AO-FLOW: LEARNING TO USE AND OPTIMIZE GENERATED VIEWS FOR 3D ASSETS

Anonymous authors

ABSTRACT

Orbit videos provide candidate observations for single-image 3D generation, yet different regions need different views and adjacent frames can repeat the same error. AO-Flow first adapts a 3D receiver to use the full 49-frame video: sparse-structure sampling retains region-specific source identities and attention masses, and Shape rereads those frames with its own queries and a soft spatial prior. Sequential supervision trains the source producer before freezing it for Shape adaptation. AO-Flow then fixes the receiver and uses final-asset feedback to update only video attention LoRA. The regional interface reaches H-F 0.639 versus 0.613 for a global four-source reader at similar receiver cost; full 49-frame reading attains slightly higher H-F at greater cost. Asset feedback raises H-F to 0.669 while trading some input-view agreement. Inference uses one image, one video and one unselected asset.

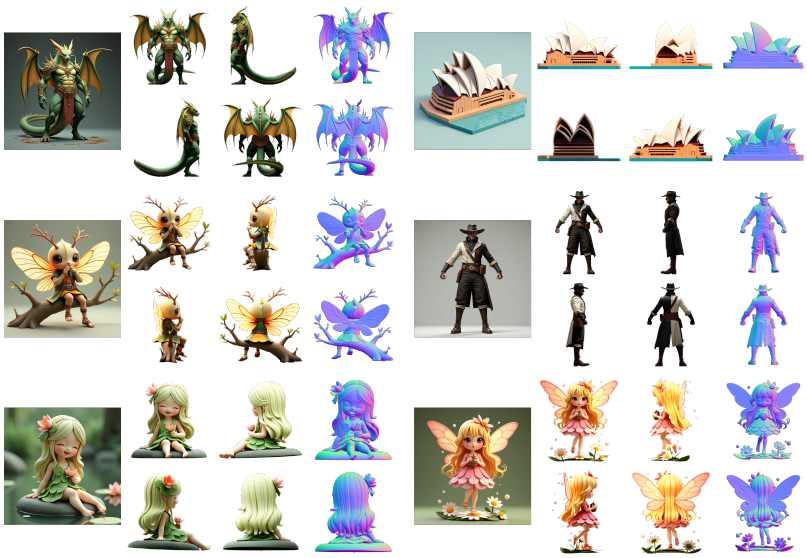

Figure 1: **Single-image asset examples.** Each group pairs an input image with four generated views and front/back normals of its matched mesh.

1 INTRODUCTION

Generating a complete 3D asset from one image requires preserving visible structure and inferring hidden surfaces. Figure 1 illustrates this setting with input images, generated views and normal renders of matched meshes. Orbit-video priors supply candidate views (Chen et al., 2024; Voleti et al., 2024). A narrow attachment or opening may appear only at an intermediate angle, providing information that canonical views miss.

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.

Different regions can benefit from different frames, while adjacent frames may repeat the same geometric error. A shared subset can discard useful local observations, and additional frames introduce redundancy and conflicts. The challenge is to expose useful regional content to the 3D generator and improve the video that supplies it.

Multi-image routing and evidential selection address which views individual 3D queries should use (Qu et al., 2026; Sun et al., 2026; Zhou et al., 2026). We study how source choices pass between generation stages. Sparse-structure (SS) sampling establishes the support on which Shape refines local geometry (Xiang et al., 2025). This motivates retaining regional sources identified during coarse generation as candidates for refinement, while Shape recomputes content attention for its current queries.

AO-Flow exposes all 49 frames from an adapted orbit-video source (Anonymous, 2026) to its geometry receiver. During its actual sampling trajectory, SS reads the feature bank and records source identities and attention masses for fixed spatial regions. Shape reads each region's retained sources with a soft spatial bias, preserving the native input-image condition. Fixed latent-grid addresses define these regions before mesh generation. We train SS first, freeze its memory producer, then adapt Shape using memories from actual SS sampling.

Receiver adaptation changes how a generated video is read but not which videos are sampled. The source must therefore provide views that support the final asset objective (Xie et al., 2024; An et al., 2026). We score final-asset geometry, unobserved-surface coverage and input preservation, then use Flow-GRPO (Liu et al., 2025a) to update video attention LoRA with the adapted receiver and orbit adapter fixed. During video post-training, reference meshes enter only the reward. Normal inference uses one image, one video and one asset, without reference-based selection.

Our contributions are threefold:

1. A cross-stage regional interface. We transfer region-specific source identities and attention masses from full-video SS sampling to Shape, which reads the retained candidates with its own content attention. Sequential adaptation trains Shape using sampled memories after freezing the SS producer.

2. Asset-oriented video post-training. We formulate video optimization around the quality of the final generated asset and apply Flow-GRPO to video attention LoRA. Fixing the adapted receiver keeps the downstream mapping stable while the policy learns from asset feedback.

3. Controlled evaluation and mechanism analysis. We specify matched receiver and policy comparisons, local evidence interventions and stage attribution to distinguish interface design from video-source optimization. Independent checks of parts, connections and duplication complement surface metrics, input preservation and cost accounting, testing whether higher rewards correspond to better asset structure.

2 Related Work

2.1 Video priors and image-to-3D generation

Single-image generators infer unseen geometry from learned priors: InstantMesh combines multi-view diffusion with sparse-view reconstruction (Xu et al., 2024), Hi3DGen uses predicted normals as a geometry bridge (Ye et al., 2025), and TRELLIS/TRELLIS.2 sample structured 3D latents (Xiang et al., 2024; 2025). V3D and SV3D use video diffusion to synthesize candidate views (Chen et al., 2024; Voleti et al., 2024); V23D adapts an orbit-video source and conditions TRELLIS.2 on selected views (Anonymous, 2026). Ouroboros3D couples view generation and reconstruction recursively (Wen et al., 2025). AO-Flow keeps the full generated orbit available to the 3D receiver and later optimizes the video source using final assets.

2.2 Multi-view conditioning and regional information transfer

The video-to-3D interface determines which observations each region can use. ReconViaGen builds global and per-view conditions (Chang et al., 2025); UMI3D and ROAR-3D route 3D queries with attention-based view choices (Qu et al., 2026; Sun et al., 2026). Stream3D retains query-level obser-

vation evidence across sequential views (Zhou et al., 2026), while Pixal3D constructs pixel-aligned conditions under known cameras (Li et al., 2026). Ink3D instead uses geometry to guide video synthesis and texture transfer (Han et al., 2026). AO-Flow transfers regional source IDs and original attention masses from actual SS sampling to Shape, which recomputes content attention using its current query and a soft nominal-camera prior. This cross-stage reuse differs from selecting one fixed global view bundle or treating stored source weights as final Shape attention.

2.3 Geometry feedback and generative post-training

DDPO and Flow-GRPO optimize diffusion or flow generators with sample-level rewards (Black et al., 2023; Liu et al., 2025a); Flow-DPO fits preferences instead (Liu et al., 2025b). Geometry-aware video objectives include reconstruction consistency in Carve3D, video preferences in VideoGPA, latent camera/reprojection rewards in VGGRPO, and geometry-derived preferences in FixAnything (Xie et al., 2024; Du et al., 2026; An et al., 2026; Vuong et al., 2026). Nabla-R2D3 directly aligns a 3D generator using image rewards (Liu et al., 2025c). AO-Flow instead holds the adapted 3D receiver fixed and updates only video attention LoRA from final-asset feedback. Same-reward Asset-DPO, geometry-video RL, and offline best-of-four imitation separate the value of the asset target from the online update rule.

3 Method

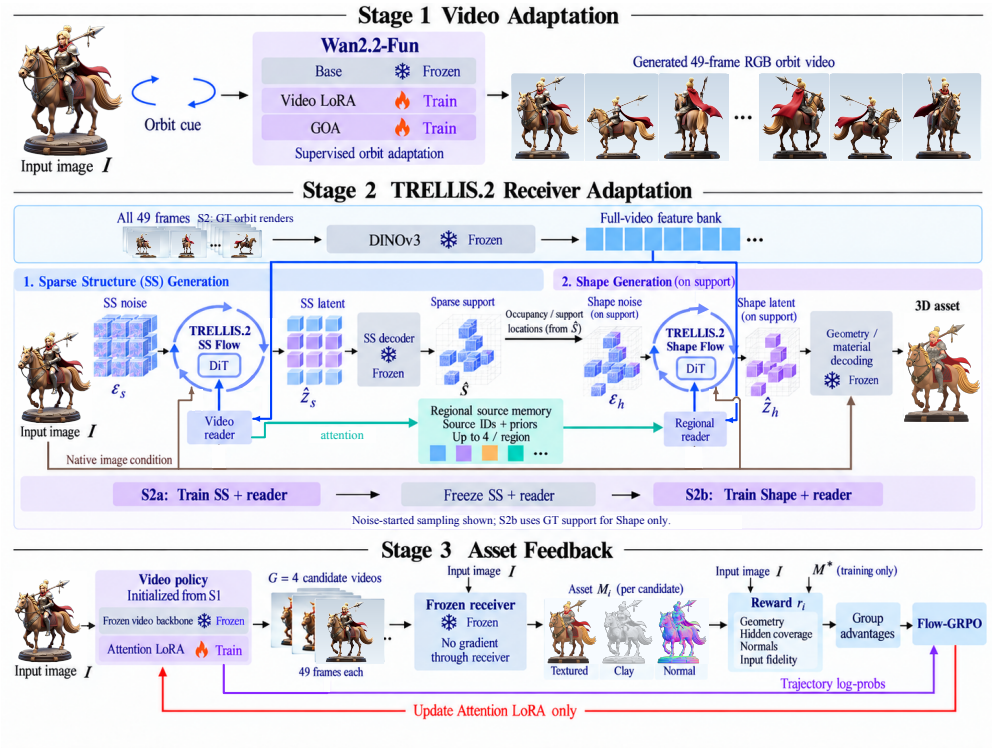

Figure 2: **AO-Flow overview.** Stage 1 generates a 49-frame orbit; Stage 2 adapts SS and Shape with regional source memory; Stage 3 freezes the receiver and updates video attention LoRA from asset feedback.

3.1 Problem formulation and framework overview

From one image I and nominal orbit c, adapted Wan2.2-Fun generates a 49-frame video $V = \{x_j\}_{j=0}^{48}$ (Wan Team, 2025; aigc-apps, 2026). The 3D receiver samples sparse structure and refines geometry with regional video sources, yielding asset M:

$$V \sim p_\theta(\cdot \mid I, c), \qquad M = F_\phi(I, V; \xi), \tag{1}$$

Here ϕ denotes adapted SS/Shape denoisers and readers, ξ receiver randomness, and θ the video attention LoRA optimized in Stage 3. Encoders, material generation and decoders remain fixed; generator and receiver share the nominal orbit convention.

Video source and asset receiver. The video source uses V23D's video LoRA and Geometry-aware Orbit Adapter (GOA) (Anonymous, 2026; Hu et al., 2021). GOA adds orbit-conditioned residuals to high-noise video blocks. TRELLIS.2 Sparse Structure (SS) Flow predicts support and Shape Flow generates geometry on it (Xiang et al., 2025). Our SS reader uses all 49 frames and records regional sources for Shape. Both stages retain the native DINO condition from I; fixed material generation and decoding also use I and the generated shape latent.

Training and inference. Stage 1 supplies the orbit-adapted video initializer. Stage 2 trains SS and its reader, freezes that memory producer, then trains Shape and its reader. Stage 3 fixes the receiver and updates video attention LoRA from final-asset rewards. Inference uses one image, video and asset, without reference meshes or candidate selection. Appendix A specifies the input-aligned coordinate adapter for the soft spatial prior.

3.2 Cross-stage regional video-to-3D interface

Coarse structure generation and local shape refinement draw on different information from a generated orbit. Our interface retains region-specific source choices from the actual SS sampling trajectory and passes them to Shape, which recomputes content attention over those candidates. Regional routing (Qu et al., 2026; Sun et al., 2026) and source-index memory (Zhou et al., 2026) are related; Appendix J compares their mechanisms with our cross-stage interface.

Full-video reading in SS. Frozen DINOv3 provides $N = 1024$ patch features per 512^2 frame (Siméoni et al., 2025); this complete bank remains available throughout geometry generation. Separate width-128 SS and Shape readers query it after native image cross-attention. Position and nominal-camera encodings provide soft cues; the input retains its native special tokens.

SS attends to all valid video patches and a learned null entry. Repeated orbit endpoints split one angle's prior weight; Appendix F specifies normalization and masks.

Regional source memory. SS has a fixed 16^3 latent-query grid. At the last actual positive model call of the same SS sampling trajectory, we aggregate attention by source frame and retain each region's top-D candidates:

$$e_{rj} = \sum_p A^s_{rjp}, \qquad J_r = \mathrm{TopD}_j(e_{rj}), \qquad m_{r\perp} = A^s_{r\perp} + \sum_{j \notin J_r} e_{rj}. \tag{2}$$

Here e_{rj} is frame j's SS attention mass, J_r the retained set and $m_{r\perp}$ the original null plus discarded mass. Up to $D = 4$ positive-mass sources are retained per region without renormalization; empty slots have zero mass. All 4,096 addresses remain available regardless of predicted occupancy. Regions can retain different frames while image content stays in the full bank; Appendix F gives tie-breaking rules.

Regional reading in Shape. For a Shape support cell $\mathbf{c}_h$ on the native 32^3 grid, we obtain its parent address by the componentwise lookup $\lfloor \mathbf{c}_h/2 \rfloor$ and denote the corresponding SS row by r. This lookup links latent-grid cells; an address can cover multiple surfaces. For the query at the cell's native center x, the reader recomputes content logits $\ell^h_{xjp} = q^h_x(k^h_{jp})^T/\sqrt{128}$ over every valid patch in each retained frame. The stored SS masses serve as priors, so Shape's final allocation can change with its own state.

Pixel-aligned conditioning relates 3D locations to image features (Li et al., 2026). Our nominal camera and coordinate adapter give nearby patches a bounded spatial preference, not a hard crop. Stored SS masses and null allocation serve as priors; Shape's current content logits determine final attention. The reader injects the result through a zero-initialized projection with a relative residual cap. Figure 3 shows addresses, retention and rereading; Appendix F gives the spatial bias, residual bound, guidance and negative-branch diagnostic.

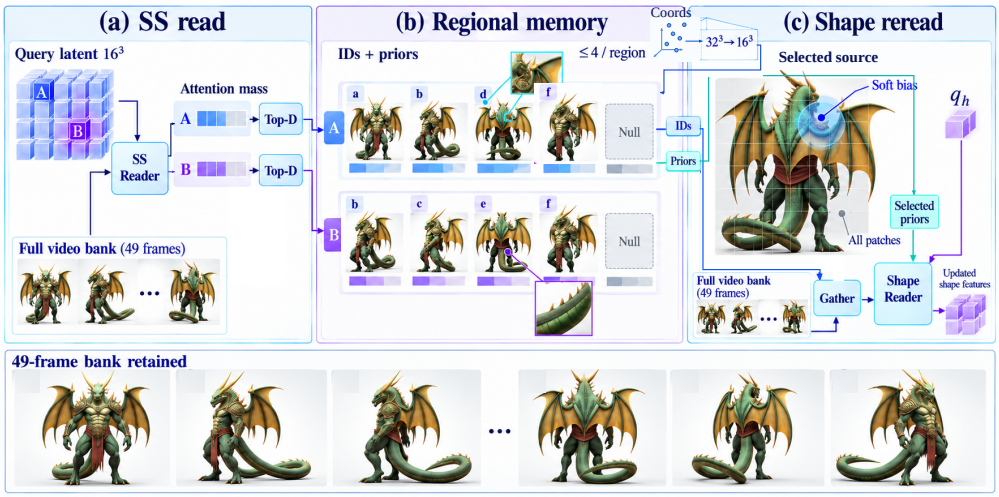

Figure 3: **Cross-stage regional conditioning.** SS reads 49 frames and retains up to four source IDs and masses per region, with discarded mass entering null. Shape rereads these sources with its current query and a soft spatial prior.

