V23D: LEVERAGING VIDEO DIFFUSION PRIORS FOR HIGH-FIDELITY MULTI-VIEW 3D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image 3D generation must infer self-occluded regions from incomplete evidence. Orbit videos offer complementary views, but a shared keyframe subset can omit cues essential to individual regions. We present **V23D**, a two-phase video-prior-to-3D framework that combines geometry-aware orbit generation with region-specific conditioning. In Phase I, we adapt Wan2.2 with noise-decoupled LoRA and a **Geometry-aware Orbit Adapter (GOA)** to generate 49-frame turntable priors, reducing FVD from **205.8 to 176.8**. In Phase II, we introduce a **Regional Source Adapter (RSA)** that connects the complete orbit to coarse-to-fine 3D generation. The sparse-structure stage records regional source-frame indices and their original attention masses; the shape stage uses independent current queries to re-read the referenced patches, while receiving sparse support separately. This design preserves source access across generation stages rather than passing only fixed fused features. On HY3D-Bench-test, V23D achieves **0.2418 ULIP-I** and **0.3827 Uni3D-I**, outperforming the evaluated baselines. Compared with the four-view OFA configuration, RSA improves F-score@1% from **57.4% to 61.1%**, and achieves **62.7%** on Toys4K in cross-dataset evaluation. Matched full-orbit comparisons further favor RSA over all-frame aggregation, instantaneous regional routing, and fixed-feature transfer, supporting the value of cross-stage source preservation and state-dependent patch re-reading.

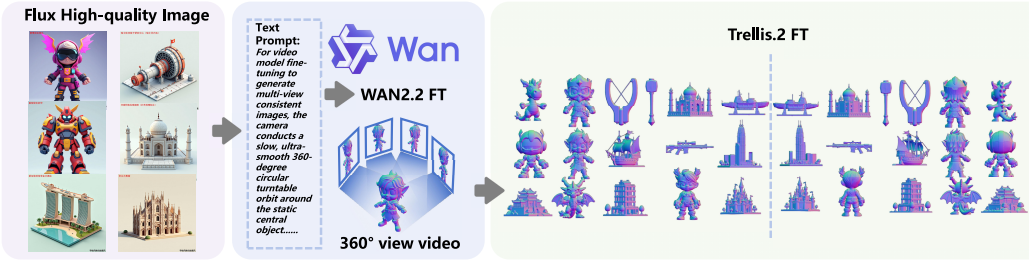

Figure 1: **V23D video-to-3D generation.** An input image is expanded into a turntable video for 3D generation with TRELLIS.2.

1 INTRODUCTION

Single-image 3D generation supports content creation and simulation, but one view leaves self-occluded geometry underdetermined. Diffusion-guided optimization (Poole et al., 2022; Wang et al., 2023; Chen et al., 2023), feed-forward reconstruction (Hong et al., 2024; Tang et al., 2024; Xu et al., 2024; Tochilkin et al., 2024), and structured latent generators (Xiang et al., 2025b;a; Zhang et al., 2024; Ye et al., 2025) supply shape priors. Multi-view diffusion (Shi et al., 2023; Liu et al., 2024; Long et al., 2023; Voleti et al., 2024) and video models (Blattmann et al., 2023; Wan et al., 2025; Chen et al., 2024; Huang et al., 2025) offer complementary generated views. Using those views requires a conditioning interface that connects image evidence to the regions being generated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

An orbit provides uneven evidence across regions: a handle obscured in selected keyframes may appear clearly between them. Its generated views can also disagree. Regional routing and coarse-to-fine view reuse address related selection problems (Qu et al., 2026; Sun et al., 2026; Wang et al., 2020; Lin et al., 2026). Our focus is the information transferred from sparse-structure (SS) generation to Shape refinement. SS selects where to read; Shape decides what to extract. Retaining source addresses and masses lets current fine-stage queries retrieve content from the original patches.

V23D combines an adapted video prior with this source-based interface. Phase I uses noise-decoupled LoRA and the **Geometry-aware Orbit Adapter (GOA)** for 49-frame turntable generation. In Phase II, the **Regional Source Adapter (RSA)** makes all frames available in a shared patch bank. SS sampling records regional source indices and retained attention masses; Shape receives these records alongside sparse support and re-reads the referenced patches with independent queries. We hypothesize that late SS states provide useful structural context for selecting sources, while Shape states benefit from extracting their details anew. RSA operates within the TRELLIS.2 hierarchy and flow-matching objective (Lipman et al., 2022; Song et al., 2020).

Our contributions are:

- **Geometry-aware video prior.** GOA complements noise-decoupled LoRA for orbit generation, providing video conditions for coarse-to-fine 3D synthesis.
- **Regional Source Adapter (RSA).** We propose a cross-stage interface that records regional source indices and original attention masses from the complete orbit, then enables independent Shape queries to re-read the source patches.
- **Leading benchmark performance and empirical validation.** V23D achieves the best results among the evaluated baselines, leading image–shape alignment on HY3D-Bench and surface accuracy on both HY3D-Bench and Toys4K. Matched full-orbit ablations further demonstrate the benefits of cross-stage source transfer and state-dependent patch re-reading.

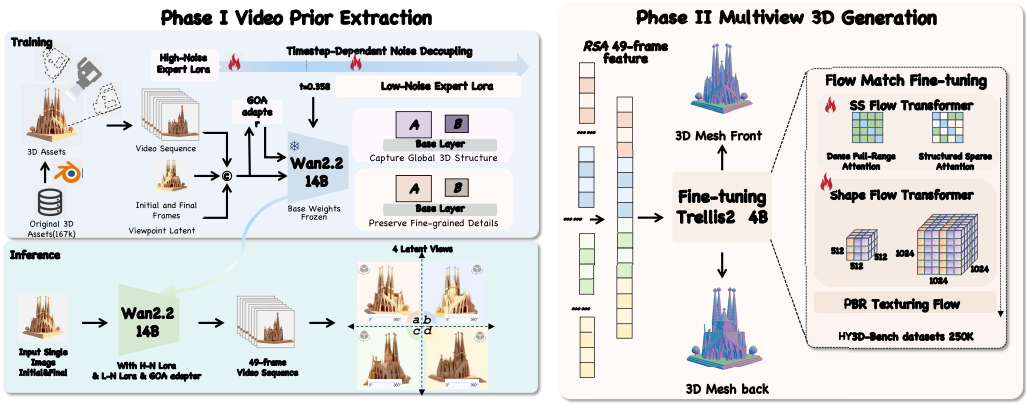

Figure 2: **V23D backbone and RSA insertion site.** Phase I generates a 49-frame orbit with Wan2.2, LoRA, and GOA. The four-view tile depicts the OFA reference; RSA instead reads the full-orbit feature bank to condition TRELLIS.2 SS and Shape generation (Fig. 3).

