SpatialVID: A Large-Scale Video Dataset with Spatial Annotations

Supplementary Material

Overview

This supplementary material provides extended details and additional results complementing the main paper. We elaborate on the data curation pipeline (Sec. A), present in-depth analyses of dataset statistics and semantic properties (Sec. C), and describe additional validation tasks and implementation details (Sec. D). An overview of the video filtering and annotation process is shown in Fig. 1.

A. Details of Our Curation Pipeline

Fig. 1 presents the overall data flow of our video curation process, through which the raw videos are progressively refined into the final SpatialVID dataset. The pipeline integrates automatic quality filtering, geometric annotation, and semantic captioning. Figure 2 further illustrates the duration and quantity distribution of the collected raw data, covering diverse indoor and outdoor scenes such as walking, train rides, and drone flights.

A.1. Score Filtering

To ensure data quality, we apply four complementary filters based on *aesthetics*, *luminance*, *text content*, and *motion intensity*. These filters remove visually poor or unsuitable clips before any geometric or semantic processing, significantly improving the robustness of downstream pose estimation.

Aesthetic Filtering. To quantitatively assess visual appeal, we use a CLIP + MLP aesthetic score predictor [6]. The model assigns a score from 0 to 10, with higher values indicating better quality. For each video clip, the score is averaged across the first, middle, and last frames. Clips with an average score below 4.0 are considered insufficiently appealing and discarded Fig. 3a).

Luminance Filtering. Luminance is calculated for the first, middle, and last frames using the standard formula $L = 0.2126R + 0.7152G + 0.0722B$, where R, G, and B are the respective channel values. Clips with average luminance outside the range [20, 140], either too dark or too bright, are excluded, ensuring that only clips with proper exposure are retained (Fig. 3b).

OCR filter. For text detection, we use the latest release of PaddleOCR [2], which offers high accuracy and robust multilingual support. We processe the first, middle, and last frames of each clip to detect text regions, computing the ratio of text area to frame size. Clips where the text area exceeded 30% are removed, as these are considered informational rather than visual (Fig. 3c).

Motion Filtering. We use lightweight VMAF [16], which is integrated into FFmpeg and outputs a valid motion score ranging from 2.0 to 14.0 (Fig. 3d).

A.2. Geometry Pipeline

We employ **MegaSaM** as our primary geometric reconstruction engine. As shown in Fig. 4, it achieves superior accuracy and robustness compared with DROID-SLAM [8], COLMAP [5], and Fast3R [13], while being faster than MonST3R [14]. Unlike VGGT [10], MegaSaM also maintains stability in feature-sparse scenes. We utilize three trajectory statistics for quality assessment:

Move Distance (MoveDist). Total camera travel distance, computed as the sum of Euclidean distances between consecutive camera centers.

Rotation Angle (RotAngle). Cumulative angular displacement across frames, capturing the extent of viewpoint change.

Trajectory Turns (TrajTurns). Number of significant turns, estimated by counting local extrema in the sequence of orientation angles relative to a start–end reference direction.

A.3. Caption Pipeline

The captioning process integrates visual-language reasoning and structured text generation in two stages.

Stage 1: Visual Parsing. We use Gemini-2.0-flash [7] to analyze sampled frames (1 fps), producing an initial description of camera motion and a summary of scene layout. The prompting format is illustrated in Fig. 5.

Stage 2: Language Refinement. The outputs, along with calibrated camera poses, are then refined using Qwen3-30B-A3B [12]. This stage yields (1) concise scene abstracts, (2) immersive shot-level narratives, and (3) structured semantic tags describing scene type, lighting, weather, crowd density, and time of day. Additionally, *Motion Trends* labels (e.g., pan, dolly, rotate, steady) summarize camera dynamics. Distributions of these tags are shown in Fig. 8. The prompt template used for the large-language refinement is shown in Fig. 6.

A.4. Instruction Examples

Examples of motion instructions are illustrated in Fig. 7. Each instruction corresponds to a specific type of camera movement derived from pose dynamics, providing an interpretable bridge between geometric motion and textual representation. The visualizations demonstrate how camera translations and rotations are systematically mapped to human-readable motion terms, ensuring clarity and consistency across clips.

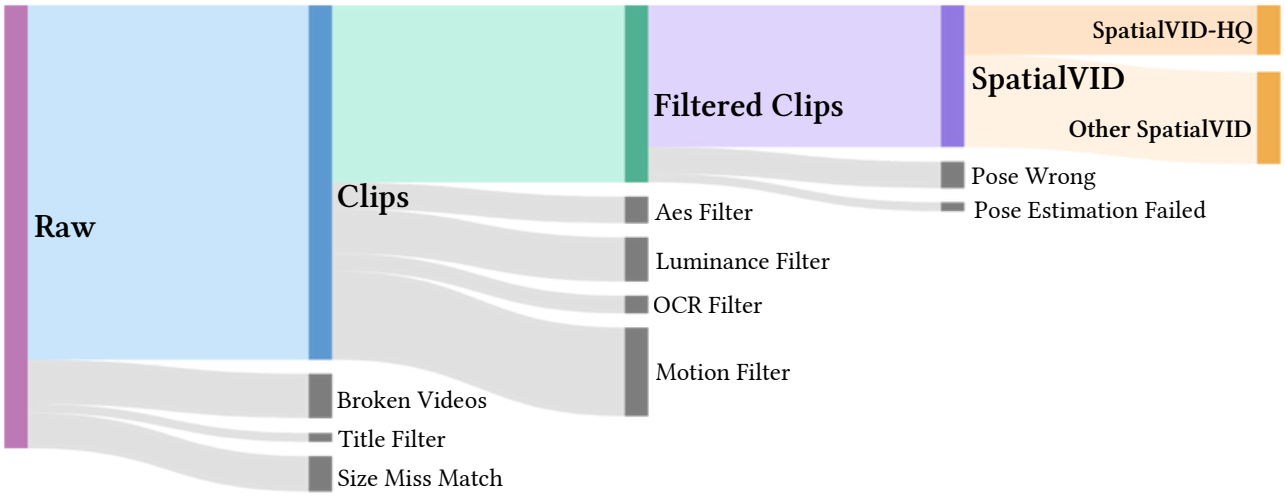


Figure 1. **The data flow of video filtering**. Raw videos are first pre-filtered to exclude content with quality defects, incorrect dimensions, or irrelevant titles. The remaining videos are segmented into clips, which are then ranked via a hierarchical scoring strategy integrating aesthetics metrics, luminance, OCR, and motion values. High-scoring clips undergo a dual annotation pipeline to capture both spatial structure and semantic information, yielding the final SpatialVID dataset. This pipeline is also employed to curate a high-quality subset (SpatialVID-HQ) with a more balanced category distribution.

