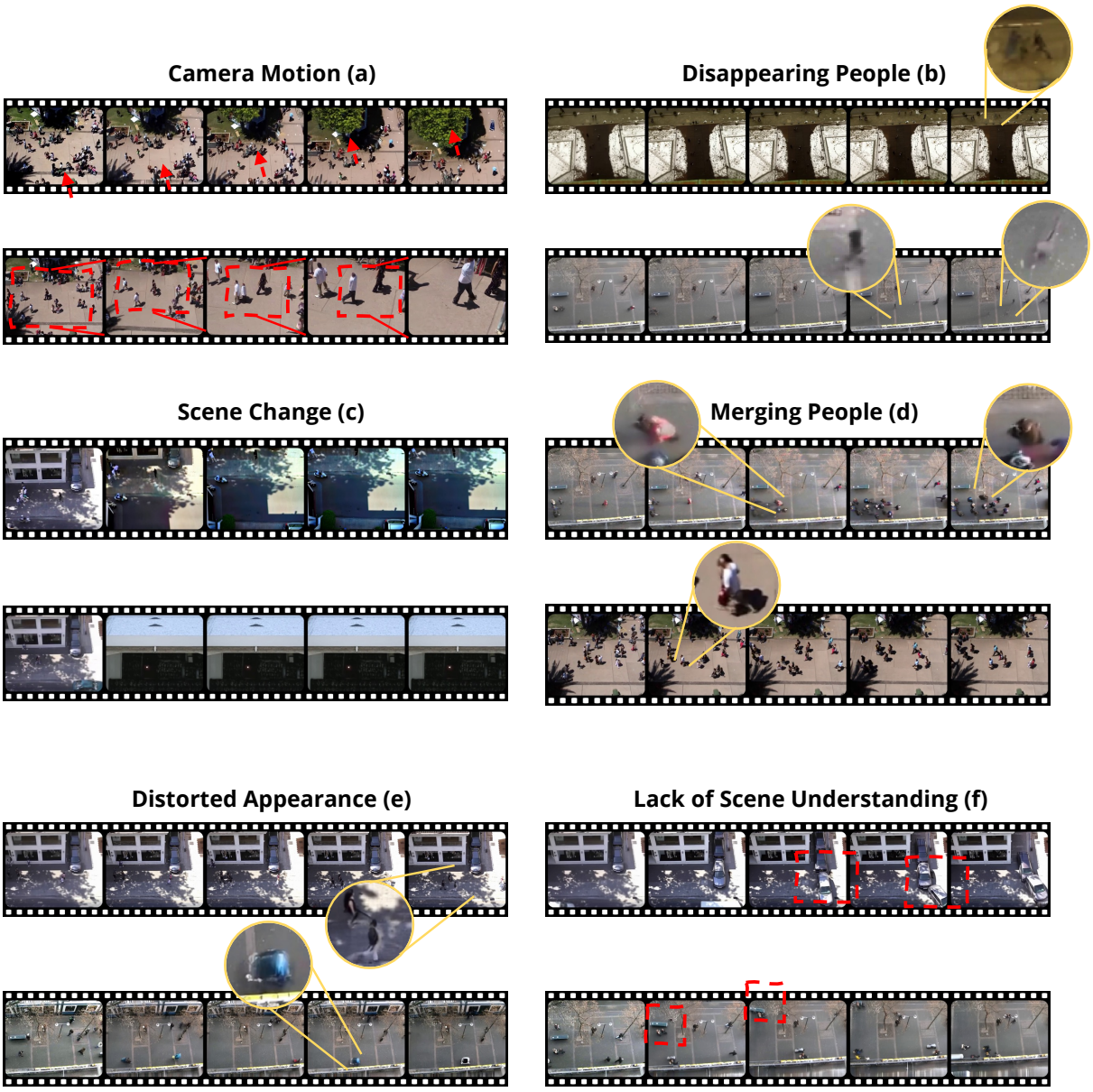


Figure 11. Common failure modes observed in I2V generations. (a) The camera perspective may pan (top) or zoom (bottom) despite the request for a stationary view in the positive and negative prompts. We filter out these videos as it prevents using the ETH/UCY homography matrices. (b) People spontaneously disappear from one frame to another or become ghostly. (c) The scene many abruptly change despite starting off as the scene given by the image condition. (d) People who begin as separate agents may merge into one another, which results in disappearing MOT track IDs. (e) People may have elongated or distorted appearance (top). Objects may appear that do not look like either people or vehicles (bottom). (f) A car at the curb which should remain parked moves forward as if driving on a road (top); a bench begins to move as if it is some type of vehicle (bottom).