2 Related Work

Image-conditioned 3D generation combines shape priors with visual evidence. Score Distillation Sampling (SDS) (Poole et al., 2022; Lin et al., 2023; Chen et al., 2023), feed-forward reconstruction (Hong et al., 2024; Tang et al., 2024; Xu et al., 2024), and structured latent generators (Xiang et al., 2025b;a) provide different generative foundations. Multi-view diffusion (Shi et al., 2023; Long et al., 2023; Voleti et al., 2024) and video priors (Chen et al., 2024; Wan et al., 2025; Huang et al., 2025) expand the available generated views. V23D studies their regional use across coarse-to-fine generation.

Regional routing and source reuse take several forms. UMI3D (Qu et al., 2026) routes voxel groups to image features using attention-derived scores. ROAR-3D (Sun et al., 2026) selects views per token and block, including in its geometric refinement stage. Stream3D (Zhou et al., 2026) stores query-level frame indices across stream chunks and aggregates their evidence into a bounded conditioning bundle. PatchmatchNet (Wang et al., 2020) reuses coarse pixel-wise view weights for finer-scale matching of geometrically warped features. These precedents motivate examining both the transferred representation and the subsequent reading operation.

Stage-specific conditions also connect reconstruction and generation. ReconViaGen (Chang et al., 2025) aggregates VGGT features into global SS and local per-view conditions, which fine-stage queries read, with rendering-aware velocity compensation for alignment. Mix3R (Lin et al., 2026) jointly predicts sparse structure, point maps, and poses, then biases fine-stage DINO-token attention using voxel–patch overlap. RSA transfers region-addressed source indices and original retained masses recorded during SS sampling. Shape queries independently re-read original patches from those fixed sources, with sparse support supplied separately. The comparisons vary when sources are selected, what crosses stages, and whether Shape re-reads the original patches.

3 Method

3.1 Overview

V23D maps an RGB image I to a 3D asset through a 49-frame orbit $\mathcal{V}$. Phase I adapts Wan2.2 with noise-decoupled LoRA and GOA. Phase II uses RSA to condition TRELLIS.2: SS sampling records regional source references, and current Shape queries re-read their original patches. RSA modifies the conditioning path while preserving the latent hierarchy, sparse-support decoder, and asset-generation stack (Figs. 2–3).

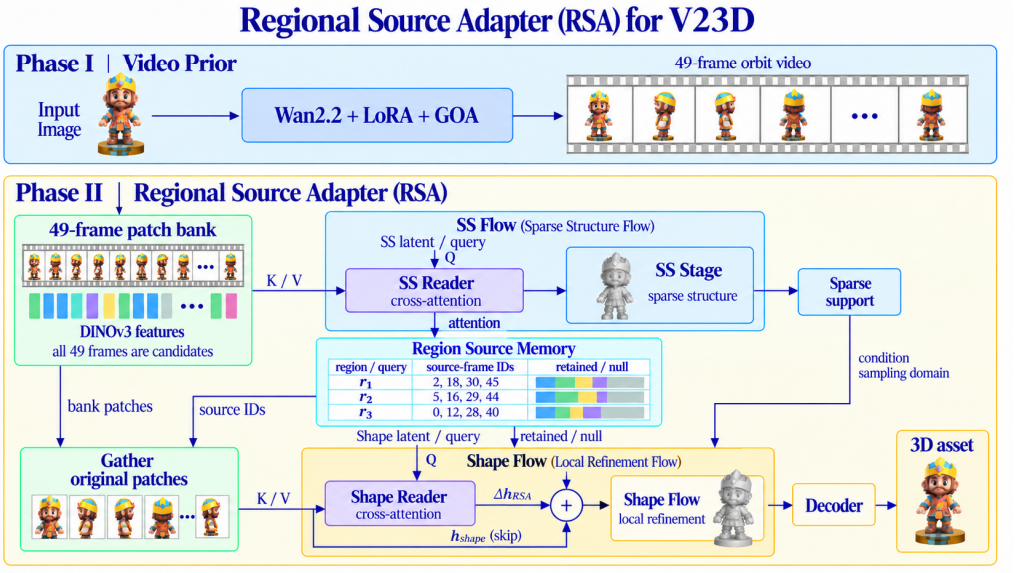

Figure 3: **Regional Source Adapter: Read, Record, and Re-read.** The SS and Shape Readers condition their respective generation stages. SS attention records source IDs, retained masses, and a null remainder. Sparse support defines the Shape sampling domain; independent Shape queries re-read original patches gathered from the recorded sources. Object thumbnails, IDs, and weights are schematic.

3.2 Geometry-aware Video Prior Extraction

Video adaptation. The 14B image-conditioned Wan2.2 prior (Wan et al., 2025) is adapted to rendered turntable sequences. For a target video latent Z_0 and image, endpoint-frame, and orbit

conditions y, conditional flow matching minimizes

$$\mathcal{L}_{\text{video}} = \mathbb{E}_{t,Z_0,\epsilon}\left[w(t)\,\|v_\phi(x_t,t,y) - (\epsilon - Z_0)\|_2^2\right]. \tag{1}$$

High-noise and low-noise LoRA branches specialize in structure and appearance, with an expert-switching boundary $t_b = 0.358$ under the video sampler's time convention.

Geometry-aware Orbit Adapter. GOA conditions high-noise DiT features on image, orbit, and endpoint-frame tokens through a residual,

$$h_l' = h_l + \Delta h_l^{\text{LoRA}} + \gamma_l \Delta h_l^{\text{geo}}. \tag{2}$$

The video objective includes orbit-closure, silhouette, and camera auxiliary terms. Their definitions and the 512^2 training/1024^2 inference schedule appear in Appendix B.1.

3.3 Regional Source Adapter (RSA)

Full-orbit patch bank. A frozen DINOv3 encoder (Siméoni et al., 2025) encodes all $F = 49$ frames into $F_{f,p} = E(I_f)_p$. The bank $\mathcal{B} = \{(f, p, F_{f,p}, u_{f,p}, c_f)\}$ stores original features, frame/patch IDs, processed patch coordinates, and orbit metadata. K/V consume feature vectors; the remaining fields provide indexing and coordinate provenance. Both attention stages use zero spatial bias. SS pooling, when present, maintains correspondence with the original patches used by Shape.