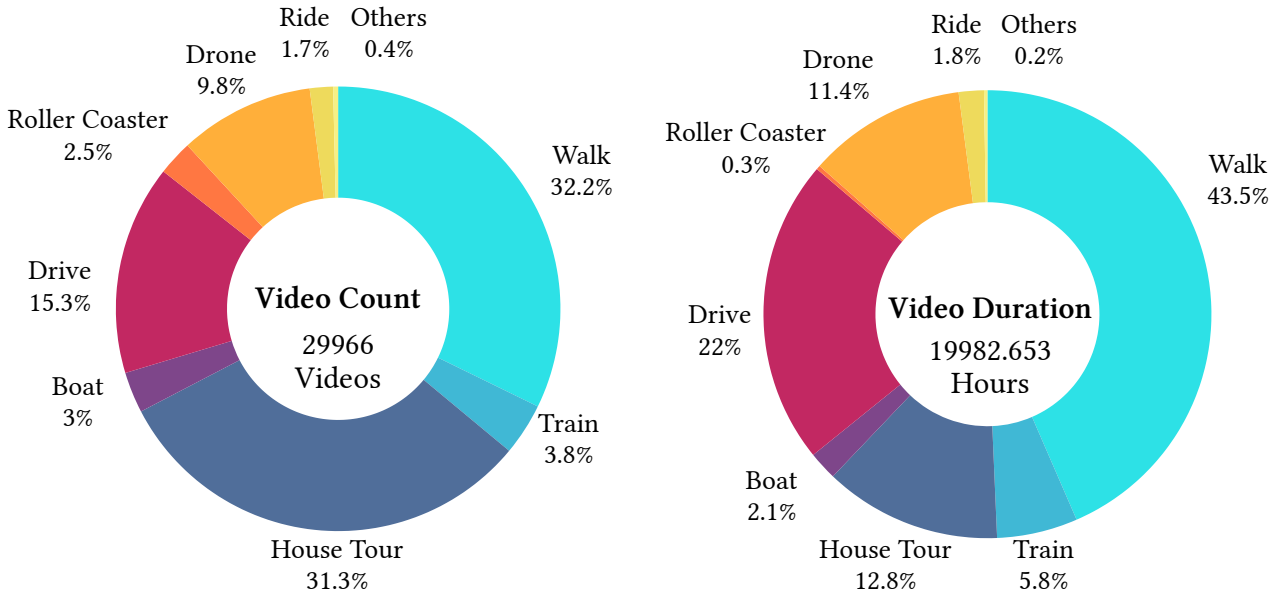
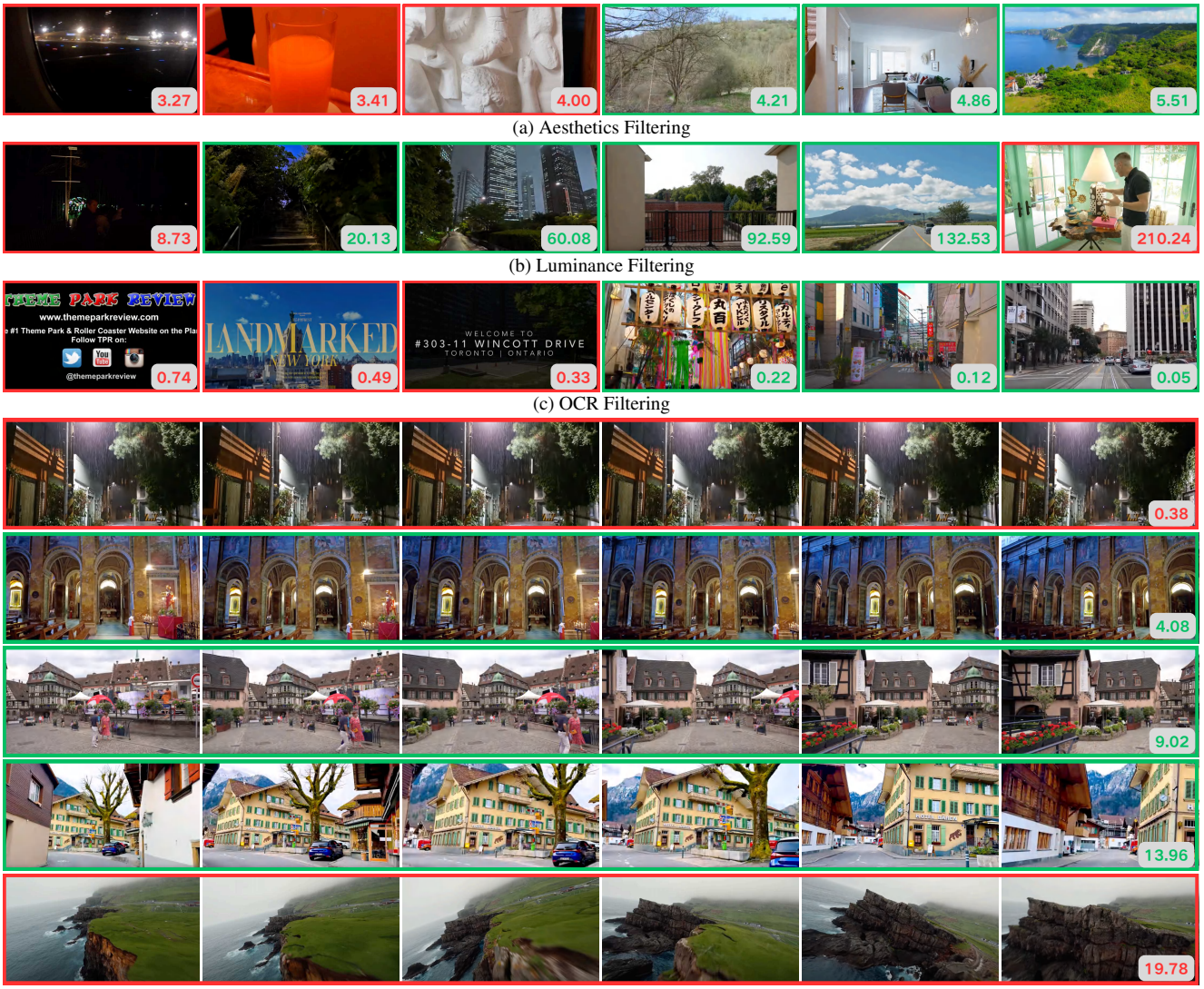


Figure 2. **Statistics of pre-filtered videos** (*aka*. the "Clips" in the Fig. 1). The left panel shows the quantity distribution of raw videos, while the right panel presents the duration distribution. These charts illustrate the variety of shooting contexts, including indoor (house tour) and outdoor (walking, train, drone, etc.) scenarios, demonstrating broad coverage of different shooting carriers and environments.

B. Safety Considerations

Our dataset is collected from publicly available web sources and covers a wide range of real-world scenarios. In the released version, all data are provided with anonymized identifiers and limited metadata exposure.

We emphasize transparent and responsible dataset release and usage, promoting ethical and accountable research practices. Access to the dataset is controlled through a gated release process. Applicants are required to provide relevant identity and research-use information before access is granted. They must also explicitly commit to complying with applicable platform policies and adhering to the dataset access requirements.



(a) Aesthetics Filtering

(b) Luminance Filtering

(c) OCR Filtering

(d) Motion Filtering

Figure 3. **Video filtering strategies**. Videos are filtered based on various quality criteria (*Aesthetics*, *Luminance*, *OCR*, and *Motion*). The number in the bottom-right corner of each clip represents its score for the corresponding quality filter. Clips with green boxes are retained, while those with red boxes are discarded due to scores below the threshold.

C. Details of Dataset Analysis

This section provides extended analyses of the SpatialVID dataset, focusing on semantic composition, caption statistics, and qualitative examples. We examine the distributions of camera motion and scene attributes, analyze caption diversity, and visualize representative samples from the dataset.

C.1. Semantic Analysis

Camera Motion Distribution. Fig. 8 shows the distribution of camera motion directions across SpatialVID and its high-quality subset, SpatialVID-HQ. The original dataset contains a wide range of motion types, including forward, lateral, and rotational movements, but their distribution is not well-balanced. In contrast, SpatialVID-HQ displays a more balanced distribution, mitigating bias toward any single motion direction and offering improved diversity for motion-conditioned generation or control tasks.