3.3 Sequential supervised adaptation

The Shape reader must learn to use memories produced by SS. Updating that producer during Shape adaptation would change the source priors supplied to the consumer. We therefore train the two stages sequentially, with independent reader parameters. Stage 1 first adapts video LoRA and GOA under the coordinate/rendering convention, providing the common video initializer π_0^{13} used in controlled comparisons.

S2a: adapt the structure producer. The primary Stage 2 condition bank uses GT orbit renders. Frozen 3D encoders provide clean latent targets z_0^a, where $a = s$ denotes SS and $a = h$ denotes Shape. For either stage, the flow target is $\epsilon - z_0^a$ at $z_t^a = (1-t)z_0^a + t\epsilon$, with loss normalized by valid latent scalars. Flow time and Gaussian noise are sampled independently by stage; the full objective is in Appendix F. $\mathcal{C}^s$ comprises the native input-image condition and full video bank; $\mathcal{C}^h$ uses the native image condition and selected video sources with their frozen SS priors. S2a updates SS and its reader. The video generator, encoders, material modules and decoders remain fixed throughout Stage 2.

S2b: adapt Shape to sampled memories. After S2a, we freeze the complete SS denoiser, its reader, and the sampling and memory rules. For each S2b record, this producer runs an actual noise-started sampling trajectory using I, the condition video and an independent SS seed. GT support, GT SS latents and reference depth do not enter this memory-producing call. Shape and its reader then learn on reference support, using a separately sampled Shape time and noise. Reference support defines only the supervised regression domain; inference uses predicted support. SS sampling, source selection and memory are detached conditioning, so predicted-support omissions remain a distinct failure mode.

The primary adaptation is flow-only; the optional surface-distribution auxiliary has zero default weight. Generated-condition Shape continuation also freezes SS. Memory-production costs, auxiliary supervision and complete schedules are in Appendices F, F.3 and E.

3.4 Asset-oriented video post-training

Receiver adaptation changes how generated conditions are consumed while leaving the video sampling policy fixed. We therefore freeze the adapted receiver and optimize the video source using final-asset feedback. Stage 3 updates attention LoRA in both video noise experts; the video base weights, FFN LoRA, GOA, encoders, SS/Shape denoisers and readers, material modules and de-

coders are fixed. For training tuples (I, c, M^*), with reference asset M^*, the objective is

$$J(\theta) = \mathbb{E}_{(I,c,M^*)\sim\mathcal{D}_{\mathrm{tr}}}\mathbb{E}_{V\sim p_\theta,\xi}\left[R_{\mathrm{asset}}(I, F_\phi(I, V; \xi), M^*)\right]. \tag{3}$$

In Stage 3, M^* enters only the reward. Asset generation, source selection, decoding and rendering stay outside the video gradient graph. Each candidate video produces its own regional memory through the fixed receiver.

Final-asset reward. Let P, Q be area-sampled points on generated and reference surfaces in the shared canonical frame. Global F-score F_τ balances surface precision and recall at distance threshold τ. To emphasize completion beyond the input view, reference depth identifies unobserved samples Q_h. Coverage $C_h(P, Q_h)$ counts reference hidden samples near any predicted point; it measures recall and differs from the regional H-F evaluation metric. Distance-gated symmetric normal agreement N_τ rewards locally matching surface directions using absolute normal dot products. Together these terms give

$$R_{\mathrm{asset}} = w_F F_\tau + w_H C_h + w_N N_\tau - \lambda_o[C_{\mathrm{obs}} - \delta]_+, \quad w_F + w_H + w_N = 1. \tag{4}$$

The input cost $C_{\mathrm{obs}} = 0.5(1 - \mathrm{IoU}) + 0.25(1 - \cos(e_M, e_I))$ compares the observed-camera asset render with I using silhouette IoU and frozen DINO foreground descriptors. Only cost exceeding tolerance δ is penalized. The default weights are $(w_F, w_H, w_N) = (0.45, 0.35, 0.20)$, with $\lambda_o = 2$, $\delta = 0.08$ and $\tau = 0.01$ in unit-diagonal coordinates. Appendix A defines visibility, empty hidden sets and failure rewards. LPIPS is reserved for evaluation.

Candidate groups and relative advantages. The old policy draws $G = 4$ complete 49-frame videos per input, with each video forming one candidate. Each candidate has its own downstream memory and asset $M_{i,k} = F_\phi(I, V_i; \xi_k)$; their mean reward $\bar{R}_i = K^{-1}\sum_k R_{\mathrm{asset}}(I, M_{i,k}, M^*)$ yields group-standardized advantage $A_i = (\bar{R}_i - \mu_R)/(\sigma_R + 10^{-4})$, where μ_R, σ_R are the group's mean and population standard deviation. The primary setting uses $K = 1$ and a separately cost-matched control uses $K = 2$. Receiver seeds are paired within a group; failures remain in the group, and equal-score groups have zero reward advantage (Appendix C).

Video policy update. Video post-training applies single-update Flow-GRPO (Liu et al., 2025a) to 20 scored positive-variance transitions per trajectory. Group advantages, sampled states and assets are detached; current-policy transition log scores retain gradients to video attention LoRA. A normalized reference KL regularizes drift from Stage 1. Current and old parameters coincide before the sole update, so ratio clipping is inactive without suppressing the policy gradient. The exact surrogate, score normalization and sampler are in Appendices C and F.5.

Reward checks and deployment. Independent structure checks and controlled defects test whether surface rewards miss connections or duplicated parts (Appendix G). Training uses the scored stochastic sampler; normal UniPC deployment returns one unselected asset from one image and one video, without reference geometry, reward evaluation or online updates.

4 Experiments

4.1 Experimental setup and overall comparison

Data and training. Identity-disjoint Objaverse-derived splits contain 10,000/500/500 training/validation/test objects, plus 200 external GSO objects for transfer (Deitke et al., 2023; Downs et al., 2022). Each has one fixed input image; reference meshes supply targets. Wan2.2-Fun generates a 49-frame orbit on a common 512^2 canvas for the TRELLIS.2-4B 512 receiver and frozen DINOv3. S1, S2a and S2b each run 10,000 updates; S2a/b use effective batch eight and reader/denoiser rates $10^{-4}/10^{-5}$, with SS frozen before S2b. S3 runs 1,000 updates at 10^{-6} with $G = 4$ videos, $K = 1$ asset per video and one update per new group; only video attention LoRA changes. Appendix K gives full schedules.

Comparisons and metrics. We compare Hi3DGen, TRELLIS, TripoSG, Hunyuan3D-2.1, TRELLIS.2, InstantMesh, Pixal3D and V23D on paired single-image inputs (Ye et al., 2025; Xiang et al.,

2024; Li et al., 2025; Tencent Hunyuan3D Team, 2025; Xiang et al., 2025; Xu et al., 2024; Li et al., 2026; Anonymous, 2026); each retains its native generation protocol. Surface measures are unsquared Chamfer distance ($\times 100$) and F-score at 1% of reference-box diagonal (Tatarchenko et al., 2019; Mescheder et al., 2019). Reference-defined hidden F-score (H-F) measures unseen geometry; rendered-normal N-MAE and input silhouette IoU measure visible geometry (Li et al., 2026). We also report distance-gated normal consistency NC_τ (Mescheder et al., 2019), masked RGB LPIPS (Zhang et al., 2018) and failures. Each quality row uses one unselected draw per object; metric supports and evaluation transforms are defined in Appendix H.

Table 1: Single-image 3D generation on 500 test objects. F/H-F use 1%; CD is unsquared distance $\times 100$. N-MAE measures rendered-normal angular error.

Method	F↑	H-F↑	CD↓	Input IoU↑	N-MAE (°)↓	Fail (%)↓
Hi3DGen	0.718	0.586	1.689	0.916	22.6	1.4
TRELLIS	0.626	0.504	2.021	0.907	26.4	2.2
TripoSG	0.729	0.609	1.612	0.918	21.8	1.2
Hunyuan3D 2.1	0.741	0.626	1.548	0.921	20.8	1.0
TRELLIS.2	0.654	0.524	1.897	0.914	25.1	1.8
InstantMesh	0.598	0.466	2.248	0.902	29.2	2.6
Pixal3D (paper version)	0.752	0.649	1.476	0.913	19.6	1.6
V23D	0.683	0.562	1.766	0.897	23.7	2.0
AO-Flow (S2 only)	0.744	0.639	1.527	0.911	20.2	1.4
AO-Flow (full)	0.767	0.669	1.430	0.906	18.7	1.4

AO-Flow achieves the highest F-score and H-F among the compared complete systems (Table 1). Its H-F of 0.669 exceeds Pixal3D's 0.649, while the S2-only receiver reaches 0.639. Input-view fidelity remains a complementary objective: Pixal3D better preserves the silhouette, whereas AO-Flow improves complete and hidden geometry. Figure 4 compares normal-view geometry across three character inputs, with matched front/back views and local detail crops. GSO and input-corruption results are in Appendix E.



Figure 4: Selected normal-view comparison on three characters with one RGB input per group. TRELLIS.2, Pixal3D and AO-Flow show front/back geometry and local crops; the complete method matrix is in Appendix Figure 8.

4.2 Cross-stage interface analysis

We fix the video initializer and disable video RL to isolate the receiver (Table 2). Apart from Fixed-4, all configurations share the same trained SS49 producer and receive matched Shape adaptation. Global-4 selects one source set for the whole object; Region-4 retains a different set for each latent region. Shape-selected-4 instead chooses sources using its current queries. Uniform-prior reading keeps the regional IDs, total real-source mass and null mass unchanged, varying only the relative source weights. Appendix I specifies these controls.

The regional interface reaches H-F 0.639 versus 0.613 for Global-4. Reusing SS source memory improves H-F by 0.45 points over Shape-selected-4, while avoiding its repeated full-bank selection

Table 2: Interface controls before video post-training. SS49 controls share the producer; Shape consumers are separately adapted. Time and memory cover the 3D receiver.

Interface	H-F↑	Input IoU↑	Receiver (s)↓	Peak (GB)↓
Fixed-4	0.589	0.914	44.8	22.4
SS49 / Shape-no-video	0.596	0.914	85.3	26.1
Global-4	0.613	0.912	97.4	31.7
Full-49	0.645	0.906	166.2	47.8
Shape-selected-4	0.635	0.912	118.6	38.4
Region-4, uniform prior	0.630	0.911	98.1	32.0
Region-4, no spatial bias	0.624	0.908	97.8	32.0
Region-4 (full interface)	0.639	0.911	98.2	32.0

probes and reducing receiver cost. This comparison supports carrying region-specific source information from coarse structure generation into shape refinement. When aggregate retained real and null prior mass are matched, regional routing retains a 0.6-point H-F advantage over global routing (Appendix Table 12). This narrower control isolates source organization from the aggregate mass partition; it does not equate Shape's final attention posterior or injected residual.

With the source IDs, total retained mass and null mass held fixed, preserving the relative SS weights adds 0.95 H-F points over uniform priors. This supports transferring graded source priors together with source identities; Shape then recomputes content attention for its own queries. Removing spatial bias reduces H-F by 1.54 points. These controls examine cross-stage source routing, relative source weighting and spatial guidance as distinct design choices.

Dense Full-49 reading remains slightly higher at 0.645, with greater receiver time and memory. Region-4 trades a small H-F reduction for lower receiver cost. Figure 5 holds the generated 49-frame video fixed across five interface settings and compares the same rear-crown region on one character.

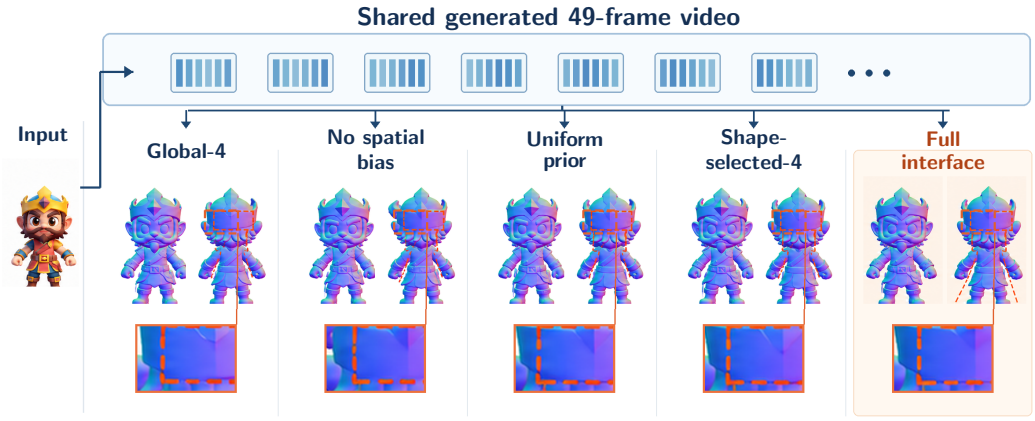

Figure 5: One-object interface comparison with a shared generated 49-frame video and no video RL. The schematic strip represents the fixed video bank; five separately adapted Shape consumers use the same SS49 producer. Aligned front/back normals and exact crops of each row's boxed rear-crown region show local crown-rim shape. Full interface corresponds to Region-4 in Table 2.

4.3 Asset-oriented video post-training

All policy comparisons freeze the same regional receiver. Continued SFT and Best-of-4 SFT test additional supervision and reward-based training-data selection. Video-reward and geometry-video RL use the same Flow-GRPO update rule, changing the feedback signal; the latter computes its reward from video geometry. Asset-DPO uses the same final-asset objective but fits preferences from a fixed offline candidate pool. Best-of-4 selection occurs only during training: deployment always returns one unselected asset.

Table 3: Video post-training with one fixed Region-4 receiver. Struct. pass applies our 100-object checklist. GPU-hours include candidate generation, scoring, selection where used, and updates; shared S1/S2 costs are excluded.

Video policy	H-F↑	Input IoU↑	RGB LPIPS↓	Struct. pass (%)↑	GPU-h↓
Initializer	0.639	0.911	0.162	66	0
Continued SFT	0.648	0.914	0.158	68	148
Best-of-4 SFT	0.656	0.912	0.161	72	426
Video-reward RL	0.650	0.919	0.151	69	268
Geometry-video RL	0.663	0.913	0.159	77	552
Asset-DPO	0.667	0.910	0.160	80	790
Asset Flow-GRPO	0.669	0.906	0.169	82	864

Final-asset feedback raises H-F from 0.639 to 0.669, a gain of 2.95 points, and achieves a structure-checklist pass rate of 82%, compared with 77% for geometry-video feedback (Table 3). Asset-DPO reaches H-F 0.667 with higher input IoU and lower incremental GPU-hours in this comparison, separating the shared asset objective from the online policy update. Structure is checked on a fixed 100-object test subset using our blinded checklist of part completeness, connections and extra parts, held separate from the training reward. The modest input-fidelity loss makes preservation an explicit trade-off. Source memories are recomputed for each policy's video. With the video fixed, generated-condition Shape continuation reaches H-F 0.647 versus 0.639 for S2; this changes the receiver, not the policy (Appendix Table 10). Matched policy views and cost details appear in Appendix Figure 9 and Appendix E.

4.4 Reward design and training dynamics

Table 4: Asset-reward ablations with fixed receiver and update budget. NC_τ is distance-gated normal consistency; Struct. pass applies our checklist to the same 100 test objects.

Reward	H-F↑	NC_τ ↑	Input IoU↑	Struct. pass (%)↑
Full reward	0.669	0.677	0.906	82
w/o hidden coverage	0.653	0.674	0.908	74
w/o normal consistency	0.665	0.650	0.907	77
w/o input penalty	0.678	0.680	0.873	76

At fixed receiver and update budget, removing hidden coverage weakens completion and removing normal agreement weakens surface consistency. Omitting the input penalty increases H-F but reduces input IoU and structural pass rate; the full reward gives the highest independent checklist pass rate (Table 4).

On a fixed 100-object validation probe, Asset Flow-GRPO raises common reward and H-F over 1,000 updates with a modest IoU decline; the paired policy trajectories are in Appendix Figure 10.

Receiver timing starts from a cached video; post-training GPU-hours include sampling, scoring and updates. Cost components and structural checks are in Appendix E.