SS reading. Within an SS block, cross-attention (Vaswani et al., 2017) uses the current state at coarse 3D address r. For head a, $q_r^{SS,a} = W_Q^{SS,a}\,\text{LN}(h_r^{SS})$, while K/V are stage-specific projections of bank features:

$$A_r^{SS,a} = \text{softmax}_{(f,p)\in\mathcal{P}^{SS},\perp}\left[\left\{\frac{q_r^{SS,a} K_{f,p}^{SS,a\top}}{\sqrt{d}} + b_{r,f,p}\right\}, \ell_{r\perp}^a\right]. \tag{3}$$

Here d is the per-head key width, $\mathcal{P}^{SS}$ indexes valid frame–patch pairs, $b = 0$, and $\ell_{r\perp}^a$ is a null logit with zero value. Padding is masked. Each head normalizes over all valid pairs and null before heads are averaged into $\bar{A}^{SS}$. Projected weighted values form the block's video residual.

Regional source records. The designated Reader's last positive-condition evaluation on the SS sampling trajectory writes a full-grid source table. Positive-call records are isolated from CFG negative calls. At each coarse address,

$$\begin{aligned} m_{r,f} &= \sum_{p\in P_f^{SS}} \bar{A}_{r,f,p}^{SS}, \qquad J_r = \text{TopD}_f(m_{r,f}), \\ m_{r\perp} &= 1 - \sum_{f\in J_r} m_{r,f}, \quad \mathcal{M}_r = (J_r, \{m_{r,f}\}_{f\in J_r}, m_{r\perp}). \end{aligned} \tag{4}$$

The initial design budget $D = 4$ applies independently to each region. Valid sources retain their original masses; padded entries contribute zero. The table includes unoccupied coarse addresses. Its null remainder combines original null attention with discarded frame mass. Consequently, $s_r = \sum_{f\in J_r} m_{r,f}$ reflects attention concentration and truncation, rather than calibrated visibility.

Independent Shape re-reading. The SS decoder supplies sparse support $\mathcal{S}$, independently of $\mathcal{M}$. A Shape coordinate maps to $r(q) = \text{Addr}_{SS}(x_q;\Gamma)$, where Γ specifies canonical bounds, axes, grids, and boundary rules. The current Shape state provides a new query; $J_{r(q)}$ gathers original patches from its retained frames:

$$\begin{aligned} q_q^H &= W_Q^H\,\text{LN}(h_q^H), \\ \beta_{q,f,p} &= \text{softmax}_{p\in P_f^H}\left(\frac{q_q^H K_{f,p}^{H\top}}{\sqrt{d}} + b_{q,f,p}^H\right), \quad f \in J_{r(q)}. \end{aligned} \tag{5}$$

In single-head notation, $b^H = 0$ and the residual is

$$\Delta h_q^H = W_O^H \left[\sum_{f \in J_{r(q)}} m_{r(q),f} \sum_{p \in P_f^H} \beta_{q,f,p} V_{f,p}^H \right]. \tag{6}$$

During Shape sampling, J/m stay fixed while current queries recompute within-frame patch attention. The retained mass $s_{r(q)}$ scales the video contribution. Empty reads return zero, and $W_O^H(0) = 0$ ensures a zero additional residual at zero retained mass. Algorithm 1 summarizes the sampling sequence; Table 6 summarizes the adopted interface design.

Cross-stage information. RSA separates a stable choice of source views from state-dependent feature extraction. Coarse SS states select regional frame references, while each evolving Shape state determines which patches to read within those frames. Passing indices and masses preserves access to the original feature bank across stages. Sparse support defines where Shape tokens exist; source memory defines which video cues condition them. This division aligns source transfer with the coarse-to-fine generation hierarchy.

3.4 Conditioning and Optimization

RSA retains the native single-image condition, disables SS-only GoT, and reads the complete 49-frame bank without OFA's quadrant selection. Stage-specific Readers occupy the final four conditioning blocks of the SS and Shape flows. Shape obtains video K/V through the recorded source sets. With $x_\tau = (1-\tau)x_0 + \tau\epsilon$, the stage objectives are

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{SS}} + \mathcal{L}_{\text{Shape}}, \qquad \mathcal{L}_* = \mathbb{E}_{\tau,x_0,\epsilon}\left[\| v_\theta^*(x_\tau, \tau, \text{cond}_*) - (\epsilon - x_0) \|_2^2 \right], \tag{7}$$

where $* \in \{\text{SS}, \text{Shape}\}$. The adaptation sequence is II-A SS/Reader training, source caching from actual SS trajectories, then II-B Shape/Reader training. Orbit conditions are paired with 3D targets and tagged by generated or rendered origin. GT support can define Shape supervision, while sampled SS trajectories supply source records. Cached IDs and masses are fixed conditions in II-B. Appendix A.1 defines gradient handling, cache identity, and receiver configuration fields.

Bank storage is $O(FPd_B)$ for P patches per frame of encoder width d_B; regional records occupy $O(RD)$ for R coarse regions. Runtime accounting includes full-frame encoding, SS reading, source gather, Shape sampling, and decoding.

4 Experiments

4.1 Experimental Setup

Datasets and conditions. RSA pairs a complete 49-frame orbit bank with each 3D target; generated and rendered conditions retain separate provenance tags. SS sampling supplies source records for subsequent Shape adaptation (Sec. 3.4). Phase I uses approximately 167K assets from Objaverse (Deitke et al., 2023) and Texverse (Zhang et al., 2025). Blender renders each asset as a 49-frame 360° turntable sequence with controlled camera trajectories, uniform illumination, and fixed object pose. Elevation jitter around the isometric angle broadens the viewing conditions. The four-view OFA reference interface uses approximately 250K HY3D-Bench training samples with GT-rendered perspective views. We select one view per azimuth quadrant from a 24-view pool, with field-of-view augmentation from 10° to 90°. At inference, generated orbit videos replace the GT pool while preserving the four-view selection and OFA token layout. The shared encoder and OFA condition layout connect GT training views with generated inference views. The four-view OFA reference architecture is provided in Fig. 9 in the appendix.

Training. Phase I initializes from image-conditioned Wan2.2 (Wan et al., 2025). We freeze the video backbone and train high- and low-noise LoRA branches with GOA in selected high-noise DiT blocks. The high-noise branch adapts coarse structure and orbit dynamics, while the low-noise branch refines appearance. Their separate updates accommodate the different roles of early and late denoising. Training uses 512×512 images, followed by zero-shot 1024×1024 inference. The video/OFA

configurations use 48 NVIDIA H20 GPUs: DeepSpeed ZeRO-2 trains video adapters, while AdamW fine-tunes the TRELLIS.2 SS and Shape flows at learning rate 5×10^{-5} and effective batch size 384. DINOv3 features form the OFA global and local conditions. RSA trains the SS Reader on complete 49-frame banks in II-A, records regional sources, then adapts the Shape Reader with fixed records in II-B (Sec. 3.4).