Caption Length and Enrichment. We provide multi-level captions for each video, including motion-oriented descriptions, concise scene summaries, and immersive narratives. To evaluate caption quality, we compare the length distributions of original and enhanced captions (Fig. 9). Both motion and scene captions show notable length increases after refinement, reflecting richer context and more detailed spatial reasoning introduced by our LLM-based generation

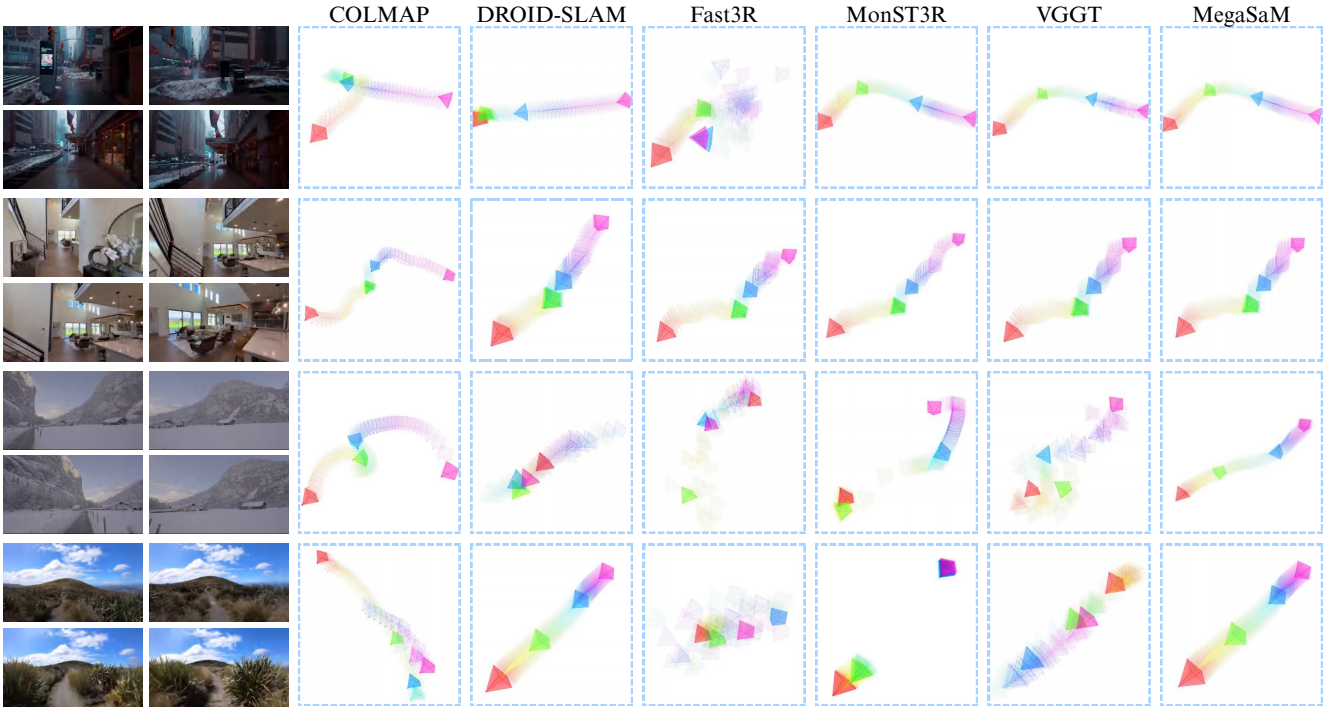


Figure 4. **Comparison of MegaSaM with other SLAM/3D reconstruction methods**. We visualize the trajectories predicted by six representative methods. The color order ROYGBV corresponds to the progression from the initial to the final time step.

You are given a sequence of video frames in chronological order. Analyze them carefully and generate two distinct captions based on the following instructions:

1. Camera Motion Caption:
From the perspective of the camera operator, describe the entire motion trajectory of the camera throughout the clip using precise cinematography terminology (e.g., static, pan, tilt, dolly, handheld, crane, aerial, zoom, etc.).
Do NOT assume the camera starts in a "static" position just because it appears stationary in the first frame.Only describe the camera as stationary if there is no visual change across multiple consecutive frames.
Instead, focus on changes between frames to infer movement. Describe motion state transitions, not frame-by-frame repetition (e.g., do not say “the camera moves forward again” if it’s continuous). For example:
- Starting with a dolly forward along a straight path,
- Then transitioning into a slow right-hand pan,
- Or shifting from handheld walking movement to a stationary pivot tilt.
Include brief environmental context where relevant to clarify direction or intent (e.g., "The camera dollies forward through a narrow alleyway, then smoothly turns left at the intersection").
Keep the final caption concise, between 50–100 words, focused only on motion and its evolution over time.

2. Scene Description:
Provide a rich, holistic description of the visual content. Include:
- Main subjects and dynamic objects: who or what is present, and what they are doing (e.g., a cyclist rides past from left to right, a group of people gather near a bench),
- Background/environment: setting (urban street, forest trail, indoor space), notable landmarks or structures,
- Lighting and atmosphere: time of day, weather conditions, mood (e.g., golden-hour lighting, overcast sky casting soft shadows, neon-lit nighttime scene),
- Overall tone or emotion conveyed by the scene.

Avoid focusing on individual frames—describe the general impression and ongoing activity across the entire clip. Aim for around 100 words, balancing detail and conciseness.

Output Format:
Do not include any explanations or extra text before or after your response.

Begin directly with:
1. Camera Motion Caption: ...
followed by
2. Scene Description: ...

Figure 5. **Visual-language model (VLM) prompt**. The template guides Gemini-2.0-flash to describe both the visual content and coarse camera motion in natural language, enabling structured parsing of dynamic scenes.

process.

Hierarchical Scene Tags. Fig. 10 visualizes the structured semantic attributes extracted from enhanced captions. The sunburst chart summarizes the distribution of five primary attributes—weather, time of day, crowd density, lighting, and scene type. The hierarchical scene-type branch

You are given a video sequence with camera trajectory data representing the camera's movement through a scene.

The data consists of:
Camera Motion Caption: A basic description of how the camera moves.
Scene Description: A detailed visual summary of the environment.
Camera position data: Three lines, representing the sequence of the camera's x-coordinate, y-coordinate, and z-coordinate. These values are derived from normalized 3D pose data using the following formula: poses = np.round(poses / (max_value - min_value) / min_abs_value).astype(int); Each value is then multiplied by 1,000,000 and rounded to the nearest integer.
Motion intensity: An integer that indicates the level of camera movement, where a value of 0 means the camera is static, 1 indicates slight movement, and 2 or higher represents normal or noticeable motion. In tasks such as Optimized Camera Motion Caption and Main Motion Trend Summary, this intensity value should be used to qualify the degree of motion described — for example, using "slight forward translate" when the intensity is 1.

Your Tasks:
1. Optimized Camera Motion Caption
Generate a refined motion caption **from the perspective of the camera itself**, using only the **camera position data** to determine movement direction and dynamics.