5 Conclusion

AO-Flow uses generated video as regional evidence for single-image 3D generation. SS preserves where useful local observations arose; Shape rereads them using its own state after sequential receiver adaptation. Final-asset feedback then changes which videos are sampled by updating attention LoRA while the receiver stays fixed. The controlled comparisons connect these two effects to geometry, input preservation and cost: regional reading improves on global source selection, while dense reading remains a higher-cost quality reference, and asset-based video updates trade some input-view agreement for completion. Inference still needs one image, one video and one unselected asset. Correlated view errors, sparse-support omissions and nominal camera error delimit the present approach.

AI USE STATEMENT

Generative AI assisted with conceptual development, literature retrieval, reward formulation, mathematical exposition, experimental design, manuscript drafting, plotting code, figure and table preparation, and LaTeX assembly. Responsibility for the method, source attribution, data provenance, and empirical claims rests with the authors.

REPRODUCIBILITY STATEMENT

The companion research and figure-source packages include point-sample reward calculations, a per-object result schema, and the pipeline for rendering experimental tables, curves, and numerical prose macros. Shared result identifiers connect the comparisons and ablations to their source records. Reproduction requires object-level splits, checkpoint and configuration hashes, random seeds, sampler settings, and component timings, following Appendix B.

ETHICS AND DATA INTEGRITY

The data protocol requires appropriately licensed assets and object-level separation between training and held-out evaluation. Reference geometry is restricted to supervised training, training rewards, and offline evaluation; held-out generation and deployment use the learned models alone. Dataset-specific completion can suppress plausible shape alternatives, and several hidden geometries may be consistent with the same input image.

REFERENCES

aigc-apps. VideoX-Fun. Software repository, 2026. URL https://github.com/aigc-apps/VideoX-Fun/tree/bce42b301f59af70f543ac964792e45954c36b2b. Source snapshot at commit bce42b301f59af70f543ac964792e45954c36b2b, accessed 2026-09-23.

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. VGGRPO: Towards world-consistent video generation with 4D latent reward. *arXiv preprint arXiv:2603.26599*, 2026. URL https://arxiv.org/abs/2603.26599.

Anonymous. V23D: Leveraging video diffusion priors for high-fidelity multi-view 3D generation. Anonymized manuscript, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. URL https://arxiv.org/abs/2305.13301.

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. doi: 10.48550/arXiv.2510.23306. URL https://arxiv.org/abs/2510.23306v1. Version 1.

Wenyue Chen, Wenjue Chen, Peng Li, Qinghe Wang, Xu Jia, Heliang Zheng, Rongfei Jia, Yuan Liu, and Ronggang Wang. Know3D: Prompting 3D generation with knowledge from vision-language models. *arXiv preprint arXiv:2603.22782v2*, 2026. URL https://arxiv.org/abs/2603.22782v2.

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. URL https://arxiv.org/abs/2403.06738.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A universe of 10M+ 3D objects. *arXiv preprint arXiv:2307.05663*, 2023. URL https://arxiv.org/abs/2307.05663.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. *arXiv preprint arXiv:2204.11918*, 2022. URL https://arxiv.org/abs/2204.11918.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling geometry priors for 3D-consistent video generation. *arXiv preprint arXiv:2601.23286v4*, 2026. doi: 10.48550/arXiv.2601.23286. URL https://arxiv.org/abs/2601.23286v4.

Yue Han, Chong Li, Zhening Liu, Cong Huang, Fang Deng, Yong Liu, Fangyun Wei, and Yan Lu. Ink3D: Sculpting 3D assets with extremely complex textures via video generative models. *arXiv preprint arXiv:2607.01222v1*, 2026. doi: 10.48550/arXiv.2607.01222. URL https://arxiv.org/abs/2607.01222v1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3D: Pixel-aligned 3D generation from images. *arXiv preprint arXiv:2605.10922v1*, 2026. doi: 10.48550/arXiv.2605.10922. URL https://arxiv.org/abs/2605.10922v1.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models, 2025. URL https://arxiv.org/abs/2502.06608.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025a. URL https://arxiv.org/abs/2505.05470.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82155–82192. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2750. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.

Qingming Liu, Zhen Liu, Dinghuai Zhang, and Kui Jia. Nabla-R2D3: Effective and efficient 3D diffusion alignment with 2D rewards. *arXiv preprint arXiv:2506.15684*, 2025c. URL https://arxiv.org/abs/2506.15684.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://www.cvlibs.net/publications/Mescheder2019CVPR.pdf.

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. doi: 10.48550/arXiv.2607.24298. URL https://arxiv.org/abs/2607.24298v1. Version 1.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

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. doi: 10.48550/arXiv.2605.21121. URL https://arxiv.org/abs/2605.21121v1. Version 1.

Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3D reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1905.03678.

Tencent Hunyuan3D Team. Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material, 2025. URL https://arxiv.org/abs/2506.15442.

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. URL https://arxiv.org/abs/2403.12008.

Khiem Vuong, Deva Ramanan, and Srinivasa Narasimhan. FixAnything: 3D-consistent rendering refinement via video generative priors. *arXiv preprint arXiv:2608.23549v1*, 2026. doi: 10.48550/arXiv.2608.23549. URL https://arxiv.org/abs/2608.23549v1.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. Official model repository, 2025. URL https://github.com/Wan-Video/Wan2.2. Repository inspected September 14, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. URL https://arxiv.org/abs/2503.11651.

Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, and Lu Sheng. Ouroboros3D: Image-to-3D generation via 3D-aware recursive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21631–21641, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wen_Ouroboros3D_Image-to-3D_Generation_via_3D-aware_Recursive_Diffusion_CVPR_2025_paper.html.

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. *arXiv preprint arXiv:2412.01506*, 2024. URL https://arxiv.org/abs/2412.01506.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025. URL https://arxiv.org/abs/2512.14692.

Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E. Kaufman. Carve3D: Improving multi-view reconstruction consistency for diffusion models with RL finetuning. *arXiv preprint arXiv:2312.13980v2*, 2024. doi: 10.48550/arXiv.2312.13980. URL https://arxiv.org/abs/2312.13980v2.

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. URL https://arxiv.org/abs/2404.07191.

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 (ICCV)*, pp. 25050–25061, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ye_Hi3DGen_High-fidelity_3D_Geometry_Generation_from_Images_via_Normal_Bridging_ICCV_2025_paper.html.

Junliang Ye, Zehuan Huang, Yansong Qu, Chunshi Wang, Yunhan Yang, Yang Li, Yawei Luo, Zhuo Chen, Sheng Lu, Jun Zhu, and Chunchao Guo. UniVerse3D: Emerging properties of unified multimodal models in 3D understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 613–623, June 2026. CVF Open Access.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018. URL https://arxiv.org/abs/1801.03924.

Kaichen Zhou, Zeyang Bai, Xinhai Chang, Mengyu Wang, Paul Pu Liang, and Fangneng Zhan. Stream3D: Sequential multi-view 3D generation via evidential memory. *arXiv preprint arXiv:2605.21472v5*, 2026. doi: 10.48550/arXiv.2605.21472. URL https://arxiv.org/abs/2605.21472v5.

A METRIC AND REWARD DEFINITIONS

A.1 COORDINATE AND SAMPLING CONVENTIONS

Center and normalize each reference once, using its bounding-box center b and diagonal L, and align its selected input azimuth α by $X_A = R_z(-\alpha)(X_{\text{raw}} - b)/L$. Inputs and orbit renders use this AO frame, with the rotation applied during reference preparation. New 3D targets use $X_N = \gamma X_A$, $\gamma = 0.9$, through a fixed native-axis adapter. This margin acts on geometry while latent-channel and Gaussian-noise scales remain fixed. Convert decoded geometry back by the fixed inverse adapter and $1/\gamma$ before measuring distances in X_A; query material attributes in their original native coordinates. Image projection applies the registered crop/resize/padding transform to the camera projection of x_N/γ. Each system uses a fixed registered adapter for its axes and units, shared evaluation cameras and a single reference normalization, with output alignment kept independent of test geometry.

We follow point-sampled surface evaluation with Chamfer distance and thresholded precision, recall and F-score (Tatarchenko et al., 2019). The reward protocol uses 16,384 surface samples per mesh. Evaluation uses 100,000 independently sampled points per mesh at thresholds 0.005, 0.01, and 0.02 of the reference diagonal. Area-weighted surface sampling avoids bias toward tessellation density. The reference sample set is fixed within each candidate group, with new point samples drawn for final evaluation.

Let $d_Q(p) = \min_{q \in Q} \|p - q\|_2$ and $d_P(q) = \min_{p \in P} \|q - p\|_2$. Define

$$\text{Prec}_\tau = |P|^{-1} \sum_{p \in P} \mathbf{1}[d_Q(p) < \tau], \tag{5}$$

$$\text{Rec}_\tau = |Q|^{-1} \sum_{q \in Q} \mathbf{1}[d_P(q) < \tau], \tag{6}$$

$$F_\tau(P, Q) = \frac{2\,\text{Prec}_\tau\,\text{Rec}_\tau}{\text{Prec}_\tau + \text{Rec}_\tau + 10^{-12}}, \tag{7}$$

$$\text{CD}(P, Q) = \tfrac{1}{2}\left(|P|^{-1} \sum_p d_Q(p) + |Q|^{-1} \sum_q d_P(q)\right). \tag{8}$$

The displayed CD is $100\,\text{CD}$, without squaring distances. Comparisons use this same distance and normalization convention.

A.2 REFERENCE-ANCHORED HIDDEN-REGION EVALUATION

Render reference depth at the observed camera. Mark a reference sample visible if it projects inside the object mask and agrees with the z-buffer within a prescribed depth tolerance. Record the image resolution and tolerance. Samples outside the camera frustum are unobserved as well; distinguish this from self-occlusion in breakdowns. A fixed boundary exclusion can avoid numerical ambiguity, but must be defined before measuring results and identical across methods.

Partition Q into Q_v, Q_h. For each candidate point p, find its nearest reference point $q^*(p)$ and set $P_h = \{p : q^*(p) \in Q_h\}$. We apply F-score to this paper's reference-defined hidden region and report $\text{HF} = F_\tau(P_h, Q_h)$. If $Q_h \neq \varnothing$ but $P_h = \varnothing$, define H-F as zero without evaluating an empty precision denominator; this includes valid meshes that predict no hidden-region points. Only $Q_h = \varnothing$ makes H-F inapplicable for every method on that reference object. This reference-anchored assignment keeps the region definition fixed across methods. Predicted hidden outliers affect precision; global F-score and CD retain errors outside this region.

The training coverage term is

$$C_h(P, Q_h) = \frac{1}{|Q_h|} \sum_{q \in Q_h} \mathbf{1}\left[\min_{p \in P} \|q - p\|_2 < \tau\right]. \tag{9}$$

It counts recovered hidden reference samples against *all* predicted samples. It measures missing-surface recovery, while global F-score also accounts for excess surface. Regional H-F evaluates both precision and recall on the hidden region. The independent structure checks assess duplication, connections and part semantics. If the reference hidden set is empty, its training-reward weight is reassigned to the global F-score term.

A.3 NORMALS AND OBSERVATION COSTS

Normal consistency compares normals at nearest surface points (Mescheder et al., 2019). We add a distance gate so that only pairs within τ contribute agreement. For surface normals n_p, n_q, define

$$N_\tau = \tfrac{1}{2}\left[\frac{1}{|P|}\sum_p \mathbf{1}[d_Q(p) < \tau]|n_p^\top n_{q^*(p)}| + \frac{1}{|Q|}\sum_q \mathbf{1}[d_P(q) < \tau]|n_q^\top n_{p^*(q)}|\right]. \tag{10}$$

We denote this gated metric by $\mathrm{NC}_\tau = N_\tau$ in the tables. All normals are unit length. Pairs outside the distance gate contribute zero. Absolute agreement treats the two normal orientations equally, accommodating orientation-ambiguous open surfaces; separate oriented-normal diagnostics use valid orientation labels.

The input-preservation render follows Appendix L.2. The cost used in Eq. 4 is

$$C_{\mathrm{obs}} = \tfrac{1}{2}(1 - \mathrm{IoU}(B_M, B_I)) + \tfrac{1}{4}\left(1 - \cos(e(\mathcal{R}(M, c_0)), e(I))\right). \tag{11}$$

For the rendered training inputs, B_I is reference input alpha thresholded at .5 and used only by the reward. B_M is the generated asset's opaque coverage at input camera c_0, likewise thresholded at .5; $\mathcal{R}$ follows the common renderer contract. Empty input foreground is a data error. Backgrounds, cameras and descriptor weights are shared across candidates.

Descriptor e uses frozen DINO spatial tokens, excluding special tokens, with the foreground-area-weighted patch pooling and L2 normalization of Appendix G. Each 16×16 patch is weighted by its binary-foreground fraction. Descriptor images cross the common uint8 boundary once; LPIPS retains its float path. Map clipped cosine similarity to distance $(1 - \cos)/2$. Empty candidate foreground or finite pooled norm at most 10^{-12} receives maximal descriptor distance one; nonfinite encoder outputs are recorded as runtime errors. Report silhouette and descriptor terms separately as input-view agreement measures, alongside hidden-geometry and structural evaluation.

Masked LPIPS (Zhang et al., 2018) provides a separate evaluation of perceptual appearance distance. Its network, foreground crop, camera and lighting are fixed across methods; report its valid-output denominator.

Frozen asset status and geometry diagnostics. Every complete-asset consumer applies C8: input-aligned azimuths $0, 45, \ldots, 315°$, elevation $15°$, radius 1.8, FOV $40°$ and 512^2 resolution, with 4×4 subpixels. The input camera is included. Check all covered nearest-surface subpixels before averaging/thresholding, then freeze the status used by reward, ranking and metrics. Expanded inspection uses a common new protocol for all methods. C8 status covers the inspected nearest surfaces. Valid black or declared defaults retain their registered material status; missing support and damaged textures are material failures. Gray rendering provides a separate geometry diagnostic.

Primary F/H-F/NC_τ/O-IoU and complete-asset Structure are *asset-utility* outcomes: geometry or material generation failure receives zero, while CD and LPIPS are undefined. Failed assets receive reward -3 in the main setting and -4 when $\lambda_o = 4$, below the valid-output lower bound $-\lambda_o(1 - \delta)$. Failures remain in bounded-score denominators; only the common reference-empty hidden set makes H-F inapplicable. Table 1 pairs valid-asset CD with failure rate.

Separately, geometry F/H-F/NC_τ/O-IoU and gray Structure use the unchanged mesh for material-only failures and assign zero to geometry failures, retaining all attempted assets in their denominators. Geometry CD is conditional on valid geometry and reports its own support. These diagnostics separate geometric change from material validity while retaining complete-asset status. Report geometry, material and complete-asset failure rates separately. Infrastructure and reference-data errors follow the common retry rule and remain unresolved execution records if the retry fails. Legitimate open surfaces are evaluated using the same surface metrics.

B TRAINING CONFIGURATION AND OPTIMIZATION

B.1 CONFIGURATION

Table 5: Stage 3 video-policy optimization configuration.

Item	Setting
Video initialization	π_0^{13}: Wan2.2-Fun adapted to the registered orbit/coordinate contract
Stage 3 trainable parameters	Attention LoRA in both noise experts; FFN LoRA and GOA frozen
LoRA rank	128
Stage 3 fixed downstream	SS producer, regional memory rules, Shape consumer and separate 128-wide readers; DINO, VAEs, materials and decoders
Video / conditions	All 49 DINO patch grids for SS; regional $D = 4$ sources for Shape; actual I conditions both stages and materials
Geometry sampling	16^3 SS addresses; requested 32^3 Shape support; positive-only $g = 1$, no CFG rescale
Policy group	$G = 4$ complete independently sampled videos per input
Downstream sampling	$K = 1$ shared seed; $K = 2$ asset-call-matched ablation
Reward	$(w_F, w_H, w_N) = (0.45, 0.35, 0.20)$; $\lambda_o = 2$; $\delta = 0.08$
Surface threshold	$\tau = 0.01$ of the canonical diagonal
Policy optimization	AdamW; learning rate 10^{-6}; 1,000 updates; complete settings in Appendix K
Ratio clipping / reference	$\varepsilon = 0.2$ (inactive in one pass); normalized reference KL weight $\beta = 0.01$
Sampling steps	20-step stochastic rollout; positive-variance scored transitions
Deployment	Manifest-bound deterministic sampler; same step count for every compared policy
Model randomness	Full-run plan: three training replicas and four evaluation draws per object/replica

Supervised adaptation configuration. Stage 1 continues the inherited LoRA/GOA under the common X_A render and camera convention to obtain π_0^{13}. S2a trains SS and its reader from 49 GT views and frozen-encoder targets. After 10,000 updates at effective batch eight, freeze the complete SS producer. S2b obtains memory from its actual noise-started sampling, independently of Shape training time, and trains Shape plus its separate reader for another 10,000 updates at batch eight. GT support defines only Shape's supervised regression domain; memory comes from noise-started SS sampling. Both flows use $z_t = (1-t)z_0 + t\epsilon$ and target $\epsilon - z_0$, normalized by valid latent scalars and object. SS sampling, top-D selection and memory caches are detached during S2b.