Baselines and metrics. Video comparisons include Wan2.2, Wan2.1, and HunyuanVideo1.5. The 3D comparison includes Hi3DGen (Ye et al., 2025), Step1X-3D, TRELLIS (Xiang et al., 2025b), TRELLIS.2 (Xiang et al., 2025a), Hunyuan3D-2.1 (Team, 2025), InstantMesh (Xu et al., 2024), and Hunyuan3D-2MV. On HY3D-Bench, video evaluation compares generated sequences with GT renderings using PSNR, SSIM, LPIPS, and FVD. PSNR and SSIM capture pixel-level agreement, LPIPS measures perceptual distance, and FVD summarizes video-distribution differences. ULIP-I and Uni3D-I measure image–3D embedding agreement on validation and test splits. The F-score@1% columns use a candidate distance threshold of 1% of the GT bounding-box diagonal; Toys4K follows single-image-to-3D generation rather than VAE reconstruction. Multi-view renderings complement these scores by exposing backside completion and local shape differences; direct surface and cost protocols are specified in Appendix A.2.

4.2 Generation Quality and Consistency

Orbit-video quality. Table 1 evaluates the intermediate video conditions before 3D generation. V23D has the best values across all four metrics, reaching PSNR 25.98 and FVD 176.8, compared with Wan2.2's 24.96 and 205.8. The gains span frame-level similarity and sequence-level distribution quality. Figures 4 and 5 show the corresponding turntable sequences: large azimuth changes expose appearance drift and shape inconsistency in the compared video models, while the V23D examples retain more coherent object identity and structure. These observations motivate using adapted orbit videos as multi-view conditions for the 3D generator.

Table 1: Quantitative comparison of generated orbit videos.

	Video Metrics			
Method	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FVD (↓)**
Wan2.2	24.96	0.957	0.067	205.8
Wan2.1	24.41	0.932	0.076	232.2
HunyuanVideo1.5	23.63	0.889	0.094	288.4
V23D	**25.98**	**0.969**	**0.059**	**176.8**



Figure 4: **Orbit-video generation.** V23D preserves object appearance and structure across large azimuth changes.

Image-to-3D comparison. Table 2 compares OFA and RSA configurations with feed-forward and native 3D generation systems. V23D (RSA) scores 0.2418 ULIP-I and 0.3827 Uni3D-I on the test split, leading the listed methods on both splits. The four-view OFA row remains the reference interface. Figure 6 presents renderings and enlarged local crops. Back-facing surfaces and self-occluded parts reveal differences in completeness, connectedness, and repeated structures that are less apparent from a frontal view. The V23D examples show more coherent backside completion in these cases. Embedding scores and such local visual observations provide complementary views of output quality rather than interchangeable geometry measures.

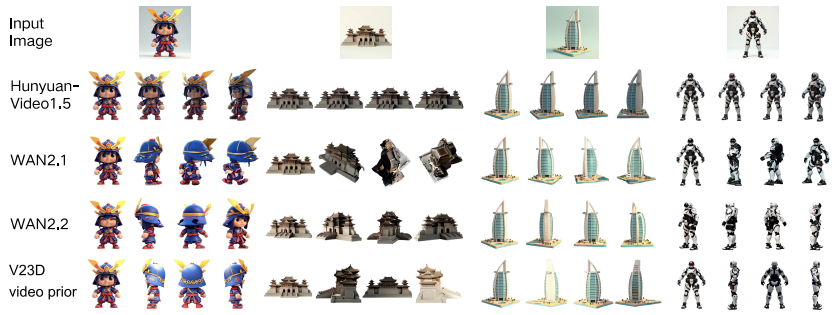

Figure 5: **Video comparison.** Generated turntable sequences illustrate cross-frame consistency against video-model baselines.

Table 2: Single-image-to-3D comparison. HY3D-Bench test/val retain ULIP-I and Uni3D-I; F-score@1% is added for HY3D-Bench-test and Toys4K test. The threshold is 1% of the GT bounding-box diagonal. Best values are **bold**; second best are underlined.

	HY3D-Bench-test			HY3D-Bench-val		Toys4K-test
Model	**ULIP-I (↑)**	**Uni3D-I (↑)**	**F-score@1% (%) ↑**	**ULIP-I (↑)**	**Uni3D-I (↑)**	**F-score@1% (%) ↑**
Hi3DGen	0.2093	0.3242	53.8	0.2040	0.3275	56.7
Step1X-3D	0.2070	0.3233	50.9	0.2033	0.3211	54.1
TRELLIS	0.2187	0.3321	52.6	0.2133	0.3410	55.3
TRELLIS.2	0.2197	0.3412	55.9	0.2102	0.3563	60.4
Hunyuan3D-2.1	0.2139	0.3569	54.7	0.2122	0.3501	58.6
InstantMesh	0.2034	0.3171	48.8	0.2021	0.3298	51.9
Hunyuan3D-2MV	0.2128	0.3510	56.2	0.2111	0.3495	60.8
V23D + OFA	0.2342	0.3721	57.4	0.2311	0.3713	60.1
V23D (RSA)	**0.2418**	**0.3827**	**61.1**	**0.2386**	**0.3812**	**62.7**

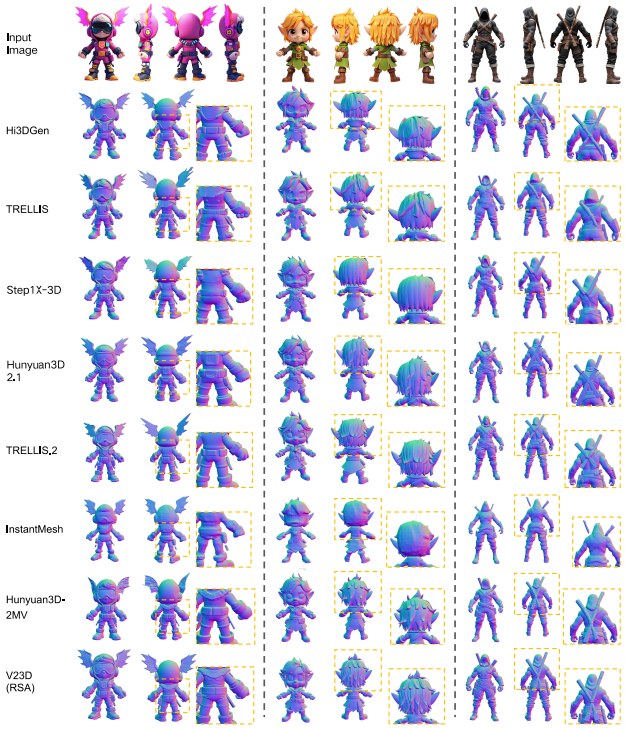

Figure 6: **Image-to-3D comparison.** Multi-view renderings and detail crops show the geometry of V23D (RSA) and competing methods.

