SceneScribe-1M: A Large-Scale Video Dataset with Comprehensive Geometric and Semantic Annotations

Supplementary Material

1. Video Source

Table 1 summarizes the raw video statistics for constructing SceneScribe-1M. The dataset draws primarily from open-source collections, including HD-VILA-100M [10], Panda-70M [4], and Koala-36M [7], which together provide over 200 million video clips totaling approximately 1M hours. In addition to these large-scale open-source datasets, SceneScribe-1M incorporates 668K self-collected clips from Pexels-Video [1], contributing an additional 672 hours of footage at the same resolution. All video clips have a resolution of at least 720p, ensuring high visual quality throughout the dataset. This diverse compilation offers extensive coverage and variety, providing a solid foundation for effective model training and evaluation.

2. LLM Templates

Video Content Assessment. To facilitate automated video content assessment, we utilize the powerful multimodal large language model Qwen2.5-VL-72B [3] as an evaluator. For systematic evaluation, we meticulously design question templates targeting six key dimensions of video quality: camera motion intensity, object motion intensity, continuity, watermark presence, camera distortion, and problematic lighting. Table 2 summarizes the assessment criteria, including the specific candidate labels employed for each dimension. These question templates (as shown in Figure 1, 2, 3, 4, 5 and 6) allow for reliable and consistent filtering of unsuitable video content. Videos are excluded from the curated dataset if they exceed thresholds in any dimension, such as exhibiting strong artifact, continuity issues, or visible watermarks.

Semantic Annotation. In Figure 7, we present the annotation template employed for generating semantic video descriptions with Qwen2.5-VL-72B. The template clearly outlines the task instructions, annotation guidelines, and objectives, ensuring high-quality and consistent annotations. Specifically, it guides the model to systematically summarize each video by detailing the scene setting, identifying primary subjects or characters, and describing significant actions. This standardized approach helps produce comprehensive and structured scene descriptions, supporting accurate semantic labeling throughout the dataset.

3. Downstream Tasks and Benchmark Settings

Monocular Depth Estimation. We employ MoGe [8] for monocular depth estimation. The model adopts a ViT encoder pre-trained on DINOv2 features and a lightweight convolutional decoder, taking single images as input and predicting affine-invariant point maps. During training, the model utilizes a combination of robust global and multi-scale local geometry losses, together with surface normal supervision and explicit masking of infinity regions. To enhance generalizability, extensive data augmentation is applied, including color jitter, Gaussian blur, JPEG compression, and random cropping with centered principal points. The learning rate is scheduled to decay every 100K iterations, and training proceeds with a batch size of 256. For evaluation, following the protocol introduced in MoGe [8], predicted depth maps are aligned to ground truth by optimizing for scale (and shift if applicable), and quantitative accuracy is measured using the absolute relative error and

Table 1. **Raw Video Statistics** for constructing SceneScribe-1M.

Type	Dataset	#Video Clips	Toal Length (h)	Resolution
Open-Source	HD-VILA-100M [10]	103M	760.3K	720p
	Panda-70M [4]	70M	167K	720p
	Koala-36M [7]	36M	172K	720p
Self-Collected	Pexels-Video [1]	668K	672	720p

Table 2. **Video Content Assessment** with Qwen2.5-VL-72B [3].

Question Type	Candidate Labels
Camera Motion Intensity	No, Slight, Strong, N/A
Object Motion Intensity	No, Slight, Strong, N/A
Continuity Judgment	Continuous, Non-Continuous
Watermarks Judgment	Yes, No
Camera Distortion Judgment	No, Slight, Strong, N/A
Problematic Light Judgment	No, Slight, Strong

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Carefully follow prompt and provide clear and detailed answers. Analyze the video carefully. Determine if the camera is moving during the video and rate the intensity of the camera motion based on the definitions below: (a) Does the video exhibit camera motion? If yes, specify the intensity; and (b) Possible Labels: No, Slight, Strong, N/A.
3. Label definitions
No: The camera is completely static or nearly immobile;
Slight: Minor camera movements (e.g., subtle handheld shaking);
Strong: Significant noticeable movements (e.g., panning, tilting, sweeping);
N/A: Unable to determine or irrelevant to the video content.
3. Label definitions
Simply give an answer with label definition without additional explanation.

Figure 1. **MLLM Instruction Template** for Camera Motion.