Use the following rules to interpret motion:
x increasing: camera moves right
x decreasing: camera moves left
y increasing: camera moves down
y decreasing: camera moves up
z increasing: camera moves forward
z decreasing: camera moves backward

Analyze the full trajectory over time to capture acceleration, deceleration, or steady motion. Integrate scene context but prioritize accuracy based on numerical data. Avoid vague phrases like "zoom out" unless it's clearly due to focal length change — here, use translation terms instead.
If motion intensity is 0, describe the fixed viewpoint and what the camera observes from that vantage point, incorporating compositional or environmental elements from the original caption. If intensity is 1, reflect subtle movement in the description (e.g., "slight right translate") without exaggerating the motion. For both cases, preserve visual context while aligning with the actual movement level. Avoid mentioning data analysis or detection explicitly — let the description itself reflect the motion state.
Target Length: 50–100 words

2. Scene Abstract Caption
Provide a single-sentence summary that captures:
- Key architectural elements
- Overall atmosphere/style
- Notable design features
Target Length: About 50 words

3. Main Motion Trend Summary
Summarize the general movement using only 1–3 short motion phrases , depending on how many are clearly present. Focus strictly on major, sustained movements — ignore minor fluctuations or brief directional changes. If only one or two movements dominate, list only those. Use directional translation terms (e.g., forward translate, left translate, upward drift)

4. Scene Keywords
Extract up to 4 keywords summarizing the key aspects of the scene. Include one term that broadly describes the scene type. Use nouns/noun phrases related to weather, place, time, lighting, scene type. Avoid adjectives/gerunds except for weather. Example: sunset, foggy, marketplace, city street, village

5. Immersive Shot Summary
Blend Optimized Camera Motion Caption and Scene Description evenly — do not focus more on the camera or the scene alone. Describe the visuals as if someone is watching a moving image unfold. Use descriptive, cinematic language that evokes imagery and emotion. Keep it concise but expressive — suitable for use in scripts, storyboards, or AI video/image generation. Target Length: 50–100 words

Given Information:
[VQA Captions]

Camera Position Data:
[Camera Poses]

Output Format:

1. Camera Motion Caption:
[From the perspective of the camera holder, with the camera as the subject. Combine camera pose information to describe]

2. Scene Abstract Caption:
[A concise one sentence summary of the scene]

3. Main Motion Trend Summary:
[keywords separated by commas, e.g., forward translate, downward tilt]

4. Scene Keywords:
[word1, word2, word3, ...] (max 5 words)

5. Immersive Shot Summary

Figure 6. **Large-language model (LLM) refinement prompt**. This instruction template conditions Qwen3-30B-A3B to generate coherent, attribute-rich captions aligned with the extracted spatial and motion cues.

covers fine-grained subcategories such as *street*, *park*, *interior*, and *vehicle*. The accompanying word cloud, shaped into the SpatialVID logo, highlights the dataset's emphasis on spatial and motion-oriented vocabulary, with frequently occurring terms like *motion*, *forward*, and *left*.

Multi-Level Caption Design. SpatialVID delivers a versatile caption suite suitable for various research needs: (1) `OptCamMotion` provides concise, machine-friendly kinematic instructions, reducing average caption length from 62.5 to 50.3 words for clean motion supervision. (2) `SceneSummary` offers compact high-level context with an average of 28.6 words. (3) `ShotImmersion` integrates

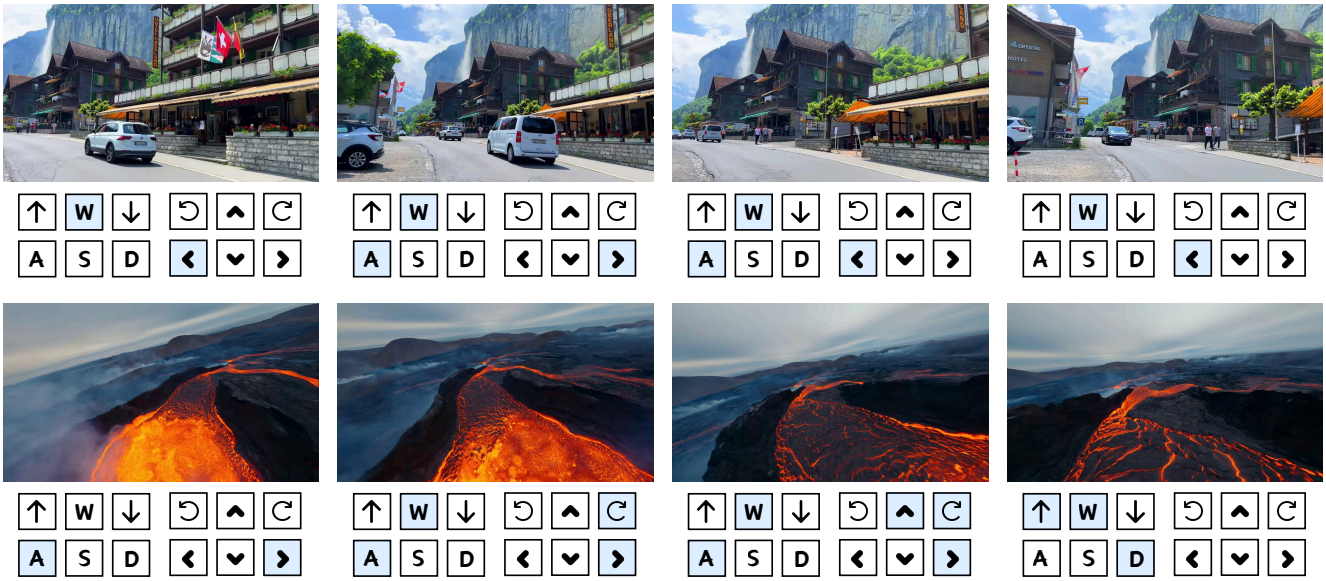


Figure 7. **Examples of motion instructions.** Keyboard-style icons denote camera motion directions. The cluster on the left corresponds to translations: **W** and **S** indicate forward and backward movement, **A** and **D** indicate left and right movement, and $\uparrow, \downarrow$ represent vertical movement. The cluster on the right corresponds to rotations: arrows denote pitch ($\wedge, \vee$) and yaw ($<, >$), while circular arrows indicate roll ($\circlearrowleft, \circlearrowright$). These visual cues intuitively describe camera operations, linking geometric motion to semantic labels.

scene semantics and camera motion into rich narratives averaging 89.7 words, supporting reasoning-intensive tasks such as video understanding and story grounding.

Overall, this structured annotation design ensures both interpretability and flexibility, enabling downstream applications ranging from camera control to multimodal spatial reasoning.

C.2. Examples of SpatialVID

We present qualitative examples in Fig. 11. The selected clips illustrate diverse motion trajectories, scene contexts, and annotation richness. Each sample contains synchronized geometry, caption, and metadata, demonstrating the spatial and semantic consistency maintained throughout the dataset.

D. Validation Tasks

This section presents the implementation details and qualitative results of validation tasks conducted using the SpatialVID dataset. We focus on three representative paradigms, camera-controlled video generation, spatial reconstruction, and view-consistent rendering, to demonstrate the utility and quality of SpatialVID annotations.

D.1. Implementation Details

Camera-Controlled Video Generation. Experiments are conducted on Sekai-Real [4], RealEstate10K [17], and our SpatialVID-HQ dataset. Since RealEstate10K does not include textual annotations, we adopt the text captions provided by CameraCtrl [3]. For fair comparison, we follow the original train/test split for RealEstate10K, randomly sample 10K training clips from Sekai-Real, and use all high-quality clips from SpatialVID-HQ. For SpatialVID-HQ, the training captions are derived from the *Immersive Shot Summary* component of our structured annotations. The DiT-based models are initialized from the publicly released TI2V-Wan2.2-5B checkpoint [9], ensuring consistent architecture and capacity across datasets. During training, the camera encoder, projector, and self-attention modules are learnable, while all remaining components are frozen. LInput frames are resized to 382×480 with a sequence length of 81 frames. We train for 20K steps using a global batch size of 32, the AdamW optimizer with an initial learning rate of 1×10^{-5}, a cosine decay schedule, and a warm-up period of 2K steps. Unless otherwise specified, camera conditioning follows the injection scheme introduced in ReCamMaster [1]. For each frame, the $3{\times}4$ camera extrinsic matrix (12 parameters) is passed through a learnable linear layer cam_encoder $\in \mathbb{R}^{12\times d}$ to project it into the feature dimension d. The resulting embedding is combined with visual tokens through a lightweight per-block *projector* ($\mathbb{R}^{d\times d}$) initialized as an identity mapping to preserve the pretrained feature scale.