4.3 Ablation Studies

The ablations examine video adaptation, OFA conditioning, and RSA source transfer and re-reading (Tables 3–5).

LoRA capacity and GOA. Table 3 separates adaptation capacity from orbit regularization. Without GOA, increasing LoRA rank from 64 through 96 to 128 improves all four video metrics, indicating that turntable generation benefits from sufficient adaptation capacity. Adding GOA at rank 128 further reduces FVD from 182.2 to 176.8 and LPIPS from 0.061 to 0.059. Figure 7 compares the corresponding videos across widely separated viewpoints. The rank-128 + GOA configuration combines the stronger LoRA adaptation with orbit-level regularization, yielding more consistent visible structures in the displayed sequences.

Table 3: Video adaptation ablation with varying LoRA rank and GOA.

Method	**PSNR** (↑)	**SSIM** (↑)	**LPIPS** (↓)	**FVD** (↓)
LoRA rank 64 *w/o* GOA	23.83	0.899	0.071	187.2
LoRA rank 96 *w/o* GOA	24.56	0.941	0.063	186.5
LoRA rank 128 *w/o* GOA	25.87	0.955	0.061	182.2
LoRA rank 128 + GOA	**25.98**	**0.969**	**0.059**	**176.8**

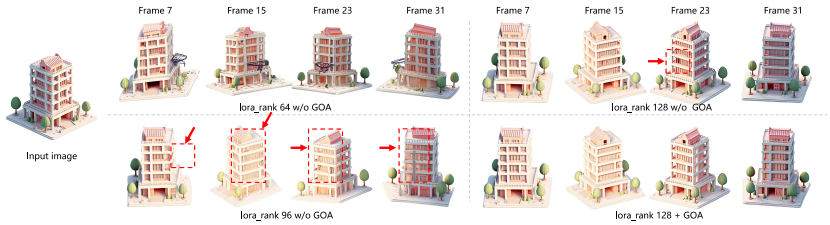

Figure 7: **Video adaptation ablation.** We compare LoRA ranks and the addition of GOA at rank 128.

SS adaptation and OFA conditioning. Table 4 analyzes the four-view generation configuration. Removing SS Flow fine-tuning produces the largest decrease in both embedding scores, supporting adaptation of the sparse-structure receiver to multi-view conditions. Replacing OFA with raw DINOv3 token concatenation also lowers both metrics, indicating a benefit from organizing the selected views into structured conditions. Removing GoT causes a larger Uni3D-I decrease than removing LvT, consistent with the global summary's role in coarse structure. LvT supplies view-specific information during refinement, and its removal also lowers both scores. Figure 8 extends the visual comparison across SS Flow fine-tuning, OFA, GOA, and RSA, showing differences in coarse completeness and local detail. Together, the table and figure cover receiver adaptation, global and local conditioning, and the two proposed adapters.

Table 4: SS fine-tuning and OFA conditioning ablations.

Strategy	**ULIP-I** (↑)	**Uni3D-I** (↑)
w/o SS Flow FT	0.2198	0.3514
w/o OFA (raw concat)	0.2268	0.3584
w/o GoT	0.2256	0.3549
w/o LvT	0.2294	0.3626
Full OFA	**0.2342**	**0.3721**



Figure 8: **Ablations across the full framework.** Visual examples illustrate controls without SS Flow fine-tuning, OFA, GOA, or RSA alongside the full model.

Regional source transfer and patch re-reading. Table 5 compares full-orbit interfaces E1–E4 with the OFA configuration E0. E1 aggregates all-frame content without regional source records. E2 chooses sources from current Shape states instead of transferring SS choices, whereas E3 transfers fixed content from the same retained sources instead of re-reading original patches. E2 refreshes routing at positive-condition Shape Reader calls and includes the additional full-bank search cost. E3 keeps the source IDs, masses, support, and compatible feature width fixed, changing what is transferred to Shape. The full RSA row exceeds each alternative on both scores. The fixed-model E5 intervention shuffles source IDs, lowering both scores while preserving receiving regions' masses, support, and Shape noise. This separates the correspondence encoded by the IDs from the quantity of conditioning. Appendix A.2 specifies matched conditions, routing costs, and fixed-support/Shape-noise controls to distinguish Shape reading from changes in coarse support.

Table 5: Interface comparison on HY3D-Bench test. F counts encoded video frames; RSA retains $D = 4$ sources per region. E0–E4 use the end-to-end protocol; E5 intervenes on E4 at inference. For F-score@1%, the candidate distance threshold is 1% of the GT bounding-box diagonal. Best values are **bold**; second best are underlined.

ID	Method	F	ULIP-I ↑	Uni3D-I ↑	F-score@1% (%) ↑
E0	V23D + OFA	4	0.2342	0.3721	57.4
E1	All-frame aggregation	49	0.2363	0.3756	58.2
E2	RSA w/o source transfer	49	<u>0.2380</u>	<u>0.3784</u>	<u>59.5</u>
E3	RSA w/o patch re-reading	49	0.2371	0.3769	59.0
E4	RSA (full)	49	**0.2418**	**0.3827**	**61.1**
E5	RSA w/ shuffled source IDs	49	0.2309	0.3672	55.8

4.4 User Study

We compare V23D (RSA) against TRELLIS.2, Hunyuan3D-2.1, and Hi3DGen on 50 randomly sampled HY3D-Bench test examples. 30 participants evaluate randomized, anonymized multi-view renderings and select their preferred result considering geometric completeness, backside plausibility, structural consistency, and overall quality. Randomization and anonymization reduce cues from method names and display order. V23D (RSA) receives 69.9% of valid votes. This perceptual comparison complements the embedding metrics by asking participants to inspect the generated assets from multiple viewpoints, including surfaces not visible in the input image.

5 Conclusion

V23D connects an orbit-video prior to coarse-to-fine 3D generation through RSA. RSA records regional sources from the complete orbit and lets Shape queries re-read their patches. Separating stable source selection from state-dependent content extraction preserves regional video access across generation stages.