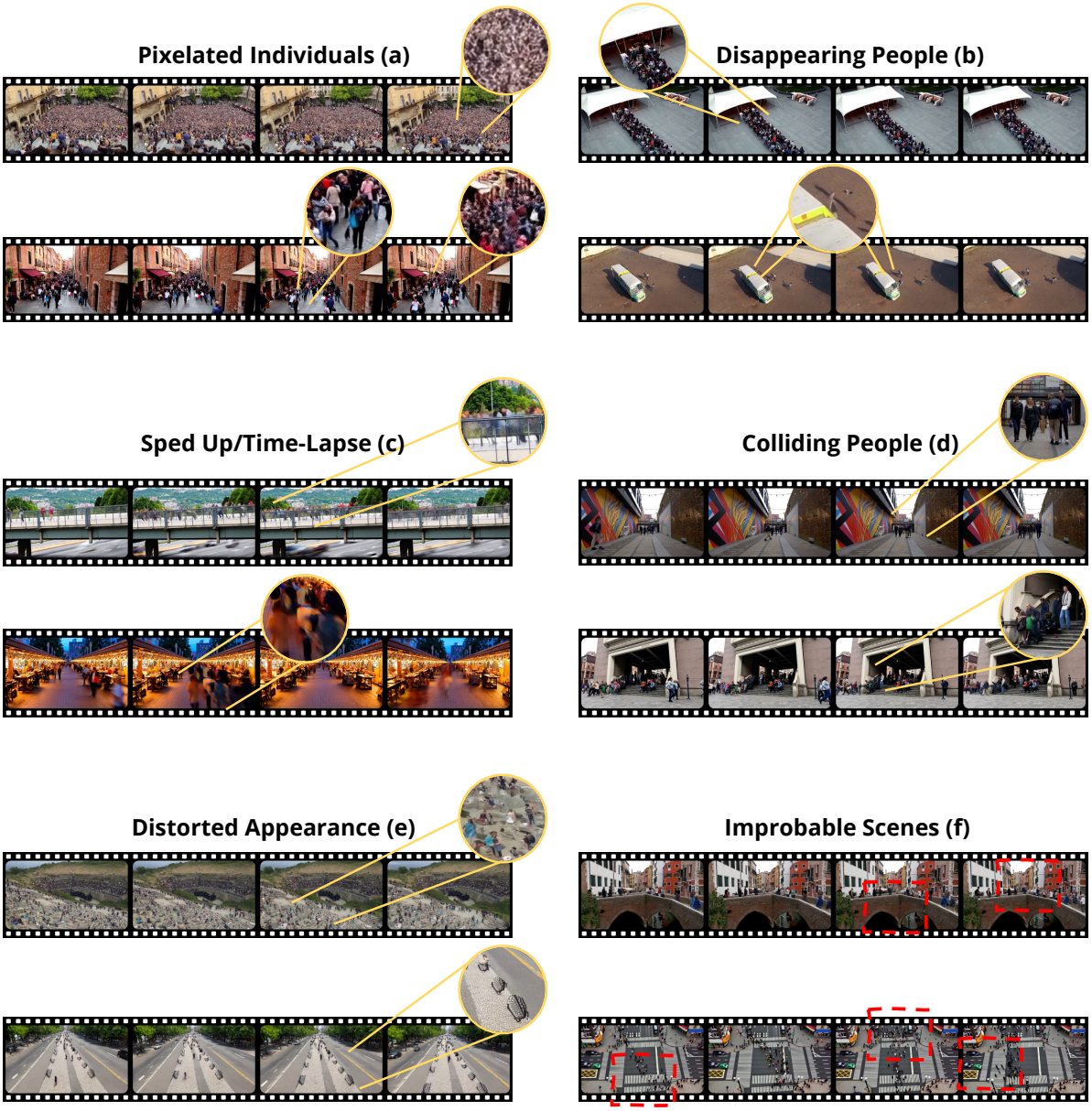


Figure 12. Common failure modes observed in T2V generations. (a) Individuals lose trackability in crowds when the depiction turns into a pixelated mass, which is more prominent for far-away people in the background than close-up people represented by more pixels. (b) Pedestrians in a dense queue disappear as they move through a bottleneck rather than re-emerging on the other side (top); an individual pedestrian in a sparse scene disappearing (bottom). (c) Undesired sped-up or time-lapse effects in generated videos cause high blurring in individual frames which prevents tracking. (d) While colliding pedestrians are more common in dense pedestrian flows (bottom), there are also examples where individuals walk directly into oncoming groups (top). (e) Scenes and/or people may have distorted appearances, which impacts the success of 3D reconstruction and tracking, respectively. (f) Scenes may be physically improbable, both in terms of 3D space (top, ill-defined perspective) or context (bottom, duplicate crosswalks).