CUT3R Fine-tuning. We follow the official fine-tuning protocol of CUT3R [11] and initialize from the publicly released `cut3r_512_dpt_4_64` checkpoint. Training is performed with a global batch size of 64 using the AdamW optimizer with a learning rate of 1×10^{-6}, a weight decay of 0.05, and a total of 6,500 iterations. We fine-tune on long video sequences ranging from 4 to 64 frames, with each frame resized such that the longer side does not exceed 512

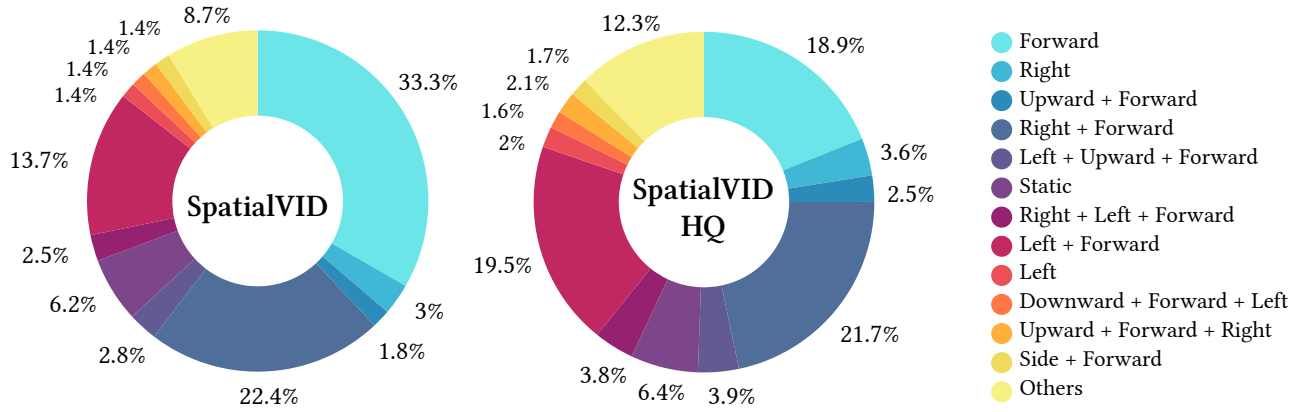
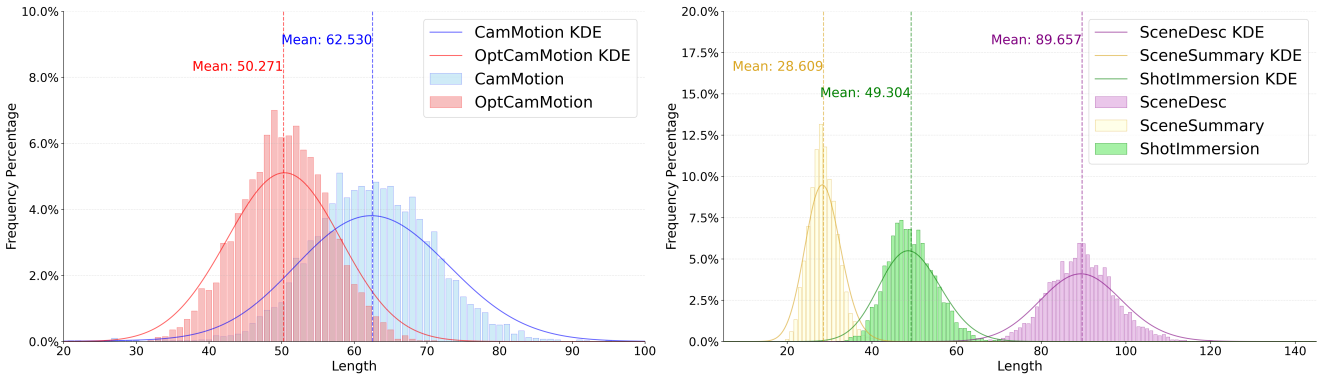


Figure 8. **Distribution of camera motion directions**. The donut charts show the distribution of camera motion directions for the SpatialVID (left) and HQ SpatialVID (right) datasets. The original SpatialVID dataset exhibits a wide range of motion patterns. In contrast, the HQ SpatialVID dataset features a more balanced distribution, addressing the overrepresentation of any single motion direction.



(a) Motion caption length distribution

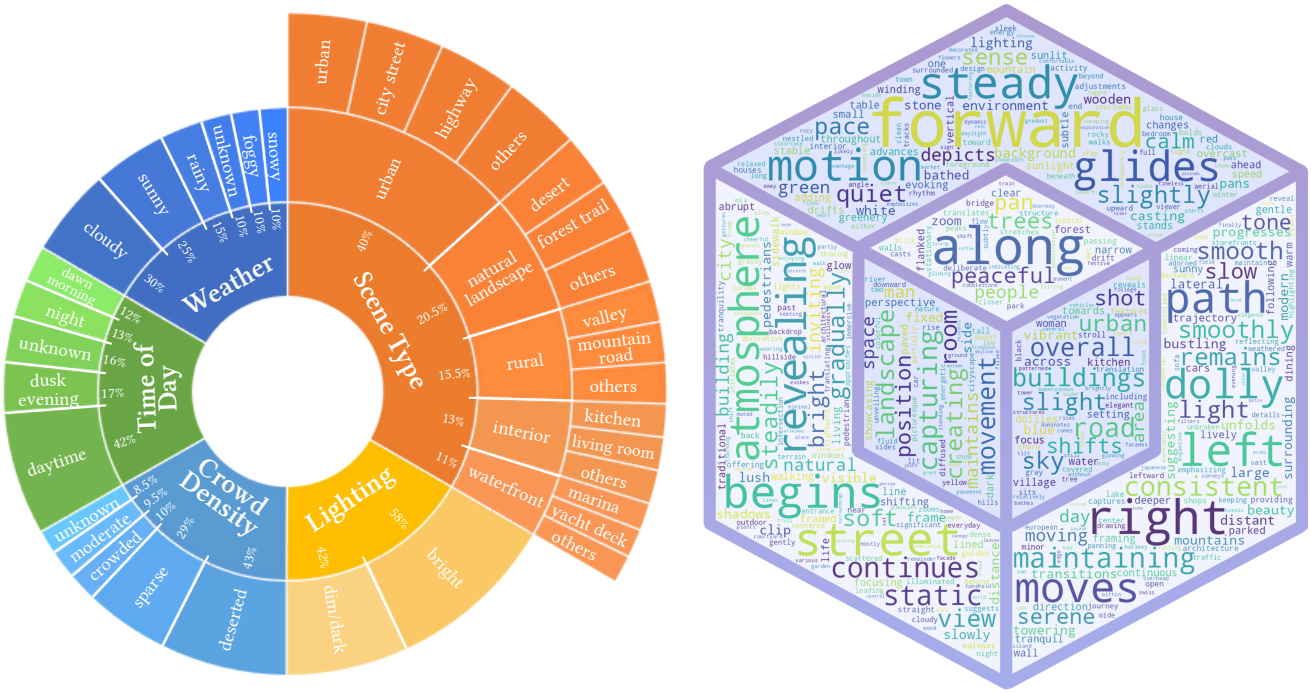
(b) Scene caption length distribution

Figure 9. **Statistical analysis of the caption data**. Fig. (a) and Fig. (b) show the length distributions for motion and scene captions, respectively, comparing the original captions to our enhanced versions. A significant increase in caption length is evident for both types after enhancement.

pixels. During training, the encoder is kept frozen while the decoder and output heads are updated to better align CUT3R's geometric predictions with the spatially consistent captions and camera poses provided by SpatialVID.

VGGT Fine-tuning. We fine-tune VGGT [10] on SpatialVID-HQ together with most of its original training data. For fair comparison, an additional model is fine-tuned using the same original data without SpatialVID-HQ. We initialize from the publicly released VGGT checkpoint and follow the original training strategy, keeping the DINO backbone frozen to ensure consistent architecture and capacity. Training is performed for 40K steps, using the AdamW optimizer with an initial learning rate of 2×10^{-4}. Unless otherwise specified, all remaining hyperparameters follow the default VGGT configuration.