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Carefully follow prompt and provide clear and detailed answers. Analyze the video carefully. Examine the scene carefully to identify if any objects or subjects (e.g., people, animals, vehicles) are moving independently from the camera. Rate the intensity and clearly list the main moving subjects:
(a) Are there moving objects or subjects in the video? If yes, specify the intensity and identify the primary moving subjects; and
(b) Possible Labels: No, Slight, Strong, N/A.
3. Label definitions
No: No detectable independent movement of subjects; the scene is static;
Slight: Minor movements (e.g., slow walking, gentle sway of trees);
Strong: Prominent movements (e.g., running people, fast vehicles);
N/A: Unable to determine or irrelevant.
4. Additional Instruction
Clearly list the primary moving subjects (e.g., people, vehicles). Simply give an answer based on the label definition without additional explanation.

Figure 2. **MLLM Instruction Template** for Object Motion.

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation. *2. Annotation Guidelines*
Carefully follow prompt and provide clear and detailed answers. Watch the video to determine if it is presented as a single continuous shot without interruptions, or if it contains cuts, scene transitions, or jumps:
(a) Is the video a single continuous shot, or does it include cuts or scene transitions? and
(b) Possible Labels: Continuous, Non-Continuous.
3. Label definitions
Continuous: No visible cuts; the footage is a seamless, uninterrupted shot;
Non-Continuous: Presence of clear cuts or transitions between scenes;
4. Additional Instruction
Simply give an answer with label definition without additional explanation.

Figure 3. **MLLM Instruction Template** for Continuity.

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Carefully follow prompt and provide clear and detailed answers. Inspect the video carefully to determine if there are any watermarks or logos present on the footage:
(a) Does the video contain a watermark? and
(b) Possible Labels: Yes, No.
3. Label definitions
Yes: Clearly visible watermark or logo is present;
No: No watermark or logo detected.
4. Additional Instruction
Simply give an answer with label definition without additional explanation.

Figure 4. **MLLM Instruction Template** for Watermarks.

the percentage of inliers within a tolerance threshold. Additional care is taken to ensure fair comparisons by masking undefined or noisy regions in the evaluation datasets as outlined in the official protocol.

3D Reconstruction. We utilize VGGT [6] for scene reconstruction and camera parameter estimation using sparse or multi-view images. VGGT is a feed-forward trans-

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Examine the video carefully to identify any camera distortion (e.g., fisheye distortion, wide-angle lens distortion, perspective distortion). Indicate the presence and severity:
(a) Does video exhibit camera distortion? If yes, specify the intensity; and
(b) Possible Labels: No, Slight, Strong, N/A.
3. Label definitions
No: No visible distortion;
Slight: Mild distortion noticeable but not significantly affecting the content;
Strong: Significant noticeable distortion (e.g., pronounced fisheye effect);
N/A: Unable to determine or irrelevant.
4. Additional Instruction
Simply give an answer with label definition without additional explanation.

Figure 5. **MLLM Instruction Template** for Camera Distortion.

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Evaluate the video for the presence of transparent objects (such as glass), noticeable reflections (mirrors, metallic surfaces), or significantly overexposed (bright, washed-out) areas. Indicate the intensity and briefly describe any prominent examples:
(a) Does the video contain transparent objects, strong reflections, or significant overexposed areas? If present, specify intensity and provide brief examples; and
(b) Possible Labels: No, Slight, Strong.
3. Label definitions
No: No identifiable transparent surfaces, reflections, or overexposed areas;
Slight: Minor reflections, transparency, or small overexposed regions not dominating the frame;
Strong: Prominent and clearly visible reflections, extensive transparent surfaces, or large areas significantly overexposed.
4. Additional Instruction
Provide concise descriptive examples (e.g., large reflective glass window, highly overexposed sky, metallic reflective surface). Simply give an answer with label definition without additional explanation.

Figure 6. **MLLM Instruction Template** for Problematic Light.

1. Task Instruction
You are an expert in computer vision, video analysis, and 3D geometry. Your detailed and precise evaluation is crucial for accurate annotation.
2. Annotation Guidelines
Carefully follow prompt and provide clear and detailed answers. Provide a concise yet thorough description of the video's contents. Be specific and factual. Clearly include the following aspects:
(a) scene setting;
(b) primary subjects or characters; and
(c) significant actions occurring.
3. Annotation Objective
Generate a textual summary of the video, detailing the main environment, primary subjects, and notable actions. The more detailed the better.