Limitations. Generated views may share errors. Fixed source sets exclude omitted views, and Shape depends on SS support. Content attention provides no calibrated surface correspondence; retained mass also reflects truncation. Encoding and reading all 49 frames increases computational cost.

6 AI USE STATEMENT

During the preparation of this work, the authors used ChatGPT to improve language, readability, and presentation. The authors reviewed and edited all model-assisted content as needed and take full responsibility for the content of the publication.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. ReconViaGen: Towards accurate multi-view 3D object reconstruction via generation. *arXiv preprint arXiv:2510.23306*, 2025. URL https://arxiv.org/abs/2510.23306v1.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22246–22256, 2023.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M Rehg. How much 3d do video foundation models encode? *arXiv preprint arXiv:2512.19949*, 2025.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Siyou Lin, Zhou Xue, Hongwen Zhang, Liang An, Dongping Li, Shaohui Jiao, and Yebin Liu. Mix3R: Mixing feed-forward reconstruction and generative 3D priors for joint multi-view aligned 3D reconstruction and pose estimation. *arXiv preprint arXiv:2605.03359*, 2026. URL https://arxiv.org/abs/2605.03359v1.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Zefan Qu, Zhenwei Wang, Gerhard Petrus Hancke, and Rynson W. H. Lau. UMI3D: Robust 3D generation on unconstrained multi-image inputs via simultaneous focus cross-attention routing. *arXiv preprint arXiv:2607.24298*, 2026. URL https://arxiv.org/abs/2607.24298v1.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Hanxiao Sun, Mingxin Yang, Shuhui Yang, Zebin He, Xintong Han, Hongbo Fu, Chunchao Guo, and Wenhan Luo. ROAR-3D: Routing arbitrary views for high-fidelity 3D generation. *arXiv preprint arXiv:2605.21121*, 2026. URL https://arxiv.org/abs/2605.21121v1.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *Computer Vision – ECCV 2024*, pp. 1–18, 2024.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. *arXiv preprint arXiv:2012.01411*, 2020. URL https://arxiv.org/abs/2012.01411v1.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025b.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25050–25061, 2025.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3d objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025.

Kaichen Zhou, Zeyang Bai, Xinhai Chang, Mengyu Wang, Paul Liang, and Fangneng Zhan. Stream3D: Sequential multi-view 3D generation via evidential memory. *arXiv preprint arXiv:2605.21472*, 2026. URL https://arxiv.org/abs/2605.21472v5.

A RSA CONFIGURATION AND EVALUATION

A.1 RSA DESIGN AND RECEIVER PROFILE

RSA connects the shared patch bank to SS and Shape through regional source records. Table 6 gives a compatible default receiver profile.

Table 6: RSA default interface and receiver profile.

Component	Adopted design
Bank	$F = 49$ frozen DINOv3 ViT-L/16 features at 512^2; original patch addresses retained
Readers	Final four SS/Shape conditioning blocks; width 1536, 12 heads ($d = 128$); separate stage projections, $b^{SS} = b^{H} = 0$
Image condition	Native single-image ports retained; SS-only GoT disabled; no OFA quadrant selection
Recording	Final positive SS Reader call; full coarse-grid source records
Sources	Initial $D = 4$ per region; original retained masses, null remainder, and zero-mass padding
Shape	Fixed J/m; independent current queries re-read original patches; support remains separate
Training	II-A SS Reader $\rightarrow$ sampled source cache $\rightarrow$ II-B Shape Reader
Sampling	TRELLIS.2 sampler and CFG settings shared across E1–E4

Receiver bindings. The bank uses the ViT-L/16 feature width of 1024 and stage-specific K/V projections into 1536-dimensional Reader space. Any SS pooling retains a map to the original patch grid for Shape gathering. The address convention Γ maps Shape coordinates to SS regions; the null value contributes zero. All interface controls share the same native-image ports and sampler settings.

Attention heads. Equations 5–6 show one head with key width d. In a multi-head Reader, each head computes its own weighted value sum; their concatenation is passed through W_O^H. All heads share the recorded J/m. The bank feature width is set by the encoder and need not equal d.

Coordinates and reads. Frame and patch indices address processed images; the preprocessing map identifies original pixels and padding. SS pooling retains an original-grid correspondence for Shape gathering. The address convention Γ specifies canonical bounds, axes, origin, grids, rounding, and boundary/empty-address handling. Camera metadata records provenance; content attention uses $b^{SS} = b^{H} = 0$.

Adaptation. II-A adapts the SS Reader and its condition projections. The resulting SS/Reader version samples complete source tables for II-B, which adapts the Shape Reader and its condition projections with SS source records fixed. Training times are sampled independently. Conditions and 3D targets share asset identity and split; generated and GT-rendered conditions carry separate tags and a mixture policy. GT support defines a supervision domain, while source records come from SS sampling. Cached J/m are fixed conditions. Frozen receiver parameters preserve gradients to Reader inputs. Initialization leaves an active gradient path, and output projections satisfy $W_O(0) = 0$. Empty/padded reads return finite zeros without an empty softmax.

Cache scope. Cache identity includes the asset/split, bank and model hashes, preprocessing, address convention, sampler/steps/CFG/seed, writing layer/call, and D/tie/padding/mass policies. Changing these fields invalidates the cache. Positive-call records are isolated from negative CFG evaluations; native-image and GoT settings are included in the receiver configuration.

Algorithm 1 RSA inference: record once, re-read during refinement.

Require: Image I, orbit $\mathcal{V}$, receiver/sampler configuration Π, address convention Γ

Encode all 49 frames into $\mathcal{B}$; obtain native image conditions only at ports specified by Π.
Sample SS with its current-state Reader; isolate positive attention from negative CFG calls.
At the designated final positive call, write full-grid $\mathcal{M}$ using Eq. 4.
Decode the SS latent into sparse support $\mathcal{S}$; initialize Shape on $\mathcal{S}$.
Map Shape coordinates to $r(q)$ using Γ; fix $J_{r(q)}$ and $m_{r(q),f}$.
for each Shape sampler step **do**
    At conditioned model evaluations, gather original patches from $J_{r(q)}$ for Reader K/V.
    Use current Shape queries in Eq. 5; inject Eq. 6 as a residual.
    Complete the model calls required by Π, then advance the sampler without rewriting $\mathcal{M}$.
end for
Decode the final Shape latent using the retained asset/material generation stack.

A.2 RSA Evaluation Protocol

Tables 5 and 7 define the comparison set and measurement protocol.