Large Reconstruction Models. All experiments are conducted on RealEstate10K [17] and our SpatialVID-HQ dataset. For fair comparison, both datasets are trained with an equal amount of 60K video clips. In each epoch, random image pairs are sampled, using two views as input and four intermediate views as supervision. All models are trained from scratch to ensure consistent architecture and capacity. Following the GS-LRM [15] training protocol, we adopt a two-stage schedule: the first stage trains at a resolution of 180×320 for 15K steps, followed by a high-resolution stage at 360×640 for an additional 45K steps, totaling 60K steps. Training is performed with a global batch size of 32 using the AdamW optimizer, an initial learning rate of 2×10^{-5}, a cosine decay schedule, and 2K warm-up steps. Unless otherwise specified, the total loss combines pixel-level, perceptual, and depth-smoothness objectives: $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 \mathcal{L}_{\text{lpips}} + \lambda_3 \mathcal{L}_{\text{reg}}$, where $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.25$. The regularization term $\mathcal{L}_{\text{reg}}$ encourages depth smoothness by penalizing abrupt depth discontinuities.



(a) Scene tags distribution

(b) World cloud

Figure 10. **(a) Distribution of scene tags**. The sunburst chart shows the distribution of categorical tags across five primary attributes: weather, time of day, crowd density, lighting, and scene type. The *scene type* attribute is hierarchical, with sub-categories for more detailed classification. The width of each sector reflects the prevalence of the corresponding tag in the dataset. **(b) Word cloud**. A word cloud shaped into the SpatialVID logo, generated from the enhanced captions. The size of each word corresponds to its frequency in the corpus. Key terms such as *motion*, *forward*, *left*, and *right* emphasize the dataset's focus on describing camera movement and spatial dynamics.

D.2. Qualitative Results

We provide additional qualitative comparisons of Camera-controlled video generation and Novel View Synthesis. The camera-controlled video generation results (Fig. 12, Fig. 13, Fig. 14) show that the model trained on *SpatialVID-HQ* precisely follows complex camera trajectories while maintaining realistic spatial continuity and dynamic visual coherence. Moreover, the model demonstrates improved prompt understanding, enabling the generation of more accurate and visually convincing environmental details such as trees and decorations. The novel view synthesis results (Fig. 15) highlight how SpatialVID supports robust geometry learning, maintaining consistent spatial layouts and detailed texture synthesis across diverse motion trajectories.

By integrating explicit 3D motion control with rich textual semantics, SpatialVIDendows physically grounded video generation, dynamic scene synthesis, and spatial intelligence tasks with robust 3D inductive biases. Comprehensive experimental results validate SpatialVID's effectiveness across diverse tasks, laying a solid foundation for research in the field of spatial intelligence.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 6

[2] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 1

[3] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 6

[4] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025. 6

[5] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*,

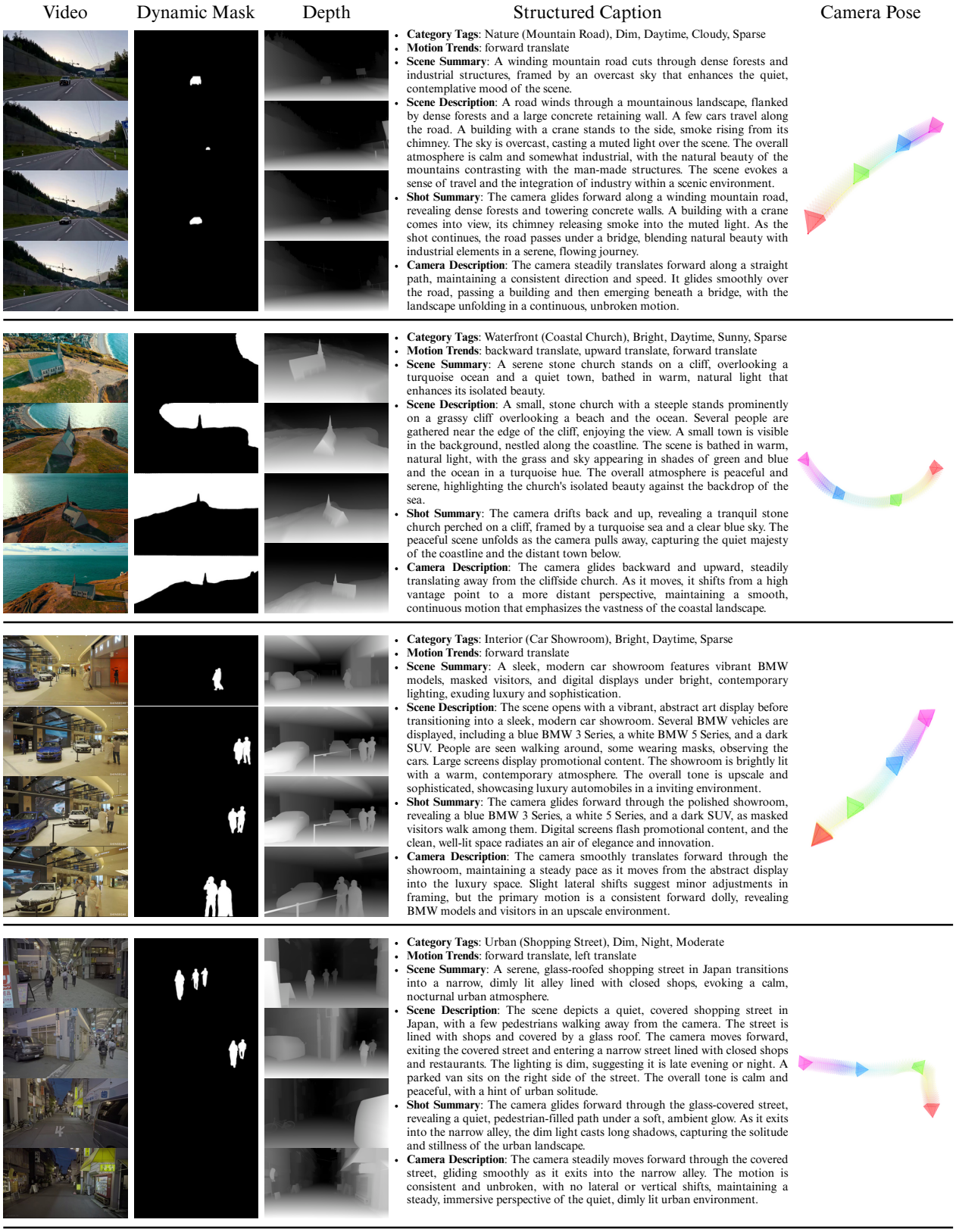


Figure 11. **Sample videos from SpatialVID**. Each example includes synchronized geometry, captions, and spatial annotations. The dataset encompasses diverse environments and camera motions, highlighting its broad coverage and multimodal consistency.