All 4,096 SS regions remain addressable, with Shape integer coordinates mapped by $\lfloor \mathbf{c}_h/2 \rfloor$ before AO-frame conversion. Each reader has independent trainable projections. The main adaptation is flow-only; the optional GT-surface auxiliary is restricted to defined correspondences within already selected sources. Generated-condition training uses flow supervision alone. GT/generated-condition continuations update only Shape and build memory from their own conditions in separate receiver branches. Each video-policy comparison fixes its receiver throughout. Disabling the video residual retains the same adapted image-conditioned denoiser. The previous four-view ACB/OFA/Qwen receiver provides a whole-system baseline.

Reward interventions. Table 6 defines the coefficients for the main reward ablations and two supplementary preservation controls. Removing hidden coverage or normal agreement sets only that coefficient to zero and preserves the remaining weights. The comparison therefore includes changes in the balance of the reward terms. Preservation coefficient, tolerance and failure reward

remain as listed. The per-object empty-hidden rule is applied afterward: the selected configuration's w_H moves to w_F when Q_h is empty; this is distinct from the no-hidden ablation.

Table 6: Reward-intervention coefficients. A removed component is set to zero without rescaling the remaining terms. The empty-reference-hidden rule is applied afterward per object.

Setting	(w_F, w_H, w_N)	λ_o	δ	R_{fail}
Full asset reward	$(9/20, 7/20, 1/5)$	2	0.08	-3
No hidden coverage	$(9/20, 0, 1/5)$	2	0.08	-3
No normal reward	$(9/20, 7/20, 0)$	2	0.08	-3
No preservation	$(9/20, 7/20, 1/5)$	0	0.08	-3
Weaker preservation	$(9/20, 7/20, 1/5)$	0.5	0.08	-3
Stronger preservation	$(9/20, 7/20, 1/5)$	4	0.08	-4

Video transition scores in Stage 3. Let $q_\theta(z_{t-1} \mid z_t, I, c)$ be the sampled stochastic video transition, with policy-independent covariance. For video-latent dimension d_{vid}, define

$$\ell_{i,t}(\theta) = d_{\text{vid}}^{-1} \log q_\theta(z_{t-1}^{(i)} \mid z_t^{(i)}, I, c), \qquad r_{i,t} = \exp(\ell_{i,t}(\theta) - \ell_{i,t}(\theta_{\text{old}})). \tag{12}$$

The reference term in Eq. 27 is the dimension-normalized transition KL to the frozen supervised video initializer. Dimension normalization makes Eq. 12 a tempered score ratio. All RL baselines use the same one-pass schedule: the detached old scores are recomputed with unchanged current parameters, so $r = 1$ at gradient evaluation and clipping is inactive. The current score retains its derivative at $r = 1$, supplying the reward gradient. Group normalization and score tempering define the practical surrogate; the reference penalty provides soft KL regularization.

Appendix F.5 defines the time grid, Gaussian mean and variance, expert routing, and conditioning convention. Transition scores use the distributions actually sampled, including the final positive-variance step. The checkpoint adapter uses the same velocity/time convention for every scored transition, including the endpoint.

SS reads all 49 frames, and its last actual positive call supplies regional source masses and indices for Shape within the existing trajectory. The image anchor, text, camera path, coordinate adapter and endpoint conditioning are identical across controlled video policies. Each video's memory is recomputed under the same frozen receiver rules.

Execution records bind coordinates, conditioning and SS capture, recomputed scores and frozen receiver/GOA parameters. Gradient checks retain the current-policy derivative at $r = 1$.

B.2 Training loop

Algorithm 1: Stage 3 asset-oriented video RL
Input: video initializer π_0^{13}, fixed sequential-S2 receiver ϕ, training references.
For each optimization round:
1. Snapshot θ_{old}; sample training images, references, and fixed orbits.
2. Draw G stochastic video trajectories per image; save transition scores.
3. Encode all 49 frames; retain their patch-source identities.
4. Draw K downstream seeds; use every seed for every candidate video.
5. Per seed, run SS, capture memory, then Shape/materials without gradients; freeze C8 status and score GK assets.
6. Average seed scores; compute detached group advantages.
7. Recompute all scores once; accumulate gradients; take one LoRA update.
8. Log failures, decomposed rewards, policy ratio, KL, and compute costs.
Output: adapted video attention LoRA. Receiver parameters stay fixed; memory varies with video and seed.

B.3 Controlled comparisons

Continued SFT uses common objects and orbit targets; best-of-four SFT and Asset-DPO include their candidate generation/scoring costs. Video-reward and geometry-video RL control the source

objective. Within each receiver, policies share π_0^{13}, coordinates, data and samplers. Global-D/Region-D/Full-49 share an all-frame SS producer and complete evidence but adapt matched Shape consumers with 128-wide readers and $g = 1$. Global priors come from the full table; Full-49 retains that prior rule with $D = 49$.

Selection changes content and prior mass. The downward mass-matched diagnostic equalizes Global/Region retained priors while each reader computes its own final attention. Fixed-4/strong 49-to-4 use the new family; probe cost is additional, and final SS4 samples its own support. The previous four-view receiver also changes initialization, coordinates, image conditioning and guidance, so its comparison concerns the whole system.

An input-proxy best-of-G reranker operates without ground truth; reference-based best-of-G is an oracle diagnostic. Its inference budget includes all generated candidates. Cross-decoder comparisons use a common camera and scale conversion without fitting to test geometry or retraining the receiver separately for each policy. Gaussian rendering scores and mesh-quality metrics are reported separately.

C WHY DOWNSTREAM DIFFERENTIABILITY IS UNNECESSARY

Consider an all-stochastic Markov trajectory $Z = (z_T, \dots, z_0)$ with density

$$p_\theta(Z \mid I, c) = p(z_T) \prod_{t=1}^{T} q_\theta(z_{t-1} \mid z_t, I, c). \tag{13}$$

During Stage 3, assume a parameter-independent initial distribution and downstream seed law; a fixed video decoder and receiver $F_\phi(I, V; \xi)$, including SS sampling, regional memory, Shape reading and coordinate rules; an integrable reward; common policy support; and conditions allowing differentiation under the integral. Let $r(Z, \xi)$ be the final asset reward, with no explicit policy-parameter dependence after the trajectory and downstream randomness are fixed. Then

$$\nabla_\theta J = \int r(Z, \xi) \nabla_\theta p_\theta(Z) p(\xi)\, dZ\, d\xi \tag{14}$$

$$= \mathbb{E}_{Z,\xi} \left[r(Z, \xi) \sum_{t=1}^{T} \nabla_\theta \log q_\theta(z_{t-1} \mid z_t, I, c) \right]. \tag{15}$$

The gradient uses transition scores, with rendering, mesh extraction, source selection and TRELLIS.2 contained in the reward evaluation. At fixed Z, ξ, memory depends on the sampled video and the fixed receiver. This identity assumes that reward inputs are independent of policy parameters once Z, ξ are fixed; hidden states from the updating Wan would violate that condition even if detached in code. The practical group-standardized, dimension-normalized surrogate differs from this unmodified likelihood-ratio estimator; its bias depends on the normalization and score scaling. One-pass clipping is inactive. Reused trajectories with multiple updates or policy-dependent deterministic transitions require separate estimators and budgets.

For a fixed pair of videos evaluated under a common downstream seed,

$$\mathrm{Var}(R_i - R_j) = \mathrm{Var}(R_i) + \mathrm{Var}(R_j) - 2\,\mathrm{Cov}(R_i, R_j). \tag{16}$$

Positive covariance can reduce score-difference variance; negative covariance can increase it.

For a fixed pair of videos and the same K independent downstream seeds, let $D_k = R(V_i, \xi_k) - R(V_j, \xi_k)$. The paired mean has variance

$$\mathrm{Var}\left(\frac{1}{K} \sum_k D_k \right) = \frac{1}{K} \left(\sigma_i^2 + \sigma_j^2 - 2\,\mathrm{Cov}_\xi(R_i, R_j) \right). \tag{17}$$

With independent seed sets the covariance term disappears. The advantage of pairing depends on its sign and is assessed through score covariance, within-input ranking agreement, and decoder calls. Candidate-specific sparse supports can consume the shared random stream at different spatial locations.

Asset-call-matched seed averaging. With $G = 4$, increasing K from one to two while halving groups and updates from 1,000 to 500 gives 2,000 rather than 4,000 videos and the same 4,000 training assets. Each seed produces its own SS trajectory, memory and asset. Apply the preservation hinge within each reward, then average the two complete rewards before computing advantages. This matches asset calls while changing the numbers of updates, video draws and visited states; evaluation is charged separately.

D Evaluation and reproducibility details

Paired comparisons use common object identities, aggregate repeated draws within objects and resample objects; training-replica variation is separate. A fixed validation subset supports diagnostics and the pilot (Appendix G.2). Receiver training ends after 10,000 S2a and 10,000 S2b updates; online RL after 1,000 updates. Asset-DPO ends at input record 1,000 and reports effective updates $U \leq 1,000$; Shape-only continuations have a separate 1,000-update budget. Validation applies predefined safety stops; each completed run uses its fixed final checkpoint, and stopped runs retain their recorded endpoint. Input selection and paired streams follow Appendix L.1.

Reward audits compare geometry and appearance interventions under common cameras. Stochastic ranking and unselected UniPC structure checks are separate populations. The latter uses one video/asset per phase for each of 40 fixed validation objects: 80 asset attempts per initializer/final pair, with frozen C8 status and blinded geometric checks. This single-replica validation diagnostic is reported separately from the 100-object held-out test checklist and repeated-seed evaluation. Qualitative comparisons preserve inputs, cameras and each asset's own crops, including failures.

Reproduction requires object/split identities, model/configuration hashes and video, downstream, point and render seeds. Records also include regional frame indices/masses, producer and consumer identities, expert routing, reward components, frozen failure status and workload. S1 continuation, S2, S3, validation/audits, final evaluation and deployment have separate cost records.

E Additional diagnostics and cost profiles

Tables 7–9 retain external transfer, receiver-by-policy attribution and the reader-work decomposition. The previous four-view receiver is a complete system reference: its S2 and post-training H-F values are 0.574 and 0.604, respectively, under its own configuration. The main tables isolate current-interface and policy comparisons; the reward geometry remains as defined above.

Table 7: Transfer without receiver retraining. GSO uses 200 protocol-rendered inputs; blur uses 500 original objects with one-pixel foreground blur. Preservation is evaluated against the clean input.

Setting	GSO H-F	GSO O-IoU	Blur H-F	Blur O-IoU
V1.2 S2	0.532	0.890	0.490	0.848
V1.2 + asset RL	0.561	0.885	0.519	0.843
Region-4 S2	0.597	0.889	0.554	0.847
Asset-DPO	0.623	0.886	0.580	0.844
Geometry-video RL	0.620	0.891	0.576	0.849
Asset Flow-GRPO	0.626	0.884	0.583	0.842

Table 8: Stage attribution under one new video initializer and shared SS producer. Each Shape receiver is fixed during its own video-RL branch. Compare within-receiver increments before their difference.

Setting	F↑	H-F↑	Geo. H-F↑	O-IoU↑	LPIPS↓	Fail (%)
Global-4 S2	0.723	0.613	0.617	0.912	0.161	1.4
Global-4 + asset RL	0.746	0.642	0.647	0.907	0.168	1.4
Region-4 S2	0.744	0.639	0.644	0.911	0.162	1.4
Region-4 + asset RL	0.767	0.669	0.674	0.906	0.169	1.4

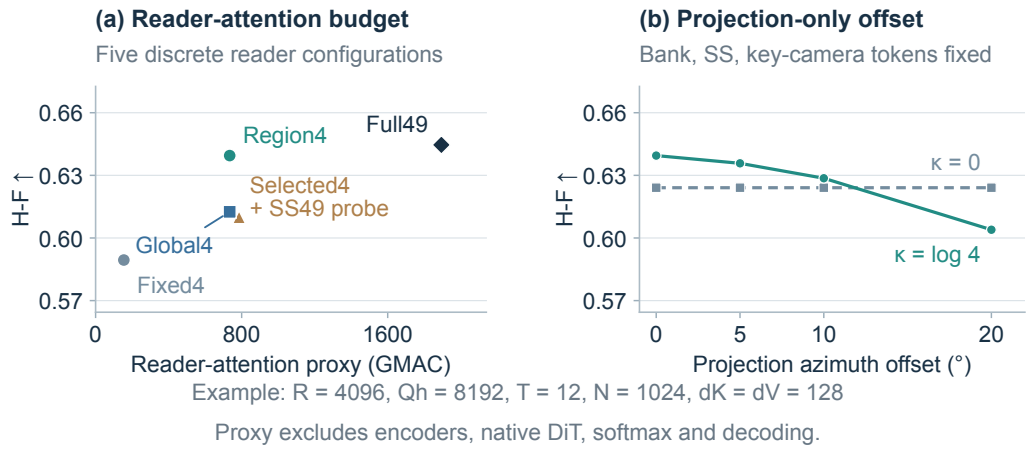

Figure 6: Reader-attention work and projection sensitivity. The proxy uses $Q_h = 8192$ and excludes the rest of the receiver. Selected-4 probes SS before global selection; it is distinct from Shape-selected-4 in Table 2. Offsets modify only the spatial-bias projection, holding video features, SS and source memory fixed.

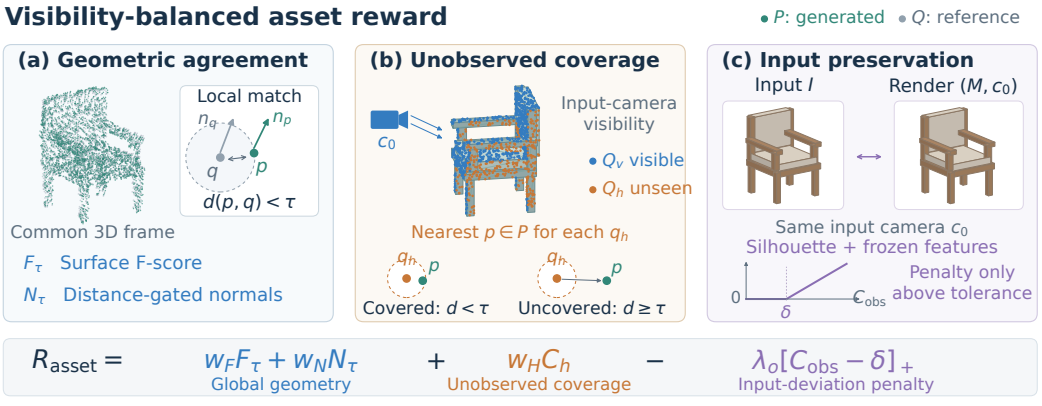

Figure 7: Geometry, reference-anchored unobserved coverage, and input preservation form the asset reward. Coverage queries all predicted points; it differs from regional H-F evaluation.

Table 9: Reader-attention work in GMAC for a common shape-query example ($Q_h = 8192$, 12 calls per stage, width 128). Probe is additional SS reading. Encoders, native DiTs, projections, backward, decoding and inspection are excluded; this is not total latency.