Shared conditions. E1–E4 use the same paired 49-frame bank, asset split, receiver initialization, native-image and SS-only GoT settings, and trainable-module budget. An asset manifest records video/3D pairing and generated versus GT-rendered conditions. Evaluation fixes cameras, alignment, sampling, region definitions, paired seeds, and failure handling. In Table 5, F counts frames encoded for video conditioning: E0 selects four globally, while E1–E5 encode all 49. The regional budget $D = 4$ is separate from F.

Aggregation and routing controls. E1 concatenates valid patch tokens from all 49 frames and uses ordinary full-bank cross-attention at the SS and Shape condition ports, without a source table. Query/output token counts, residual locations, projection widths, and head counts match E4; the full-bank Shape attention cost is counted separately. E2 selects regional sources at each positive-condition Shape Reader evaluation. Its routing attention normalizes jointly over valid frame–patch pairs and null, then sums patch weights within each frame to obtain comparable frame scores. Top-D selection retains the original frame masses; current Shape queries subsequently re-read selected patches using Eq. 5. Regional query construction, routing blocks, and sharing belong to the receiver profile. Full-bank routing and projection costs are included.

Fixed-content control. E3 retains E4's J_r, $m_{r,f}$, null remainder, and support, but supplies fixed content. At the SS recording call, define $\beta^{SS}_{r,f,p} = \bar{A}^{SS}_{r,f,p}/m_{r,f}$ for positive frame mass and zero otherwise. Then

$$c_r = \sum_{f \in J_r} m_{r,f} \sum_p \beta^{SS}_{r,f,p} V^H_{f,p}. \tag{8}$$

This definition requires SS weights on the original Shape patch grid, or an explicit pooling-to-original map. Mapped conditional weights are nonnegative and sum to one over valid original patches for positive frame mass. The value space, payload width, and residual locations match E4. For a fixed checkpoint and bank, c_r remains constant throughout one Shape sampling trajectory. During training, fixed original features and SS weights are combined with the current value projection; projected caches are refreshed whenever that projection changes. E3's budget includes its SS patch weights. E4 instead re-reads original patches with current Shape queries.

Source intervention and attribution. E5 permutes source-ID lists within compatible-length region groups at inference, with the E4 checkpoint fixed. Each receiving region retains its ordered mass vector, null, source count, support, and initial Shape noise. Records include index validity, duplicate policy, changed-ID fraction, and singleton groups. Table 5 uses the end-to-end protocol for E0–E4, with generated SS support; E3 shares E4's SS trajectory, support, and J/m. E5 is a paired intervention on E4 rather than a separately trained model. A complementary Shape-only protocol fixes support and noise across interfaces to isolate reading effects. Asset-linked case studies cover complementary views, source omissions, repeated-part confusion, and missing support.

Table 7: Evaluation metrics and cost-accounting protocol.

Measurement	Definition or breakdown
ULIP-I / Uni3D-I ↑	Matched splits, inputs, scoring versions
Surface distance ↓	Alignment, canonical units, point sampling
F-score ↑ / normals	Fixed threshold / fixed cameras and masks
Regional source statistics	Occupied-region s_r, IDs, off-four-view mass
Condition residual scale	Video increment / main-path norm by block
Online runtime / peak memory	Stage timings, hardware, precision, cache state
Offline cache cost	Construction, storage and I/O, separately timed
Failures / uncertainty	Attempted assets, paired seeds, aggregation

Geometry and source diagnostics. Surface-distance and F-score evaluations specify alignment, units, point sampling, and thresholds; normal evaluation specifies cameras, visibility masks, orientations, and empty-pixel treatment. Regions are defined independently of favorable outputs. ULIP-I and Uni3D-I complement these geometry metrics. Source statistics include occupied-region s_r, source IDs, mass outside the four selected views, and projected video-residual/main-path norm ratios with defined zero handling. Empty regions are tabulated separately. Mass-renormalization comparisons control residual amplitude.

Cost and uncertainty. Online accounting covers video generation, DINO encoding and bank storage, SS reading/sampling, recording/gather, Shape reading/sampling, and asset decoding. Exclusive stage times and end-to-end wall time are separated from offline cache construction/storage/I/O. Reports specify hardware, batch size, precision, synchronization, warm-up, cache state, attempted assets, failures, and denominators. Paired per-asset differences use uncertainty estimates matched to the asset/seed sampling structure.

B Training and Evaluation Details

This section describes video adaptation, OFA conditioning, rendering, and perceptual evaluation.

B.1 Training Details

Phase I: Geometry-aware Video Prior Extraction. We base our multi-view generator on the Wan2.2-Fun-A14B-InP architecture, a 14B-parameter image-conditioned video diffusion model with a dual-DiT design for high-noise and low-noise denoising regimes. Training is distributed across 48 NVIDIA H20 GPUs with 96GB memory, spanning 4 nodes, and uses DeepSpeed ZeRO-2 optimization. We freeze the base Wan2.2 parameters and train two noise-specific LoRA branches together with the proposed Geometry-aware Orbit Adapter (GOA). The LoRA rank is set to $r = 128$, targeting the attention layers (`q, k, v, o`) and feed-forward layers (`ffn.0, ffn.2`). The high-noise LoRA branch focuses on global structure and orbit dynamics, while the low-noise branch preserves object identity and high-frequency appearance details.

The video model is trained with a warm-restart strategy. After completing the first epoch, which contains 5,240 steps per noise stage with a learning rate of 5×10^{-5}, we reset the optimizer state and conduct a second epoch to further stabilize cross-frame consistency. GOA is inserted into selected high-noise DiT blocks and predicts lightweight orbit-geometry residuals. Its residual gate is zero-initialized to avoid disturbing the pretrained video prior at the beginning of training. During training, Blender-rendered assets provide auxiliary orbit-level supervision, including camera pose, silhouette consistency, and first-last frame closure. These signals are used only as training-time regularization, and no additional 3D supervision or reconstruction module is required at inference.

Resolution Scaling and Accelerated Inference. Phase I is trained at 512×512 resolution and then zero-shot upscaled to 1024×1024 during inference.

We also use a custom Lightning LoRA with $\alpha = 0.125$ during inference, which reduces the number of sampling steps from 50 to 20 and allows a small CFG scale of 1.0. This setting accelerates turntable video generation while suppressing high-frequency artifacts and maintaining stable object identity across the 49-frame orbit sequence.