Figure 7. **MLLM Instruction Template** for Video Captions.

former network trained on a large, diverse collection of 3D-annotated datasets, and is designed to jointly predict camera intrinsics/extrinsics, depth maps, dense point clouds, and tracking features for any number of input views in a

single forward pass. During inference, the model achieves state-of-the-art results and remarkable efficiency, generally reconstructing scenes in less than a second without requiring visual geometry-based post-processing. For training, VGGT applies multi-task supervision across camera, depth, point map, and tracking branches. The loss terms include Huber loss for camera poses, aleatoric uncertainty-weighted depth and point losses, and correspondence-based tracking loss, with outputs normalized to a canonical scale. Evaluation protocols follow standard benchmarks such as CO3Dv2 and ETH3D, assessing accuracy and completeness of point cloud reconstructions, as well as camera parameter estimation using relative rotation and translation metrics (e.g., AUC_{30}). For each test scene, predicted cameras and point clouds are aligned with ground-truth via the Umeyama algorithm, and comparisons are made to previous optimization-based and deep learning methods.

4D Reconstruction. For dynamic scene reconstruction from monocular video, we leverage MonST3R [11], a geometry-first approach that extends static pointmap representations to time-varying dynamic scenes. Given a video sequence, MonST3R predicts per-frame pointmaps and associated camera poses, using a ViT-based backbone and decoder fine-tuned on a curated mixture of synthetic and real datasets featuring dynamic motion. During global optimization, pairwise pointmap estimates are aggregated over a temporal sliding window, and camera trajectory smoothness and flow consistency losses are incorporated to enforce temporal coherence and robust alignment. This yields a dynamic global point cloud parameterized by per-frame camera poses and depth estimates, which supports temporally stable 4D reconstruction, joint video depth, and pose estimation. We apply Sintel for evaluation, using absolute-relative error and correct pixels percentage for depth, and standard pose metrics for camera trajectory.

2D Point Tracking. We adopt CoTracker3 [5] for robust point tracking across a broad range of videos. CoTracker3 is a transformer-based architecture designed to iteratively predict point tracks along with per-point confidence and visibility, enabling accurate tracking of both visible and occluded points throughout a sequence. CoTracker3 simplifies and accelerates the point tracking pipeline by utilizing efficient 4D correlation features, cross-track attention, and a unified transformer for joint updates—leading to improved tracking accuracy and efficiency over prior approaches. The model is first pre-trained on synthetic Kubric data, then effectively fine-tuned on SceneScribe-1M. Training is supervised using a discounted Huber loss for visible/occluded tracks and binary cross-entropy for confidence/visibility estimation. For evaluation, we follow the TAP-Vid benchmark protocol, measuring Average Jaccard, visibility-averaged tracking precision, and occlusion prediction accuracy on TAP-Vid.

3D Point Tracking. We employ SpatialTrackerV2 [9] for 3D point tracking in monocular videos. SpatialTrackerV2 unifies depth prediction, camera pose estimation, and object motion into a single, fully differentiable, and end-to-end pipeline. The original SpatialTrackerV2 is trained on a heterogeneous collection of 17 datasets, which cover posed RGB-D videos, synthetic sequences, and diverse real-world scenarios, leveraging different forms of self-supervision and consistency constraints to mitigate domain biases and error accumulation. During inference, query points are tracked through arbitrary motion—including occlusions and rapid dynamics—with spatial-temporal consistency and visibility/dynamic probability estimates. For evaluation, we follow the TAPVId-3D benchmark protocol, reporting metrics such as Average Jaccard, Average 3D Position Accuracy, and Occlusion Accuracy across representative scenarios. Depth and camera predictions are aligned via scale/shift as needed for fair comparison.

Text/Pose-to-Image. We employ AC3D [2] for precise camera-controllable video generation. AC3D builds upon a large-scale Video Diffusion Transformer pretrained for text-to-video synthesis, and introduces a principled method to inject camera trajectory control into video diffusion models without sacrificing generation quality. For training and evaluation, AC3D fine-tunes lightweight camera conditioning modules (ControlNet-style) atop a frozen video DiT backbone, using our SceneScribe-1M. The model is supervised with both text prompts and camera trajectories (as Plucker coordinates), optimizing for visual fidelity and accurate camera following. During inference, AC3D enables conditioning on arbitrary camera paths alongside text prompts, granting users fine-grained control over both semantics and viewpoint in the generated videos. Performance is evaluated on RealEstate10K test set using several metrics (FID, FVD, CLIP score for visual quality, translation and rotation error with respect to camera following).

References

[1] Openvideo. https://github.com/UmiMarch/OpenVideo, 2023. 1

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22875–22889, 2025. 3

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang,

et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13320–13331, 2024. 1

[5] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6013–6022, 2025. 3

[6] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 2

[7] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8428–8437, 2025. 1

[8] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5271, 2025. 1

[9] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025. 3

[10] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, 2022. 1

[11] 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. 3