Interface	H-F	SS	Shape	Probe	Total
Fixed-4	0.589	51.5	103.1	0.0	154.6
Selected-4 + probe	0.610	51.5	103.1	631.4	786.0
Global-4	0.613	631.4	103.1	0.0	734.4
Region-4	0.639	631.4	103.1	0.0	734.4
Full-49	0.645	631.4	1262.7	0.0	1894.1

Computation before duration. A branch-sample equivalent counts one video latent evaluated by its selected expert at one time under one condition. With $G = 4$, $T = 20$ and two video-CFG branches, RL requires 160 rollout, 160 current and 160 reference equivalents, current backward and one update. Detached old scores need no fourth pass; reference regularization reuses current/reference evaluations. Checkpoint replay is additional. Four time/noise draws give DPO eight current/eight reference conditional equivalents and SFT four current equivalents, excluding pools and eligibility. Measured duration also depends on batching, execution and the remaining pipeline components.

Use a component ledger

$$C = C_{\text{cache}} + \sum_a \left(N_a^0 f_a^0 + N_a^\nabla f_a^\nabla + C_a^{\text{backward}} + C_a^{\text{replay}} \right) + U c_{\text{update}} + C_{\text{check/render}} + C_{\text{I/O}}.$$

Here N_a^0, N_a^∇ count no-grad/gradient-enabled forwards; f_a depends on shape, batch, expert, precision, device and offloading. Each operation has one owner; nested timings decompose rather than duplicate cost. Report records, updates, videos and asset attempts separately, including failed attempts. Pool, checking and encoding costs remain charged when an arm performs zero updates. For DPO with N prescribed pool records and $U \leq N$ accepted updates, video-forward work is $160N + 16U$: pool generation scales with N, whereas current/reference fitting scales with U. The $4N$ video/asset attempts remain at $U = 0$; encoding, inspection, scoring, backward and replay have their own actual event counts.

Receiver and memory work. For N patches/frame, $Q_s = 4096$, $D = 4$ and $d_k = d_v = 128$, the added reader work has leading term

$$T_s Q_s (49N)(d_k + d_v) + T_h Q_h (DN)(d_k + d_v),$$

with projection, native denoising, decoding and backward work accounted separately. New SS/Shape use 12 positive-only evaluations each; texture follows its own fixed sampler. Count all 49 frame encodings, the original-input encoding, memory extraction and spatial reads. Identical native I features may be shared across geometry and material consumers without duplicate ownership. The full feature bank remains resident, and evidence is read from the existing SS trajectory. Fixed-4 selection probes, when used, are additional work.

Each S2 phase has 10,000 updates at batch eight. Independent SS seeds can require 80,000 first-time S2b memory rollouts, charged despite later caching. DINO/3D target caches also incur creation costs. Identity includes producer, input, coordinates/cameras, sampler and seed; changed producers invalidate memory, and trainable Shape projections are recomputed. Optional surface labels and contract-consistent S1 continuation have separate cost accounts.

Asset checks and reporting scope. Every scored complete asset pays for C8 geometry/material checks, including offline preference/filter pools and validation. Count eight raster views and covered subpixel material queries; reused compatible buffers are not charged twice. Deployment generation-only cost is distinct from generation-plus-QA cost. Independent structural packets, annotation, pilot trials, final evaluation and S1 continuation remain explicit accounts even when excluded from a main post-training subtotal.

A single H200 with 141 GB is the profiling target; device memory fit and component durations follow actual execution records. Each method's standalone cost includes its required offline pool; a

separate project-total account may deduplicate physical reuse. Active-device hours, peak memory and parallel wall time are different quantities.

E.1 Complete qualitative system matrix

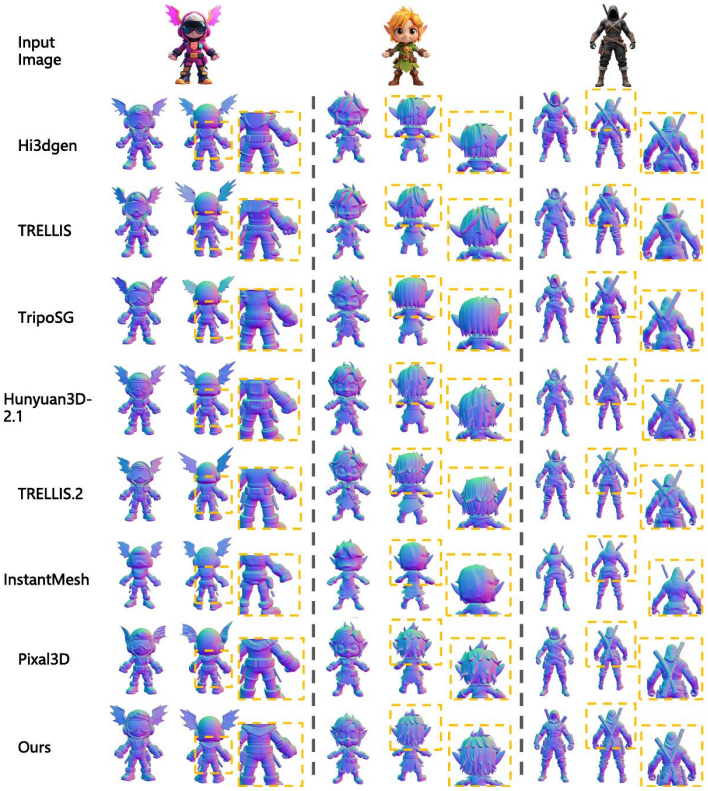

Figure 8: Complete normal-view matrix for the three character groups. Each group displays one selected RGB input, front/back geometry and the same local detail region across Hi3DGen, TRELLIS, TripoSG, Hunyuan3D 2.1, TRELLIS.2, InstantMesh, Pixal3D and AO-Flow.

F Regional interface and rollout specifications

F.1 Feature, address, and reader contract

The new receiver uses all 49 decoded RGB frames. Each is processed on the same 512^2 canvas, yielding 1,024 DINOv3 spatial patches of width 1,024. The native image sequence for actual I retains its CLS and four storage tokens; video keys contain spatial patches only. The two readers are inserted after the native image cross-attention of the last spatially indexed SS/Shape transformer block. The block binding and preprocessing are fixed before training.

The pinned SS model has resolution 16, eight latent channels, 30 blocks, hidden width 1,536, and no patchification. Its row order is $r = (16i+j)16+k$. The 512 pipeline requests 32^3 Shape support and passes those integer coordinates to Shape without an axis permutation. For integer Shape support coordinates $\mathbf{c}_h$, define parent addresses componentwise by $\lfloor \mathbf{c}_h/2 \rfloor$ and continuous cell centers by $(\mathbf{c}_h + 1/2)/32 - 1/2$. Reader positions use native unit-grid centers $u_s = (i + 1/2)/16 - 1/2$ and $u_h = (\mathbf{c}_h + 1/2)/32 - 1/2$ in $\Phi(2u)$. Projection first applies the fixed adapter $x_N = \mathcal{A}(u_h)$; address lookup never applies that adapter. These are fixed addresses, not surface points or exclusive decoder receptive fields. Full self-attention and the decoder mix spatial information. Memory retains every SS address, including regions without predicted occupancy.

For stage $a \in \{s, h\}$, all vectors are rows and the reader width is $d = 128$:

$$q_x^a = \mathrm{LN}_0(h_x)W_Q^a + \Phi(2u_x)U_x^a, \tag{18}$$

$$k_{jp}^a = \mathrm{LN}_0(D_{jp})W_K^a + \Phi(\widetilde{u}_p)U_p^a + \omega_j U_c^a, \tag{19}$$

$$v_{jp}^a = \mathrm{LN}_0(D_{jp})W_V^a, \qquad \ell_{x\perp}^a = q_x^a w_\perp^a + b_\perp^a. \tag{20}$$

LN_0 has no affine parameters and epsilon 10^{-5}. Φ concatenates sine/cosine features at frequencies $2^k\pi$, $k = 0, \ldots, 3$, independently for each coordinate. Patch centers $\widetilde{u}_p$ are normalized on the actual processed canvas. Camera features are $\omega_j = [\sin a_j, \cos a_j, \sin e, \cos e, \rho]$. Time is already represented in the native hidden state. W_Q^a maps native hidden width to 128; W_K^a, W_V^a map 1,024 to 128. Output W_O^a maps 128 back to native hidden width. SS and Shape share no trainable reader tensors. DINO features alone may be cached jointly.

Projection matrices have no bias and use Xavier initialization, except zero output projections; null weights/bias initialize to zero. Native denoisers start from the same public 512 checkpoint within matched controls. No Qwen, ACB, OFA or extra semantic gate is active in this geometry path. The legacy four-view receiver remains a separately specified system.

F.2 Mass-preserving source retention and spatial reading

For SS region r, let N_j be the number of valid patches in frame j and $\ell_{rjp}^s = q_r^s(k_{jp}^s)^T/\sqrt{128}$. Its joint patch/null attention is

$$A_{rjp}^s = \frac{(w_j/N_j)\exp(\ell_{rjp}^s)}{\exp(\ell_{r\perp}^s) + \sum_{b,p'}(w_b/N_b)\exp(\ell_{rbp'}^s)}. \tag{21}$$

The source priors use $w_0 = w_{48} = 1/96$ and $w_j = 1/48$ otherwise, splitting one nominal angle across the duplicated orbit endpoints. SS uses Eq. 21, with all-source weights summing to one before masking. Structural masks exclude padding, not predicted occlusion. Empty sources are omitted; remaining weights are not renormalized. For valid keys, the log-prior is $\log(w_j/N_j)$; zero weights are masked rather than logged. Null has zero value. At neutral logits and complete sources, its mass is $1/2$ regardless of the number of patches. This describes initialization; null mass is an attention allocation rather than a calibrated confidence.

Each SS denoising call performs a full-source read, but only the last actual positive model call contributes to memory. Under the 12-step schedule this is the call before the last update, not an additional $t = 0$ forward. The memory uses this single positive call, with $e_{rj} = \sum_p A_{rjp}^s$ computed for all regions and all frames. Region-D retains the highest positive masses, with ties broken by ascending original frame index; slots without an eligible source are padded and have zero mass. The original null and all unselected masses are summed directly into $m_{r\perp}$, avoiding subtraction round-off. Selected masses retain their original scale. Endpoints 0 and 48 share one nominal azimuth's prior weight and can both be selected, including correlated duplicate content.

For retained frame j, patch center u_p and nominal projection $\pi_j(x)$, Shape uses

$$\begin{aligned} b_{xjp} &= -\kappa \min\left(\frac{\|u_p - \pi_j(x)\|^2}{2\sigma^2}, 1\right), \\ A_{xjp}^h &= \frac{(e_{rj}/N_j)\exp(\ell_{xjp}^h + b_{xjp})}{m_{r\perp}\exp(\ell_{x\perp}^h) + \sum_{b\in J_r,p'}(e_{rb}/N_b)\exp(\ell_{xbp'}^h + b_{xbp'})}. \end{aligned} \tag{22}$$

Shape uses Eq. 22, with the full retained frames rather than a hard patch crop. Its pixel-space bias lies in $[-\log 4, 0]$, with $\sigma = 32$ pixels. Behind-camera or out-of-canvas projections set that bias to zero, without inferring visibility. If a retained source becomes invalid, its prior mass is transferred to null. Empty candidates give all-null. The bias also changes allocation to null. At fixed query/content/null logits, write $Z_0 = \sum_{j,p}(e_{rj}/N_j)\exp(\ell_{xjp}^h)$, $Z_b = \sum_{j,p}(e_{rj}/N_j)\exp(\ell_{xjp}^h + b_{xjp})$ and $n = m_{r\perp}\exp(\ell_{x\perp}^h)$. Since $Z_0/4 \le Z_b \le Z_0$, the real-source masses $p_0 = Z_0/(n + Z_0)$ and $p_b = Z_b/(n + Z_b)$ obey $p_0/(4 - 3p_0) \le p_b \le p_0$. With neutral logits and retained prior $1/24$, the lower endpoint is $1/93$. Thus the ordinary bias ablation measures its combined spatial and allocation effects; Appendix F.6 defines a separate partition-mass diagnostic. The bound compares

identical current logits, not separately learned consumers, and smaller probability mass need not imply a smaller value-vector norm. For either reader,

$$y_x^a = \sum_{j,p} A_{xjp}^a v_{jp}^a, \quad \Delta h_x^a = y_x^a W_O^a, \quad h_x' = h_x + \tfrac{1}{2}\Delta h_x^a \min\left(1, \frac{\|h_x\|_2}{\|\Delta h_x^a\|_2 + 10^{-6}}\right). \tag{23}$$

There is no separate gate. The residual is zero for $h = 0$, zero output projection or all-null. The cap bounds the hidden-feature residual; geometry and input preservation are evaluated separately. Content logits further modify the stored allocation, so retained prior mass is not a reliability upper bound.

A neutral SS row gives each interior frame mass $1/96$. Retaining four such frames leaves $1/24$ real prior and $23/24$ null in Shape before content/spatial logits. Consequently a change in D changes both accessible sources and prior odds. The same-rule D comparison measures the complete trade-off between source access and prior allocation.

F.3 Optional surface supervision

The primary method uses flow supervision only. The optional auxiliary is built from GT rendering buffers in the same processed pixel frame. For a reference Shape support cell, use only samples whose local surface identity and projected visibility can be established. Project those samples into the candidate frames already chosen by frozen SS; GT does not replace the selected indices. Each sample divides unit target mass equally among its visible corresponding patches. If every candidate is reliably occluded or out of view, assign shared null. Single-view occlusion alone is not an all-source null event.

Depth discontinuities, invalid buffers, ambiguous foreground boundaries and unresolved thin-sheet sides are excluded, not converted into confident negatives. A cell center is not itself a GT surface sample. Record attempted/retained samples and coverage for thin structures; empty target sets receive no auxiliary term. Generated conditions do not inherit pixel targets from a paired reference asset. With distributions Y_x and A_x^h including null, the optional loss is

$$\mathcal{L}_{\text{surf}} = \frac{1}{|\mathcal{B}_+|} \sum_{o \in \mathcal{B}_+} \frac{1}{|\Omega_o|} \sum_{x \in \Omega_o} \tfrac{1}{2}\|A_x^h - Y_x\|_2^2. \tag{24}$$

It is zero for an empty effective batch. Each squared-distribution term is at most one; this bounds the loss value, not its gradient contribution. The default coefficient is zero; the separate auxiliary arm uses 0.05. This auxiliary uses a Brier-style distribution loss; its geometric context is discussed in Appendix J. Gradients reach Shape queries, keys, null and their hidden-state inputs, while the SS producer and stored memory remain detached.

F.4 Sequential optimization and memory provenance

For $a \in \{s, h\}$ and condition $\mathcal{C}^a$, the normalized flow-matching loss is

$$\mathcal{L}_a = \mathbb{E}\left[\|v_{\phi_a}(z_t^a, t; \mathcal{C}^a) - (\epsilon - z_0^a)\|^2 / d_a\right], \qquad z_t^a = (1-t)z_0^a + t\epsilon, \tag{25}$$

where d_a counts valid latent scalars, $t \sim U(0,1)$ and ϵ is Gaussian noise; these are sampled independently by object and stage. S2a performs 10,000 SS/reader updates at effective batch eight; S2b separately performs 10,000 Shape/reader updates. At each phase's start, initialize fresh AdamW moments and a 500-update linear warmup to reader/denoiser rates $10^{-4}/10^{-5}$. Other optimizer settings follow Appendix K. Each object has independent uniform flow time and Gaussian noise, normalized by valid latent scalars; the two phases do not average a joint SS/Shape loss. Eight one-object microbatches contribute $1/8$ each, followed by one global gradient clip and optimizer step. Optional auxiliary terms are additionally normalized by the number of eligible objects, not by patch count over the whole batch.

After S2a freeze the complete SS denoiser, reader, coordinates, sampler and memory rule. For each S2b record, run this producer from independent Gaussian noise on actual I and the record's condition video. Neither GT support, GT SS latent nor reference depth enters this call. Its full-grid memory then conditions Shape regression at an independently sampled training time. GT Shape

support supplies supervised coordinates only; deployment uses predicted support. A finite memory from an empty predicted support can still be read on GT training support, but the SS failure is retained and inference never uses GT to repair it.