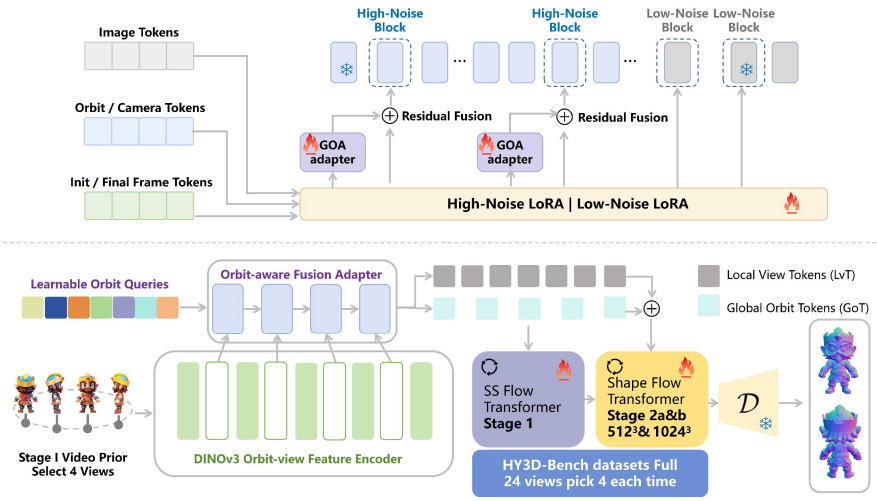

Figure 9: **Four-view OFA reference interface.** The video prior supplies four selected orbit views to the DINOv3 encoder and OFA, yielding global orbit tokens for SS and local view tokens for Shape. The upper panel shows the retained GOA video adaptation.

Phase II: OFA-conditioned 3D Sparse Transformer Tuning. For Phase II, we initialize from the official TRELLIS.2 4B sparse transformer. Different from Phase I, Phase II is trained with full-parameter fine-tuning because the 3D sparse transformer must adapt to a new global-local multi-view conditioning layout. The training conditions are constructed from the HY3D-Bench training set. For each training asset, we use a GT-rendered 24-view perspective pool and select four representative views using the azimuth-aware quadrant protocol. These selected GT-rendered views are encoded by DINOv3 (Siméoni et al., 2025), fused by the Orbit-aware Fusion Adapter (OFA), and converted into Global Orbit Tokens (GoT) and Local View Tokens (LvT).

For the OFA configuration, training uses GT-rendered multi-view conditions; Stage-I generated videos supply the inference conditions. At inference time, the GT-rendered view pool is unavailable, so the frozen Stage-I video generator first synthesizes a 49-frame orbit video from the input image. The same quadrant-based selection strategy and OFA token layout are then applied to the generated orbit views. This design allows Phase II to learn from clean and accurately posed multi-view observations while remaining compatible with generated video priors at inference.

We fine-tune the TRELLIS.2 SS Flow and Shape Flow transformers with these structured conditions. SS Flow is conditioned by GoT for coarse sparse-structure generation, while Shape Flow is conditioned by both GoT and LvT for fine geometric detail generation. All flow-matching stages are optimized using AdamW with a learning rate of 5×10^{-5} and an effective batch size of 384. Each major Phase-II flow-matching stage takes approximately 10 days to train on the same 48-GPU H20 cluster. The training pipeline adapts both the video prior and the 3D foundation model.

GOA auxiliary losses. The video training specification combines

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{video}} + 0.05\mathcal{L}_{\mathrm{orb}} + 0.05\mathcal{L}_{\mathrm{sil}} + 0.02\mathcal{L}_{\mathrm{cam}}. \tag{9}$$

$\mathcal{L}_{\mathrm{orb}}$ measures DINO-feature ℓ_2 distance between first and last orbit frames. $\mathcal{L}_{\mathrm{sil}}$ measures mean ℓ_1 distance between generated and rendered silhouettes. $\mathcal{L}_{\mathrm{cam}}$ uses azimuth/elevation ℓ_1 regression from pooled GOA features.

B.2 Data Processing and Augmentation

Phase I Turntable Video Rendering. For Phase I video prior extraction, we construct a large-scale turntable-video dataset from Objaverse (Deitke et al., 2023) and Texverse (Zhang et al., 2025), containing approximately 167K high-quality 3D assets. Each asset is rendered in Blender as a 49-frame 360° orbit sequence with controlled camera trajectories, uniform illumination, and fixed object pose. All Phase-I renderings use an orthographic camera projection to provide stable turntable supervision and reduce perspective-induced scale variation.

To improve robustness to unconstrained in-the-wild input images, we stratify the rendering pipeline into three elevation distributions: (1) **Horizontal:** 0° elevation, producing approximately 52K videos; (2) **Strict Isometric:** fixed at the standard isometric angle of 35.26°, producing approximately 52K videos; and (3) **Jittered Isometric:** 35.26° with uniformly sampled perturbations of $\pm 20^\circ$, producing approximately 63K videos. This elevation jittering acts as an important regularizer, encouraging the video prior to decouple object structure from a fixed camera elevation.

OFA Training Views. For OFA-conditioned 3D generation, we use approximately 250K HY3D-Bench training samples/assets with GT-rendered multi-view observations. Phase I uses orthographic rendering, while OFA training uses perspective views.

For each HY3D-Bench training asset, we construct a GT-rendered 24-view perspective pool. To improve scale robustness, the Field of View (FOV) is dynamically sampled from 10° to 90°. The camera distance is adjusted accordingly so that the object remains tightly framed inside the image boundary. Camera elevations are sampled from typical viewing ranges, e.g., $[-10^\circ, 45^\circ]$. These GT-rendered 24-view pools provide stable and accurately posed multi-view supervision for Phase-II fine-tuning.

Azimuth-aware Quadrant Selection. During multi-view condition construction, we set $M = 4$ and enforce an azimuth-aware quadrant selection strategy. Before selecting views, we apply a random global azimuth offset $\Delta\theta \in [0^\circ, 360^\circ)$ to prevent the model from overfitting to canonical front or side views. The full 360° azimuth space is then partitioned into four 90° quadrants, and exactly one view is sampled from each quadrant. This strategy produces four-view conditions with broad orbit coverage and mitigates the blind spots caused by naive random or adjacent-view sampling.

During Phase-II training, the candidate view pool is the GT-rendered 24-view pool of each HY3D-Bench training asset. During inference, the candidate view pool is the 49-frame orbit video generated by the frozen Stage-I video prior. In both cases, the same quadrant-selection rule and OFA token layout are used. This shared layout reduces the discrepancy between GT multi-view training conditions and generated video-prior inference conditions.

B.3 Additional Qualitative Results

Figure 10 shows additional multi-view asset comparisons for V23D (RSA).



Figure 10: **Additional V23D (RSA) qualitative results.** Multi-view comparisons of V23D (RSA) with Hunyuan3D-2MV.