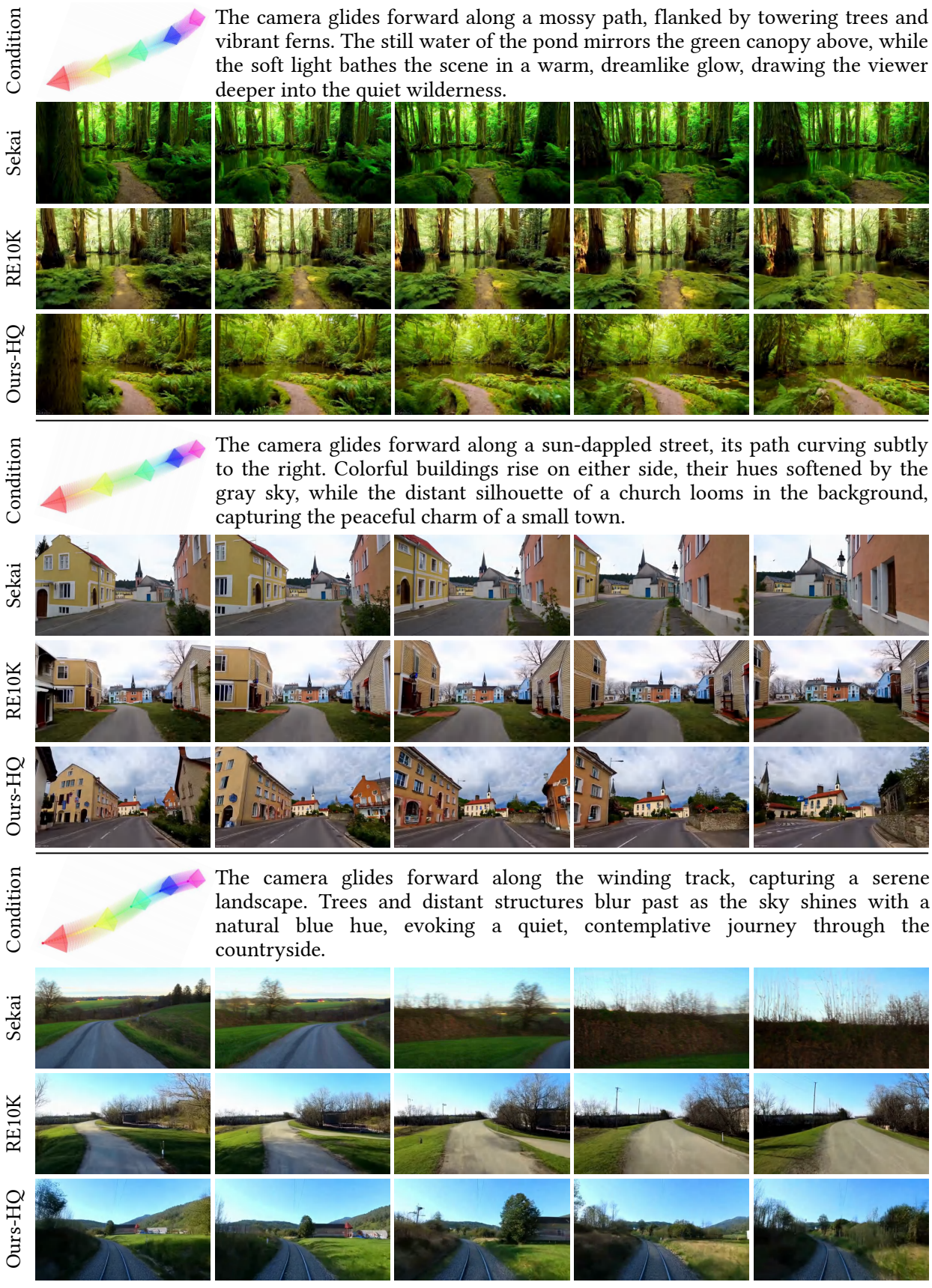


Figure 12. **Different training datasets performance on SpatialVID samples.**

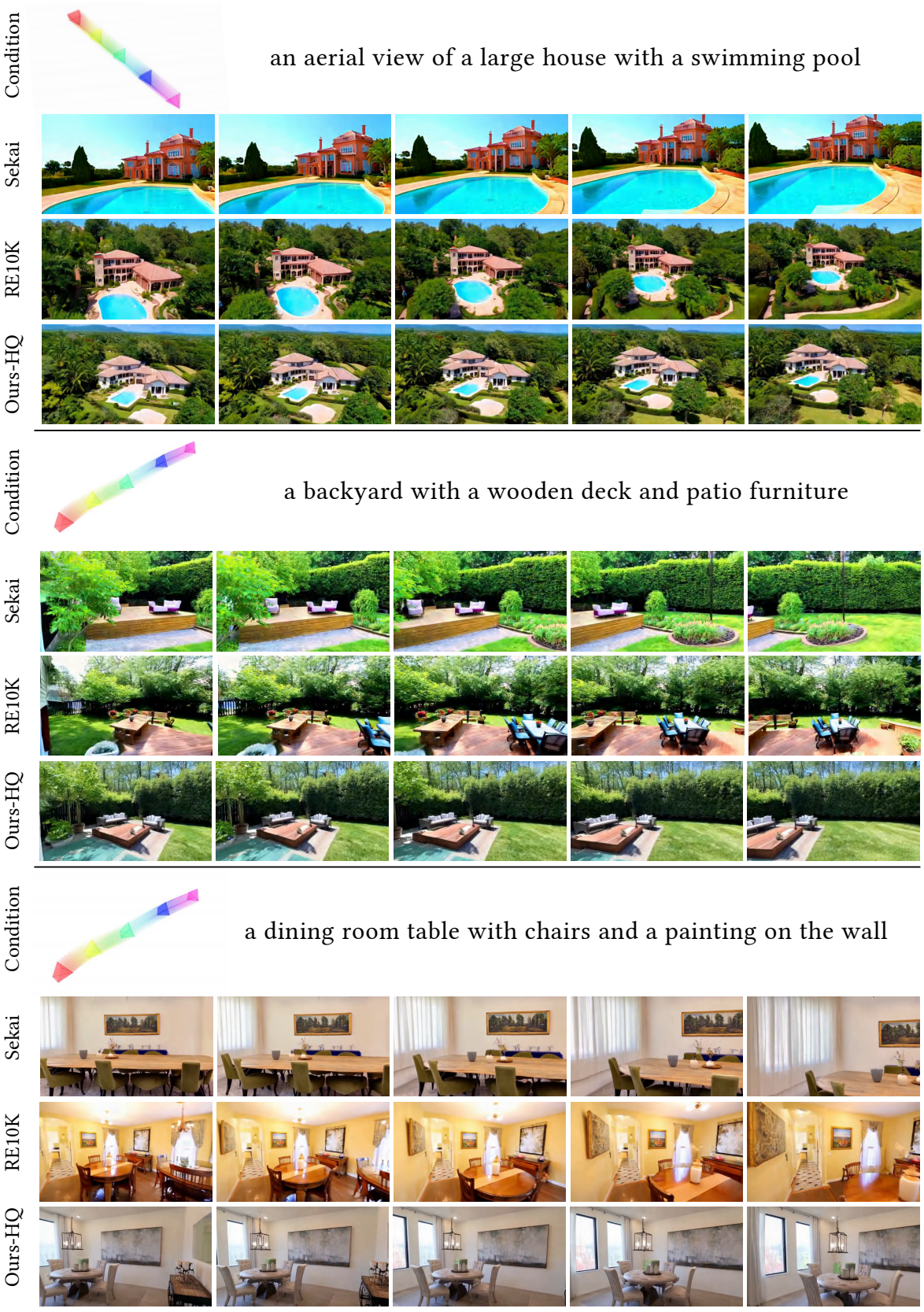


Figure 13. **Different training datasets performance on RealEstate10K samples.**

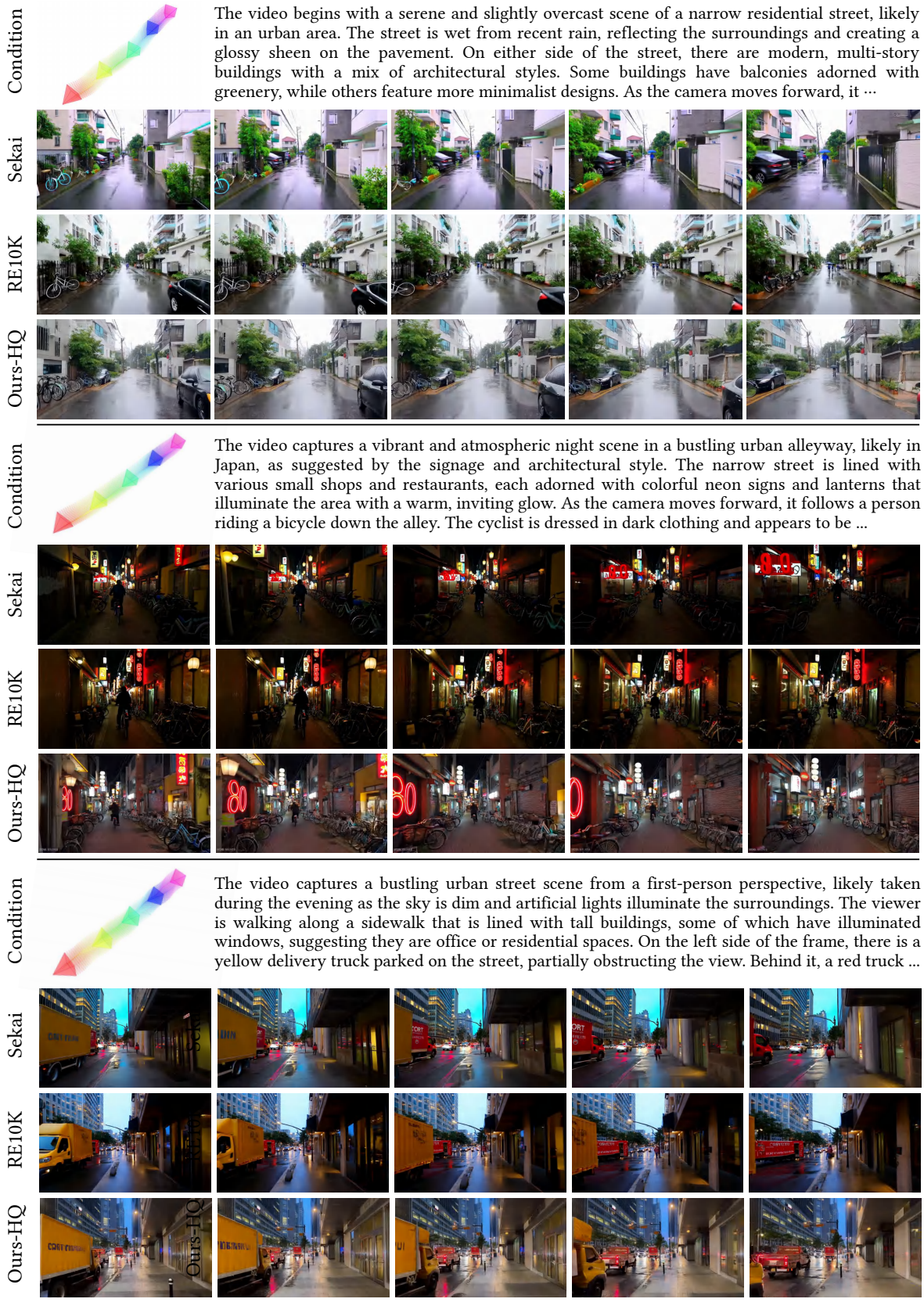


Figure 14. **Different training datasets performance on Sekai-Real samples.**

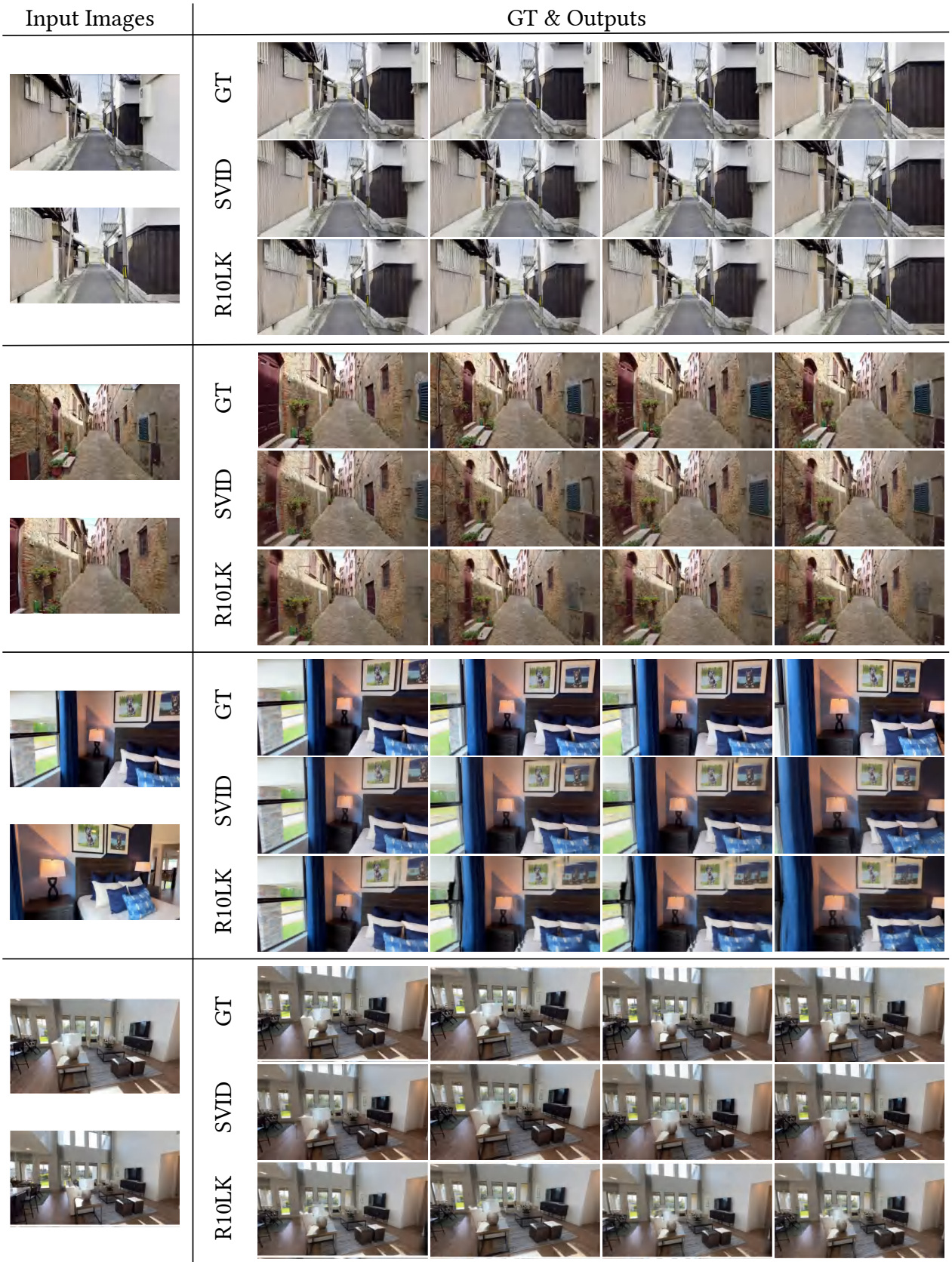


Figure 15. **GS-LRM qualitive comparison on SpatialVID**.

2016. 1

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022. 1

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[8] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34: 16558–16569, 2021. 1

[9] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6

[10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 7

[11] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 6

[12] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[13] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 1

[14] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 1

[15] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 7

[16] Li Zhi, Aaron Anne, Katsavounidis Ioannis, Moorthy Anush, and Manohara Megha. Toward a practical perceptual video quality metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, 2016. 1

[17] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 6, 7