A cache key includes processed input/video bytes, DINO/processor identity, SS checkpoint, coordinate/camera contract, sampler/guidance, extraction block/call, selection rule and SS seed. Same-key records may reuse memory; new seeds/conditions require new production. Updating Shape never changes the producer, and its changing K/V projections are recomputed rather than cached as fixed features. The 80,000 object records may therefore require up to 80,000 first SS rollouts; this is charged separately from the 20,000 optimizer updates. No gradient crosses sampling, top-D or a cache. GT/generated-condition continuations train Shape only and regenerate memory from their respective conditions (Appendix G.3).

S3 retains one input group per update, four complete videos and one asset per video, with 1,000 groups per replica. The reference policy is the complete contract-consistent S1 initializer, including attention/FFN LoRA, GOA and adapter scales. Only attention LoRA in both experts updates; FFN/GOA and the entire receiver remain fixed. Memory is recomputed for each new video without changing its producer parameters. Scores retain gradients, while assets, rewards, advantages and trajectory states are detached. Controlled arms omit inseparable Lightning modifications.

Relation to the clipped surrogate. For $r_{i,n} = \exp(\ell_{i,n}(\theta) - \ell_{i,n}(\theta_{\text{old}}))$, the general normalized-score surrogate is

$$\mathcal{L}_{\text{clip}} = -\frac{1}{GT}\sum_{i,n} \min\{r_{i,n}A_i, \text{clip}(r_{i,n}, 1-\varepsilon, 1+\varepsilon)A_i\} + \beta\mathcal{L}_{\text{ref}}, \qquad \varepsilon = 0.2. \tag{26}$$

With $\ell_{i,n}(\theta) = d_{\text{vid}}^{-1} \log q_\theta(z_{n+1}^{(i)} \mid z_n^{(i)}, I, c)$ for the actual sampled transition distribution, the one-pass gradient is

$$\nabla_\theta \mathcal{L}|_{\theta_{\text{old}}} = \left[-\frac{1}{GT}\sum_{i,n} A_i \nabla_\theta \ell_{i,n}(\theta) + \beta \nabla_\theta \mathcal{L}_{\text{ref}}(\theta) \right]_{\theta=\theta_{\text{old}}}. \tag{27}$$

Before the sole optimizer step, $\theta = \theta_{\text{old}}$ and each ratio is one. Both branches then have the same local derivative, yielding Eq. 27 with detached advantages, old scores and trajectory states. Thus clipping is inactive in this one-pass update, and reference regularization is a soft penalty. Centering advantages can make the surrogate value near zero while retaining a nonzero gradient. The update budget uses one pass per trajectory group.

F.5 Positive-variance rollout and scoring

We specify a discrete Gaussian rollout following the reverse-SDE construction of Flow-GRPO (Liu et al., 2025a). Let $n = 0, \ldots, T$, $T = 20$, $u_n = 1 - n/T$, $\tau_n = 5u_n/(1+4u_n)$ and $h_n = \tau_n - \tau_{n+1}$. At the scored steps $n < T$, define

$$s_n^2 = 0.1^2 \frac{\tau_n}{\max(1-\tau_n, 1-\tau_1)}, \qquad b_n^2 = s_n^2 h_n, \tag{28}$$

$$\mu_{\theta,n}(x) = x - h_n \left[f_\theta(x, \tau_n) + \frac{s_n^2}{2\tau_n}\{x + (1-\tau_n) f_\theta(x, \tau_n)\} \right], \tag{29}$$

$$q_{\theta,n}(x_{n+1} \mid x_n, I, c) = \mathcal{N}(\mu_{\theta,n}(x_n), b_n^2 I_{d_{\text{vid}}}). \tag{30}$$

The initial state is parameter-independent standard Gaussian. The first-step cap avoids division by zero. Every transition, including the last into $\tau_T = 0$, has positive variance and is saved and scored; there is no subsequent policy-dependent deterministic denoise. The terminal latent is decoded by the fixed video VAE. This defines the discrete stochastic policy used for training; the deployment sampler is specified separately below.

For RL rollout and transition scoring, f_θ uses the data-to-noise convention $\epsilon - z_0$, with CFG $f_{\text{neg}} + 5(f_{\text{cond}} - f_{\text{neg}})$. The default adapter uses FP32 native time $t_{\text{native}} = 1,000\tau$ without rounding or a second shift, selecting the high expert at $t_{\text{native}} \geq 900$ and the low expert otherwise. Rollout, current, old and reference share this routing. GOA remains frozen. Conditions are fixed side inputs;

the profile samples all $16 \times 13 \times 64 \times 64 = 851,968$ latent coordinates, including the first temporal slice. Batch, CFG copies, side conditions and internal sequence padding are excluded from d_{vid}. This density requires GOA and conditioning to preserve all sampled coordinates. If either overwrites coordinates, the density must instead be defined on the remaining free coordinates. The adapter also uses the stated output convention at the boundary call $t_{\mathrm{native}} = 1,000$. This guided velocity differs from Asset-DPO's conditional regression prediction (Appendix G.2).

Both recorded states are detached in score recomputation. The normalized score used in Eq. 27 is

$$\ell_{i,n}(\theta) = -\frac{1}{2d_{\mathrm{vid}}}\left[\frac{\|x_{i,n+1} - \mu_{\theta,n}(x_{i,n})\|_2^2}{b_n^2} + d_{\mathrm{vid}}\log(2\pi b_n^2)\right]. \tag{31}$$

At these old-policy states, the normalized transition reference KL is exactly

$$\mathcal{L}_{\mathrm{ref}} = \frac{1}{GT}\sum_{i,n}\frac{\|\mu_{\theta,n}(x_{i,n}) - \mu_{\theta_0,n}(x_{i,n})\|_2^2}{2d_{\mathrm{vid}}b_n^2}. \tag{32}$$

The density uses $d_{\mathrm{vid}}\ell$, while the practical ratio uses $\exp(\Delta\ell)$ and is tempered. The default one-pass gradient is Eq. 27; clipping is inactive, while group standardization and score normalization remain surrogate choices. The score identity in Appendix C applies to this stochastic policy. Deployment uses the specified 20-step Fun UniPC solver, whose integer model timesteps and initial sigma differ from the rollout (Appendix K.2). Both samplers evaluate the same trained checkpoints, with paired asset metrics and failures reported as separate populations.

F.6 Matched interface controls

Frame access and selection. Fixed-4 reads indices 0/12/24/36 through the new reader family. Selected-4 uses a frozen full-orbit SS probe and fixed selection rule to choose four sources, then runs its independently adapted four-frame SS; the probe's latent/support/memory is discarded. All 49 encoder inputs, the probe and its training dependency are included in standalone cost. This control adapts receiver-aware selection to the new reader family; Stream3D itself uses velocity fusion. A DINO-only coverage selector provides a cheaper supplementary reference.

Global-D, Region-D and Full-49 instead share one trained all49 SS producer. At the S2 initializer comparison they share videos, SS noise, predicted supports and the untruncated evidence table. Their Shape consumers are independently trained with matching architecture/updates. The existing Global-D gives each of all 4,096 fixed-grid rows one vote per retained frame ID, including rows unused by predicted Shape support. It breaks total-vote ties by frame index and rebuilds each region's priors from the *full* evidence table at the selected global indices. A locally discarded frame is not assigned artificial zero mass. Full-49 is D=49 with the same evidence priors and original null, not a different uniform-attention model. The main D comparison uses the same spatial bias; setting $\kappa = 0$ is a separate positional ablation.

Predicted-support-aware global control. A support-aware global selector uses the same sampled 32^3 support $\widehat{S}$:

$$n_r = \left|\{c \in \widehat{S} : \mathrm{parent}(c) = r\}\right|, \qquad S_j = \sum_r n_r e_{rj}, \qquad J^{\mathrm{supp}} = \mathrm{Top4}_{j:S_j>0}(S_j). \tag{33}$$

Count unique predicted cells, not GT support or merely nonempty regions; break ties by original frame index. Shape gathers each region's original e_{rj} at these common IDs, not S_j, and moves discarded mass to null without renormalization. Its native initialization, independent 10,000-update/B8 adaptation, $\eta = 0$, $g = 1$, $\kappa = \log 4$ and $\sigma = 32$ match Region-D. Training and evaluation both select from predicted support; GT support only defines Shape's training loss domain. Empty predicted support gives $n_r = 0$, so this new selector returns empty candidates/all-null. The existing ownership-vote Global-D remains unchanged: all 4,096 rows still vote during S2b, even if predicted support is empty, and its memory can condition GT-domain training. At deployment, empty SS support is a geometry failure for every arm: skip Shape without GT repair. This reuses SS and adds only support aggregation/selection and separately charged Shape training. Its support-weighted evidence proxy provides a second global selection rule for comparison with the ownership-vote heuristic.

Independently trained Shape-no-video control. To test the net addition of Shape video conditioning, keep the common SS49 producer and train another Shape backbone from the same native initialization, using matched S2b objects, GT targets/support, time/noise schedule, 10,000 updates/B8, $\eta = 0$ and $g = 1$. Actual I remains the native image condition. Hard-bypass the entire Shape video reader during both training and evaluation, excluding the bank, indices, masses, null and projection bias. The reader is omitted or held inactive outside the optimizer. Independent adaptation measures the full Shape-video path, including its reader capacity.

Its GT-support Shape loss consumes no SS memory, so no per-record memory rollout is required for this control; shared producer records used for pairing/logging retain their actual cost owner. Evaluation still uses the same video, SS seed and predicted support, followed by the fixed material path and C8 checks, with zero Shape-reader work. No extra S3 arm is required. The system retains video information through SS support. Its comparison with Region-D measures the additional effect of Shape video conditioning.

Retained-mass diagnostic. Let a_r^G, a_r^R be the total retained real masses for Global/Region choices. In the paired mass-matched diagnostic, scale each side's retained priors down to $\min(a_r^G, a_r^R)$ and transfer the removed mass to null. Both sides are all-null when the common mass is zero. No weak prior is amplified. Consumers are independently adapted under this same rule, allowing source-content comparisons at matched real *prior* mass. Final attention mass and residual norm can still differ. Region top-D maximizes the retained sum within a common eligible source universe; therefore an unmatched Global/Region comparison inherently changes both access and prior odds.

Region–source association control. Keep each target region's ordered slot masses and null fixed, but replace its ordered source-ID tuple with one from another region of the same video. Stratify all grid rows by their number of valid slots. Within each nonempty stratum containing at least two rows, sort region IDs by a hash of the preregistered record key and region ID, then cyclically shift the donor assignment by one. This is a bijection with no self-donor in that stratum. Empty rows and singleton strata remain unchanged; record both donor changes and actual source-set changes, since different regions may retain identical tuples. Training and evaluation use the same algorithm with distinct record-derived seeds fixed independently of outputs.

Only donor IDs move: the target's ranked masses retain their original $e_{r,j}$ values, and null remains fixed. Contents, camera tokens and projection metadata follow the reassigned frame IDs in the unchanged bank. These synthetic ID–prior pairs disrupt regional association. A matched Shape consumer is independently adapted and evaluated under this rule; switching the ordinary consumer only at test time instead measures dependence under distribution shift. Index reassignment adds no SS trajectory or DINO encoding; its indexing work and additional matched Shape training are counted.

Real/null partition-mass control. For a location-specific interpretation, use an additional diagnostic reader that preserves the no-bias real/null mass at the same current logits. With $L_{jp} = \log(e_{rj}/N_j) + \ell^h_{xjp}$ over valid positive-prior keys, define

$$A^{\mathrm{pm}}_{xjp} = p_0 \frac{\exp(L_{jp} + b_{xjp})}{\sum_{b,p'} \exp(L_{bp'} + b_{xbp'})}, \qquad A^{\mathrm{pm}}_{x\perp} = 1 - p_0, \tag{34}$$

where $p_0 = \sum_{j,p} \exp(L_{jp}) / [n + \sum_{j,p} \exp(L_{jp})]$. Empty real sets give all-null. Compute partitions with stable log-sum-exp and retain their gradients during Shape training, including that of p_0. Equivalently, subtract $\log(Z_b/Z_0)$ from all real logits. This content-weighted correction can be positive; the diagnostic therefore differs from the original nonpositive bias. Constant spatial bias cancels, while relative spatial reweighting remains. Independently matched Shape adaptation uses this rule throughout training and evaluation. It does not equalize independently learned trajectories, value directions, readout norms or capped residual norms.

For these attribution controls, record per-query retained prior mass, actual post-logit real mass, pre-projection readout norm, pre-cap residual norm and final injected norm, with object-level aggregation. The association control also records donor IDs, unchanged/changed tuple counts and seed identity. These records distinguish prior allocation from learned readout strength. Source-content

conclusions use the matched-prior control, and location-specific conclusions use the partition-mass control; the unadjusted comparisons measure complete configurations. The no-video control measures the net Shape-video effect, while the support-aware global control tests sensitivity to the global selection rule.

The auxiliary arm is separate from the flow-only main comparison. If it materially changes results, train a control with the same auxiliary but the Shape video residual disabled, to test query regularization without value transfer. This auxiliary comparison complements the flow-only Shape-no-video arm above. Global-D and Region-D each have their own frozen-receiver S3 arm, so their S2-to-S3 increments form the main stage-attribution comparison. The V1.2 ACB/OFA system remains a complete-system reference with its own initialization and guidance.

Content interventions. Appendix G.5 specifies the separate natural-cue token-ablation and controlled RGB-omission populations. The former edits encoded features; the latter reruns DINO on changed RGB. Full-pipeline interventions recompute SS and memory. Fixed-producer interventions preserve support, source IDs, masses/null, metadata and the saved Shape-entry RNG state, changing only Shape-source content. The two interventions measure full-pipeline sensitivity and Shape content dependence, respectively. Natural-cue and synthetic-omission results remain separate because the synthetic population does not estimate natural error frequency.

G Control objectives and evaluation protocol

G.1 Reference-assisted video controls

All primary video-optimization arms share the orbit initializer, input sequence, LoRA scope, stochastic sampler configuration and fixed adapted receiver. The 2D video reward uses the same training reference asset rendered at the four prescribed reward cameras (video indices 0/12/24/36). Let q_B be foreground IoU, $q_E = (1 + \cos(e(x), e(y)))/2$ the frozen DINO similarity, and d_L the frozen AlexNet LPIPS distance (Zhang et al., 2018) on the union-masked, white-background images. Then

$$R_{2D} = \frac{1}{4}\sum_{j=1}^{4}\left[0.25q_{B,j} + 0.35q_{E,j} + 0.40\exp(-d_{L,j}/0.25)\right]. \tag{35}$$

Generated and reference frames use their original registered positions on a fixed 512-pixel canvas. Reference alpha supplies the reference mask; the candidate foreground extractor is fixed RMBG-2.0[1], evaluated on a 1024-pixel ImageNet-normalized canvas, with its sigmoid output bilinearly resized and thresholded at 0.5. Empty candidate foreground gives zero for that frame. DINO excludes CLS/register tokens and uses foreground-area-weighted pooling followed by L2 normalization. Training LPIPS uses the scalar v0.1 AlexNet output on union-masked images; all scorers are frozen. The resulting control measures reference-assisted appearance agreement.

Our geometry-video control adapts VGGRPO's geometric-reward motivation (An et al., 2026) to RGB videos. Frozen VGGT-1B (Wang et al., 2025) estimates cameras and depths from frames $0, 4, \ldots, 48$, while the asset receiver consumes all 49 frames through its fixed regional interface. Geometry is scored on a 128-pixel grid. With depth-normalized camera centers c_i, world-frame rotations A_i, increments $v_i = c_{i+1} - c_i$ and $\omega_i = \mathrm{vee}\log(A_{i+1}A_i^\top)$, define

$$e_v = \frac{1}{11}\sum_{i=1}^{11}\frac{\|v_i - v_{i-1}\|}{\|v_i\| + \|v_{i-1}\| + 10^{-6}}, \quad e_\omega = \frac{1}{11}\sum_{i=1}^{11}\frac{\|\omega_i - \omega_{i-1}\|}{\|\omega_i\| + \|\omega_{i-1}\| + 10^{-6}}. \tag{36}$$

The motion score is $r_m = \frac{1}{2}[(1 + e_v)^{-1} + (1 + e_\omega)^{-1}]$. It is zero when total estimated rotation lies outside $[\pi, 3\pi]$ or normalized translation length is below 0.05, preventing a static sequence from receiving a perfect orbit-motion score.

