TL;DR SpatialVID curates more than 21,000 hours of raw video into ~2.7M clips (~7,089 hours) and annotates them with per‑frame camera poses, depth maps, dynamic masks, structured captions, and motion instructions — a resource aimed at scaling spatial intelligence research and data‑hungry video/3D models.

What is SpatialVID?

SpatialVID is a large‑scale dataset designed to address the shortage of diverse, high‑quality spatial annotations for real‑world dynamic videos. It focuses on rich geometric and semantic labels that are useful for 3D perception, reconstruction, and video generation systems.

Key components:

- Per‑frame camera poses (ground‑truth camera motion for dynamic scenes)
- Dense depth maps across time
- Dynamic object masks
- Structured captions
- Serialized motion instructions

References:

- Paper (arXiv): https://arxiv.org/abs/2509.09676
- GitHub: https://github.com/NJU-3DV/spatialVID
- Project page: https://nju-3dv.github.io/projects/SpatialVID/

Scale and curation pipeline

According to the paper and project materials, the pipeline:

- Collects more than 21,000 hours of raw, in‑the‑wild video.
- Filters and segments the corpus into approximately 2.7 million clips, totalling ~7,089 hours of dynamic content.
- Runs an annotation pipeline that adds camera poses, depth maps, dynamic masks, captions, and motion instructions.

The intent is to improve model generalization by increasing scene diversity, motion patterns, and annotation richness.

Note: Always consult the upstream repo for the latest statistics and updates, as numbers and splits may evolve.

Downloading the dataset

The authors provide downloads via multiple mirrors and helper utilities:

- Hugging Face organization: https://huggingface.co/SpatialVID
- ModelScope org: https://www.modelscope.cn/organization/SpatialVID
- Repo utilities: utils/download_SpatialVID.py (dataset) and utils/download_YouTube.py (raw videos)

Example (see repository for exact arguments and latest instructions):

```
git clone --recursive https://github.com/NJU-3DV/SpatialVID.git
cd SpatialVID
conda create -n SpatialVID python=3.10.13
conda activate SpatialVID
pip install -r requirements/requirements.txt

# Optional: scoring dependencies
pip install paddlepaddle-gpu==3.0.0 -i https://www.paddlepaddle.org.cn/packages/stable/cu126/
pip install -r requirements/requirements_scoring.txt

# Use helper script to download datasets (adjust prefixes/paths)
python utils/download_SpatialVID.py --help
```

Depth components can be large; expect significant storage and bandwidth requirements.

Annotation, scoring, and visualization

The repository includes scripts for scoring, annotation, captioning, and visualization:

```
# 1) Scoring (configure ROOT_VIDEO, OUTPUT_DIR in scripts/scoring.sh)
bash scripts/scoring.sh

# 2) Annotation (configure CSV, OUTPUT_DIR in scripts/annotation.sh)
```

```
bash scripts/annotation.sh

# 3) Captioning (configure CSV, SRC_DIR, OUTPUT_DIR, and API keys)
bash scripts/caption.sh
```

Visualization helpers:

- Pose visualization: viser/visualize_pose.py (uses poses.npy)
- Final annotation visualization: viser/visualize_megasam.py (uses sgd_cvd_hr.npz)

These tools help sanity check camera trajectories, depth consistency, and object dynamics before downstream training or evaluation.

Practical considerations

- Storage/IO: Plan for large download sizes and IOPS if using networked storage.
- GPU memory: Some depth and segmentation models are memory intensive; follow repo tips for environment setup.
- Licensing: Source code is Apache 2.0. Dataset and third party components may carry different or more restrictive terms — review each subset's license before commercial use.
- Reproducibility: Versions of annotation/scoring models can affect outputs; pin requirements and document your environment.

Why it matters

Scaling video datasets with spatial annotations is critical for:

- 3D reconstruction and SLAM/VO research
- Video diffusion/generation models with camera aware controls
- AR occlusion, object tracking, and scene understanding
- Policy learning for embodied agents that need spatial reasoning

SpatialVID contributes broader scene diversity and richer labels to help close the gap between lab curated benchmarks and real world dynamic video.

References

- arXiv: https://arxiv.org/abs/2509.09676 (submitted 2025 09 11)
- GitHub: https://github.com/NJU-3DV/spatialVID
- Project page: https://nju-3dv.github.io/projects/SpatialVID/
- Announcement thread (external): https://www.linkedin.com/posts/naveen-manwani-65491678_paper-alert-paper-title-spatialvid-activity-7372688618517114880-5ooQ

  Notes: Figures and numeric statistics are based on the public paper/README at publish time. Validate fit for your use case and check the repo for updates.