For each target frame, predicted depths from the other frames are reprojected with a 3-by-3 nearest-depth splat; the target itself is excluded, as is the duplicate endpoint when either closed-orbit endpoint is the target. At each foreground pixel, use clipped relative depth error $\min(|\hat{D} - D|/(D +$

[1] https://huggingface.co/briaai/RMBG-2.0

$10^{-6}\bar{d}), 1)$, assigning error one to missing coverage. Let E_i be its frame mean; foregrounds below 64 pixels receive $E_i = 1$. Set $r_g = 1 - \mathrm{mean}(\mathrm{Top3}(E))$ and

$$R_{GV} = 0.50R_{2D} + 0.25r_m + 0.25r_g. \tag{37}$$

This control receives reference 2D supervision plus video-geometry self-consistency; AO-Flow additionally scores the final asset against reference surfaces. The comparison therefore evaluates these complete objectives and their respective supervision. Invalid geometry predictions remain in the candidate pool with zero geometry terms. Infrastructure failures are retried on the same saved inputs and remain unresolved errors if scoring still fails.

Continued SFT uses the same 1,000 input records, four independent time/noise targets per update, and one optimizer step per record. Best-of-four SFT scores four initializer videos per record with the asset reward and selects a highest-reward encodable video using the fixed uniform tie rule. Frozen-VAE CFM training skips the entire update when no winner is encodable; retain the prescribed inputs and seeds and record all eligibility attempts and the actual update count. This is unthresholded reward-ranked imitation: if all assets fail but a video is encodable, a failed candidate remains eligible for imitation. The 1,000-update account assumes an encodable winner in every record. Candidate preparation and scoring are charged separately. Training uses fixed record endpoints after the separately budgeted validation pilot; compute accounts include the work specific to each objective.

G.2 Same-asset offline preference control

Asset-DPO uses the same saved initializer pool as best-of-four SFT: 1,000 input records, four complete videos and four assets per record, with the same fixed regional receiver ϕ^{13}, reward and downstream seed. The largest-reward encodable video is the preferred sample; its tie rule is identical to best-of-four SFT. The smallest-reward encodable video is rejected, with a separate fixed uniform tie seed. Only strict FP32 reward differences define preferences. Equal-score groups and groups with fewer than two encodable videos skip the optimizer step, including weight decay. Valid videos whose assets fail remain eligible rejected samples with the prescribed failure reward. Infrastructure errors follow the shared retry rule and are recorded separately from preferences.

Following the constant-coefficient flow preference objective (Liu et al., 2025b), we use the complete supervised video initializer θ_0 as the frozen reference, including its Stage 1 LoRA and GOA. Geometry-derived video preferences provide a related precedent (Du et al., 2026); here preferences use final-asset rewards. For $a \in \{w, l\}$, encode the complete preferred/rejected video as z_a. Each pair uses four independent draws (t_k, ϵ_k), shared between both videos and both models, with $t_k \sim \mathcal{U}(0, 1)$ and $\epsilon_k \sim \mathcal{N}(0, I)$:

$$z^a_{t_k} = (1 - t_k)z_a + t_k\epsilon_k, \qquad u^a_k = \epsilon_k - z_a, \tag{38}$$

$$E^a_\theta(k) = d^{-1}_{\mathrm{vid}} \| M \odot [v_\theta(z^a_{t_k}, t_k; I, c) - u^a_k] \|^2_2, \tag{39}$$

$$\Delta^a_k = E^a_\theta(k) - E^a_{\theta_0}(k), \tag{40}$$

$$\mathcal{L}_{\mathrm{ADPO}} = \frac{1}{4} \sum_{k=1}^{4} \log\big(1 + \exp\{\beta_D(\Delta^w_k - \Delta^l_k)\}\big) . \tag{41}$$

M is the common binary mask of predicted latent coordinates and $d_{\mathrm{vid}} = \sum M$. For this control, v_θ denotes the native conditional velocity prediction without a classifier-free guidance combination; this training definition is used for winner/loser and current/reference alike. CFG-5 is used for rollout (Appendix F.5); preference fitting uses the conditional velocity directly. Native time mapping, expert routing and latent scaling are identical across the four branches. Reference errors are detached; both current-model errors retain gradients. The half coefficient is absorbed into β_D, and the per-coordinate normalization is explicit. The objective is a flow preference surrogate on reference-relative regression errors. Averaging losses after the nonlinearity differs from applying one sigmoid to an averaged margin.

The default $\beta_D = 500$ uses FP32 energies and a numerically stable softplus. The three-trial selection rule and preservation constraints are fixed below (Appendix G.2). This loss supplies the complete preference update, with equal weight per eligible pair. The final checkpoint follows input record 1,000; report the actual number $U \le 1,000$ of optimizer updates. Skipped groups retain their positions in the fixed record list. Four paired draws require eight current and eight reference video

forwards per valid update, unlike winner-only SFT. All pool, reference, optimization and evaluation costs are charged separately.

Evaluation generates one unselected video and asset per input. Comparing trained Asset-DPO and online-RL policies assesses the effect of their optimization procedures under the common asset objective; preference-fitting diagnostics establish the scope of that comparison.

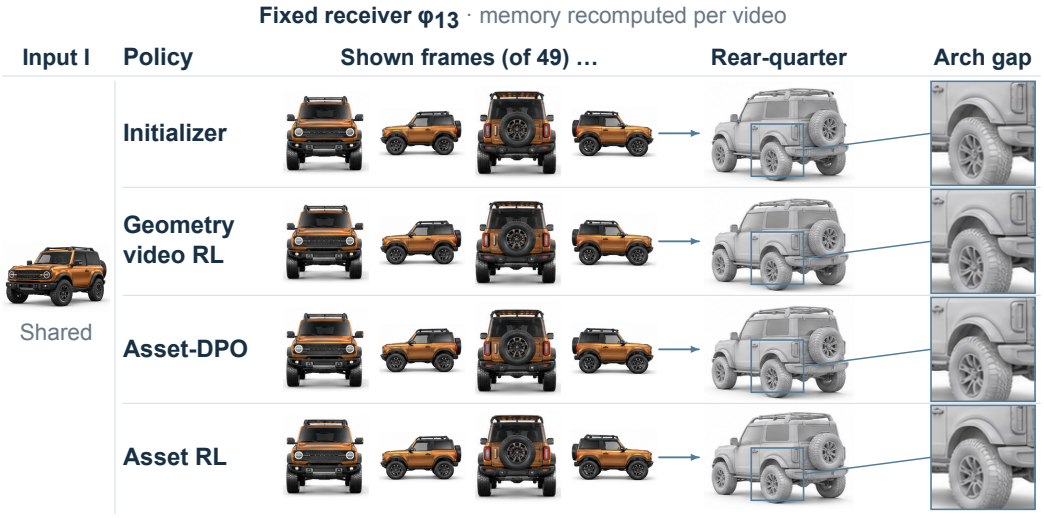

Figure 9: Video-policy qualitative pairing with one shared input and a fixed regional receiver. Each policy supplies its own 49-frame video, recomputes regional memory and yields one unselected asset. Four displayed frames, rear-quarter gray geometry and the corresponding wheel-arch crop use matched display cameras; the full 49-frame sequence remains the receiver input.

Preference fitting and learning diagnostics. For each paired draw define $m_k = \beta_D(\Delta_{l,k} - \Delta_{w,k})$, so positive margins favor the preferred video and the loss is $\mathrm{softplus}(-m_k)$. Freeze eligible pairs within the first 100 records of the existing training pool and their four time/noise draws; reuse them across checkpoints at records 0/100/250/500/750/1,000. Skipped records remain in the coverage account, without substitute pairs or new asset generation. Report mean/median and 10th/90th percentile margins, positive/zero/negative fractions, mean per-draw loss, current/reference winner/loser energies, nonfinite values, gradient norms and actual updates. Initialization gives zero margins and $\log 2$ loss while retaining parameter gradients. Compute the nonlinear loss per draw before averaging. These statistics describe fitting on the fixed training-pool subset. Separately track H-F, O-IoU, conditional LPIPS and failures on the existing 100-object normal-deployment validation probe. Charge all diagnostic forwards/evaluations, retain the fixed final checkpoint, and record each policy's own learning curves.

Validation post-training trajectories. Figure 10 uses the fixed 100-object normal-deployment validation probe. Before generating outputs, sort reference-eligible validation IDs by the SHA-256 digest of `aoflow-validation-probe-v117/` followed by the ID, break ties by ID, and retain the first 100. This cohort is separate from the primary test tables and their 100-object structure checklist. The three RL policies share the S1 initializer and the frozen Region-4 receiver. Evaluate the initial policy and checkpoints at updates $100, 200, \ldots, 1000$, using replica r1 and one unselected UniPC asset per object. Object-keyed initial-video and downstream random streams are shared across policies and checkpoints; each video produces its own source memory. Every policy is scored by the same complete asset objective in Equation 4, before group normalization, with the common failure reward. Compute the input penalty per asset before averaging. H-F and input IoU follow Appendix A; reference-empty hidden sets are excluded only from H-F. The displayed probe has 100 attempted objects and 91 reference-applicable H-F objects at each checkpoint. Lines connect checkpoint means directly. The fixed final checkpoint is update 1,000, and validation generation/scoring has a separate cost account.

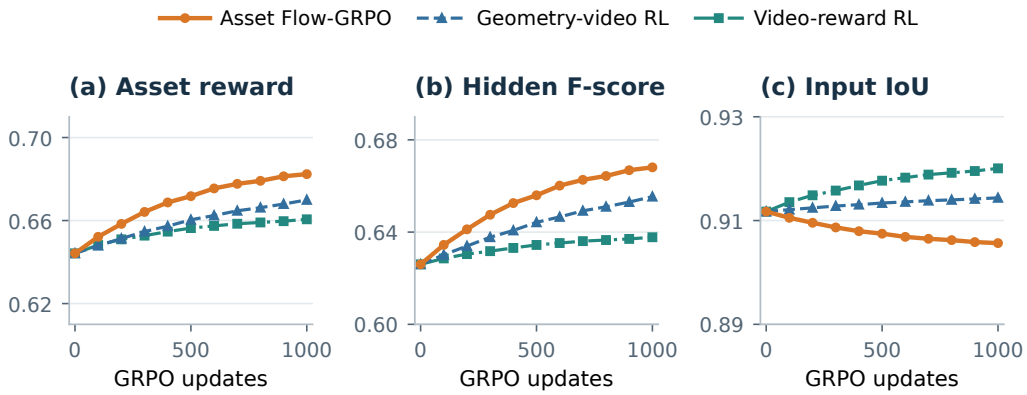

Figure 10: **Validation trajectories during video post-training.** Three RL feedback signals share the initializer and frozen regional receiver. One unselected asset per object is evaluated on a fixed 100-object probe every 100 updates: **(a)** common final-asset reward before group normalization, **(b)** hidden-region F-score, and **(c)** input-view IoU.

Validation selection and safety stops. The pilot uses the fixed 100-object validation probe and one paired deployment draw per object. Its reference is initialization under the same receiver: Region-D before video optimization for the primary controls, and the corresponding Global-D initializer for the stage comparison. Utility is

$$U = .4\,\overline{\mathrm{F}} + .4\,\overline{\text{H-F}} + .2\,\overline{\text{O-IoU}}.$$

Failures contribute zero; H-F excludes only reference-inapplicable objects. A candidate is eligible when mean O-IoU decreases by at most .02 and paired LPIPS increases by at most .02 relative to initialization. LPIPS uses candidate/reference common valid objects; its preregistered coverage requirement is at least 90 of 100. We additionally require failure-rate increase at most .02. These prespecified coverage and failure tolerances define eligibility for this comparison.

Each video-policy training arm has three pilot trials of at most 100 training records: learning rates $\{3\times10^{-7}, 10^{-6}, 3\times10^{-6}\}$, using seed 3101 and the common list in Appendix K.3. Asset-DPO replaces these with $\beta_D \in \{50, 500, 5000\}$ at fixed learning rate 10^{-6}; it receives no extra search. Choose the highest eligible U accumulated in FP64. Differences at most 10^{-8} are ties, resolved by lower learning rate or lower β_D. If none qualifies, record a failed pilot and exclude that arm from the formal comparison. All failed trials and costs remain reported. Formal replicas restart from the common initializer and use fixed final checkpoints.

Nonfinite loss/gradients/parameters, changed frozen receiver/GOA, inconsistent transition scores or incorrect expert/free-mask routing stop a run immediately. At validation input-record indices 100/250/500/750/1,000, two consecutive violations of O-IoU decrease .05, paired LPIPS increase .05, or failure-rate increase .10 relative to the same-receiver initializer trigger a recorded safety stop. Undefined evaluation values pause continuation pending recorded diagnosis. Infrastructure errors allow one retry on identical saved inputs and seeds. Retain a stopped run's recorded endpoint and status. Test-time preservation criteria are specified separately from these operational thresholds.

G.3 Matched Shape-continuation controls

Copy the complete sequentially trained regional receiver into GT-condition and generated-condition branches. Keep π_0^{13} and the entire SS memory producer fixed. The two branches share reference Shape targets and 1,000 prescribed input records; the generated branch uses one unselected initializer video per unique input. Each branch creates memory from its own actual conditions by running the frozen SS sampler. The producer uses its own sampled trajectory and support.

Only Shape and its reader update, with flow supervision and auxiliary weight zero. Four independent Shape time/noise draws are averaged per record. Use fresh matched AdamW states and

reader/denoiser rates $10^{-4}/10^{-5}$ without warmup. The common starting checkpoint and continuation step remain distinct. Condition generation, memory production and training are charged. These branches are evaluated separately from the fixed receiver used by primary video-policy comparisons.

Table 10: Shape continuation under the same fixed video: GT-support and generated-condition controls change receiver training, not the video policy. All entries use the primary 500-object layout and the common reference H-F support.

Receiver	F↑	H-F↑	Input IoU↑	CD↓
S2 receiver	0.744	0.639	0.911	1.527
GT-support Shape continuation	0.746	0.642	0.909	1.515
Generated-condition Shape continuation	0.750	0.647	0.908	1.503

The generated-condition continuation improves H-F relative to the common S2 receiver while leaving the video source fixed. It changes Shape adaptation and is therefore separate from the fixed-receiver policy comparison in Table 3.

Reference-conditioned diagnostics remain separate from image-only deployment. Pose/content failures of generated conditions are retained. Continuing SS defines a separate receiver experiment that requires refreezing the producer, rebuilding memory and readapting Shape.

G.4 Metric and perturbation implementation

Reference visibility is computed at 1024 pixels with camera-z tolerance 0.002 in unit-diagonal coordinates, with no default boundary exclusion. Mask IoU uses the full 512-pixel canvas. Evaluation LPIPS (Zhang et al., 2018) uses the spatial AlexNet v0.1 map, bilinearly resized to the canvas and averaged over the union of input and rendered foreground masks; generation failures remain distinct from evaluator errors and undefined metrics. The point-sampling and coordinate definitions in Appendix A apply to all methods.

The corruption suite uses foreground Gaussian noise with standard deviations $\{0.03, 0.06\}$ in $[0, 1]$ RGB, Gaussian blur of $\{1, 2\}$ pixels, and foreground occlusion areas of $\{5, 15\}\%$. The same saved perturbation is supplied to every method. Main corruption preservation scores use the clean input as target; actual-input reproduction is separately named, with occluded pixels excluded from that secondary diagnostic. Average the six conditions within each original object before object-level inference.

The displayed column in Table 7 is specifically the foreground-blur condition `blur_1`, with Gaussian standard deviation one pixel. This reporting profile uses the same saved blurred input for all methods and evaluates preservation against its clean counterpart. The full-suite composite requires all six condition records and has a separate population and cost account.

G.5 Structure checks and statistics

The reward audit has two phases: initializer π_0 and the prescribed final asset-RL policy, both evaluated through the same regional receiver $\bar{\phi}^{13}$. Population `audit_stochastic_v13` uses the scored 20-transition stochastic sampler, four complete videos per input and one asset per video ($K = 1$); all scorers share the same pool within each phase. The same 40 validation objects and seed indices are used across phases, while generated content may change. Pools bind policy, receiver and sampler identities. This population is distinct from normal deterministic deployment (`eval_det_v13`) and its paired sampler-transfer diagnostic (`eval_stochastic_v13`). Additional DPO and geometry-policy audits are separately budgeted extensions.

Spearman correlation assigns average ranks to exact ties. A constant score or checklist vector makes correlation undefined: the group is excluded only from that scorer's correlation mean, with defined/undefined counts retained, and remains in selection, preservation and failure summaries. Selection quantities use the exact expectation over uniformly chosen maximum-score candidates, including failures, without a checklist-based tie-break. Appendix L.1 separates the structural and selection populations. These audits measure reference-assisted ranking on validation objects; normal single-sample deployment is evaluated separately.

For each phase separately, let $s_{oi} \in [0,1]$ be candidate i's independent checklist score for object o, and $r_{oi}^{(h)}$ its score under reward scorer h. Define the maximum-reward tie set and regret by

$$T_o^{(h)} = \arg\max_i r_{oi}^{(h)}, \qquad \bar{s}_o^{(h)} = \frac{1}{|T_o^{(h)}|} \sum_{i \in T_o^{(h)}} s_{oi}, \tag{42}$$

$$\mathrm{Regret}^{(h)} = \frac{1}{N} \sum_o \left(\max_i s_{oi} - \bar{s}_o^{(h)} \right), \qquad \mathrm{RandomRegret} = \frac{1}{N} \sum_o \left(\max_i s_{oi} - \frac{1}{G} \sum_i s_{oi} \right). \tag{43}$$

The random reference is an exact expectation on the same complete pool. Selected structure, IoU and failure indicators are first averaged over each tie set, then equally over all N objects. Groups with undefined correlation remain in these summaries. Four-candidate correlations describe ordering within the sampled groups.

Unselected policy quality is $a_o = G^{-1} \sum_i s_{oi}$, with failures retained, and maximum attainable quality in that pool is $b_o = \max_i s_{oi}$. Report their object means separately from score-selected quality, and pair the phase change as $N^{-1} \sum_o (a_o^{\mathrm{post}} - a_o^{\mathrm{init}})$. Apply the same averaging to the three checklist components and failure indicators. $\bar{b} - \bar{a}$ equals expected random regret; the audit record stores both quantities and correlation support. Changes describe this stochastic audit population. Annotation records preserve the required-part/relation denominators and missing/relation/extra-instance counts underlying the quantized scores.

Prespecified audit sampling. Before candidate generation, freeze the eligible validation inventory with all three checklist components applicable. Order its IDs by the SHA-256 digest of the UTF-8 key formed by joining prefix `aoflow-audit-v1/20260917`, `object`, and the object ID with slashes; take the first 40, breaking digest ties by ID. Freeze IDs and category counts before seeing outputs. Overlap with the 100-object learning probe is allowed; both populations use validation objects exclusively. Use the first prescribed sequential-S2 replica r1 after both phases, binding its SS producer and Shape checkpoints in the audit manifest.

Audit generation uses canonical keys under `aoflow-audit-v13`, registered jointly with ordinary evaluation (Appendix L.1). Video keys include object identity and candidate index 0–3; each transition key includes the complete parent video key and transition index 0–19. Downstream, point-sampling and render keys are object-shared. Initializer/post-RL phases intentionally reuse these keys. The original full-digest rule still determines blinded annotation order, tied by object/candidate identity. The frozen registry, model identities and saved pools accompany the audit; full key serialization belongs to the companion profile.

For the 40-object audit, freeze a reference-only inventory of auditable part-instance slots and required relations before inspecting outputs. Relations comprise required connections and inspectable openings/separations. Exclude ambiguous slots and their incident relations in this inventory, identically for every method and phase; eligibility requires positive part and relation denominators and applicability of all three components. Permitted hidden alternatives must specify mappings to the same scored slots and relations, with unchanged denominators. Only alternatives with such a common mapping enter the frozen inventory.

Two raters independently inspect blinded gray/textured renders from eight fixed azimuths; a third adjudicates disagreement without reward or method labels. These *perceptual structural judgments* assess visible part completeness, attachment and separation. A one-to-one assignment maps visible candidate instances to the frozen required slots. Let n_P be the required-slot count, m the unmatched required slots, and x the unmatched candidate instances in the declared auditable part classes. Permitted alternatives count as matched instances. Let n_E count required connection/opening relations and b count those violated. Each required relation contributes at most one error: a missing endpoint, an absent or incorrectly located/paired attachment, or a closed required opening. Multiple defects on the same relation are not summed; one missing part can also violate its incident relations. Additional connections enter the relation count when they violate a declared separation/opening; extra-part counts concern unmatched instances.

Define the three error fractions and component scores as

$$e_P = m/n_P, \qquad e_C = b/n_E, \qquad e_D = \min(x/n_P, 1), \qquad f(e) = 1 - \tfrac{1}{4}\lceil 4e \rceil, \quad 0 \le e \le 1, \tag{44}$$

and $s_{oi} = [f(e_P) + f(e_C) + f(e_D)]/3$. Thus errors $0, (0, .25], (.25, .5], (.5, .75], (.75, 1]$ yield $1, .75, .5, .25, 0$. Extras exceeding n_P saturate the nonduplication error at one. Only explicitly required visible apertures and separations enter n_E; legitimate open boundaries retain their reference-defined interpretation.

Worked checklist cases. Consider four required parts—body, handle, spout and base—with three required body attachments and one open spout outlet ($n_P = n_E = 4$). A duplicate handle adds one unmatched instance while all required relations remain correct: $(e_P, e_C, e_D) = (0, 0, .25)$, component scores $(1, 1, .75)$ and $s = 11/12$. A present but disconnected handle instead violates its one required attachment: errors $(0, .25, 0)$, components $(1, .75, 1)$ and $s = 11/12$. A handle attached at either of two predeclared admissible hidden body sites, with the other requirements satisfied, is an acceptable alternative: all errors are zero and $s = 1$. These worked cases specify the checklist calculation. Six extras in this inventory saturate at $e_D = 1$.

A generation-failed complete asset receives three zero utility components and $s = 0$, remains in the pool and is counted as failed. For a material-only failure with valid geometry, separately record gray-geometry checklist components without changing its full-asset failure. An all-failed pool has zero regret, selected failure one and undefined correlation; report these quantities together. Renderer/infrastructure exceptions follow the shared retry rule and retain their separate error status. Preserve the original annotations, excluded-item inventory, agreement and adjudication records. Unresolved render ambiguity remains in the records for the frozen inventory.

Mix both phases before blind annotation, hiding policy, stage, reward and gate values. Aggregate the 320 candidate assets into object-level changes. Report phase-specific undefined correlations, score support, selected-output preservation and failures alongside means. Retain undefined-rank counts when comparing phases, since narrower post-RL pools can increase them. Distinguish reward ordering, candidate selection and policy learning throughout.

A ten-object controlled-defect audit uses the first ten separable-part-eligible audit objects ranked by SHA-256 of slash-joined prefix `aoflow-defect-v1/20260917` and object ID. Parts and intervention directions are fixed from references before viewing candidates. It applies predeclared part deletion, duplication, disconnection, texture-only and light-only changes. Total reward may respond to texture through preservation even when geometry is unchanged; natural-candidate correlation and controlled-defect sensitivity remain separate diagnostics.

Local source interventions. Two validation-only populations have distinct eligibility and interpretation. Freeze object/part inventories, acceptable attachment alternatives, masks and intervention records before inspecting any downstream comparison output. Rank objects by SHA-256 of `aoflow-local-v13/20260924/` plus object ID and take the first eligible objects, using lexicographic reference-part IDs to choose one target per object. The actual input stays unchanged; paired conditions share component seed keys and retain their unselected outputs.

Natural rare-cue panel. Screen at most 500 validation objects, one unselected initializer video each, for at most 40 eligible objects. The reference target part has less than 5% surface visibility in the input camera. Two raters inspect each generated frame against the reference checklist, judging target identity, visible attachment and shape before seeing generated assets; a third adjudicates disagreements. Require an appearance-consistent cue in 1–4 of the 49 frames, none at 0/48, with at least 64 visible target pixels per cue frame. Unresolved judgments and insufficient sham area are screening exclusions, counted before any 3D comparison. The resulting population conditions the diagnostic on the prescribed visible cue.

Raters outline the target directly on the generated frames to obtain the intervention masks. Select 16×16 DINO patches with at least 50% target coverage (at least one patch). Erase all cue frames' selected features by replacing them with the corresponding patch features of one white 512-pixel canvas processed by the same DINO pipeline. The sham replaces the same number of foreground patches outside the target mask dilated by 16 pixels, choosing greatest center distance first and row-major ties. Patch position, nominal camera and frame ID stay fixed; the edit changes both key and value source features. DINO self-attention can distribute target information into unedited patches,

so this ablation measures dependence on the selected cue tokens while other encoded cues remain available.

The full-pipeline branch reruns SS and Shape from the edited bank. The fixed-producer branch retains the original SS latent/support, full source-mass table, candidate IDs, retained masses, null and all sampler identities; only the Shape bank changes. Save and restore the original Shape-entry RNG state after SS. The full rerun restores the original pre-SS state; candidate-specific supports can map the shared stream to different voxelwise noise. Texture still receives actual I. Report original, erased and sham conditions with no intervention-specific fitting. The fixed-memory branch measures consumer dependence, while the full rerun includes changes in SS support and selection.

Controlled repetition panel. A separate reference-only population takes at most 20 hash-ranked validation objects with a separable hidden part visible over at least 64 pixels in eight interior orbit frames. Render clean GT conditions, then delete that part and rerender under the identical camera/light. Composite the changed image only within the union target mask, dilated by two pixels. Sort eligible interior frames by the object/frame hash under `aoflow-repeat-v13/20260924`; nested doses edit the first 0/1/4/8 frames. Reencode the changed RGB. The full branch recomputes SS/memory; the fixed branch retains every clean-producer field above. These are oracle-condition omission tests, separate from generated-video deployment.

Endpoints and denominators. The primary endpoint is blinded, adjudicated binary correctness of the prespecified local attachment; also record part presence and extra-instance count. Local F@1% uses the reference part plus its attachment neighborhood within .02 normalized diagonals, and prediction samples inside their fixed reference box expanded by .02. A nonempty reference with empty predicted region scores zero. Use the common 100,000-point surface-sampling protocol; an empty local reference set is reference-only N/A for all arms and is counted separately. Separately report the fraction of reference target-part 32-grid cells present in predicted SS support, source retention in their fixed parent regions, real prior/posterior mass and residual norms. This requires the fixed native adapter, without feeding reference support to the producer.

Each object has equal weight in original–erased, original–sham and difference-of-change contrasts. Keep screened/eligible counts, edited frames/patches, reference applicability, failures, completed/unresolved records and paired support. Candidate failures receive zero asset utility; material-only failures retain separate valid-geometry scores. Infrastructure errors or unresolved annotations block the affected paired export until resolved; keep their status separate from generation failures. Object eligibility follows the frozen screening inventory. Charge screening, annotation, condition construction, both inference branches and C8. The fixed inventories and raw annotations accompany each reported comparison.

The full-run protocol uses three paired replicas with explicit (S2a, S2b, S3) seed triples $(1101, 1201, 2101)$, $(1102, 1202, 2102)$ and $(1103, 1203, 2103)$. Separate stage namespaces drive the registered random streams; matched arms share their prescribed within-replica pairing keys, with separate streams for each stage. Evaluation uses four normal inference draws per object and replica. These repeat one fixed input; Appendix L.1 binds input selection and component seeds. Fixed native systems use matched evaluation panels for inference-draw variation. Metrics are averaged within object/replica, then across replicas, then objects. H-F excludes only the common reference-inapplicable objects; CD and LPIPS retain separate valid-output denominators.

For either conditional metric, first average its valid draws within each method–object–replica cell. A cell with no valid draw is undefined. A method's standalone estimate averages its defined replica cells equally for each object, then averages defined objects equally; each defined object has equal weight. For a paired contrast, intersect defined cells across *all* compared methods and retain only objects having a common defined cell in *every* prescribed replica. Compute the contrast within those paired cells, average all replicas equally, then average retained objects equally, reporting excluded objects and each metric's valid-draw/cell/object counts. Each arm averages its own valid draws within the common supported cells, yielding a conditional comparison. No retained objects means an undefined contrast. For six-condition composites, a cell must first have a defined conditional mean in every condition; incomplete cells stay undefined before the same intersection rule.

Primary paired H-F contrasts compare Region-D with Global-D before video optimization, each receiver's own S2-to-S3 change, their difference in changes, and asset RL versus geometric video RL or Asset-DPO under the same regional receiver. Spatial and retained-mass controls have separate

contrast identities. The comparison includes both the online versus fixed-offline sampling schedule and the optimization objective. Report preservation, failures, geometric diagnostics and cost alongside complete-asset quality. Deployment and stochastic populations remain separate. The full-run plan uses 10,000 object bootstrap resamples with a recorded PCG64 seed for percentile 95% intervals conditional on the fitted replicas; replica values and sample SD are separate. Intervals are marginal percentile intervals for the specified contrasts.

A preservation-compatible improvement requires a positive geometry gain and preservation/failure tolerances fixed before the final test. Otherwise, interpret the results as paired geometry–preservation trade-offs. Specify test-time criteria separately from validation-pilot rejection tolerances.

H Reporting denominators and additional metrics

The four main tables use one training replica and one unselected deployment draw per object. The primary set contains 500 objects; external GSO contains 200. Repeated-seed evaluation is specified separately above. Each record retains attempt counts, geometry/material failures, reference-only hidden-region applicability and conditional metric support. Our blinded checklist pass rate uses the 100-object test structure cohort and requires all reference-defined part, relation and extra-part items to pass. This task-specific test cohort is separate from the 40-object validation diagnostics (Appendix L.1).

Input-view normal error. Following rendered-normal evaluation in Pixal3D (Li et al., 2026), we measure input-view mean normal angular error (N-MAE). Generated and reference geometry are rendered at the fixed input camera under their registered coordinate transforms. Use the common 512^2 canvas, the nearest-surface geometry normals and camera transforms in Appendix L.2. Orient normals consistently toward the camera and normalize each vector. On the intersection of predicted and reference foreground masks, compute $\arccos(\mathrm{clip}(n_p^\top n_*, -1, 1))$ in degrees, average pixels within each object, and then give each supported object equal weight. This compares oriented geometry normals directly in camera space, using the signed dot product on overlapping pixels. Surface NC_τ instead uses distance-gated 3D nearest neighbors, while LPIPS (Zhang et al., 2018) compares learned image features. Empty mask intersections and complete-asset failures have undefined N-MAE; report their counts and the separate silhouette IoU/failure rate. Native system coordinate adapters are fixed before evaluation. The reference normal target comes from the reference mesh.

CD, N-MAE and RGB LPIPS use their declared conditional supports; F/H-F/NC_τ/IoU remain complete-asset utility metrics with generation failures scored as zero. H-F excludes only reference-inapplicable objects. Geometry-only diagnostics retain valid geometry after material failure alongside the unchanged complete-asset status. Repeated methods in different tables use the same evaluated outputs.

Table 11: Evaluation support for Tables 1–4. N is the attempted-object count; H denotes reference-applicable H-F objects. CD, N-MAE and LP give conditional object counts; Struct. is the test-subset size, and F counts complete-asset failures.

Configuration	N	H	CD	N-MAE	LP	Struct.	F
Hi3DGen	500	454	493	493	493	0	7
TRELLIS	500	454	489	489	489	0	11
TripoSG	500	454	494	494	494	0	6
Hunyuan3D 2.1	500	454	495	495	495	0	5
TRELLIS.2	500	454	491	491	491	0	9
InstantMesh	500	454	487	487	487	0	13
Pixal3D (paper version)	500	454	492	492	492	0	8
V23D	500	454	490	490	490	0	10
AO-Flow S2	500	454	493	493	493	100	7
AO-Flow	500	454	493	493	493	100	7
Fixed-4	500	454	493	–	493	0	7
SS49 / Shape-no-video	500	454	493	–	493	0	7
Global-4	500	454	493	–	493	0	7
Full-49	500	454	493	–	493	0	7
Shape-selected-4	500	454	493	–	493	0	7
Region-4, uniform prior	500	454	493	–	493	0	7
Region-4, no spatial bias	500	454	493	–	493	0	7
Continued SFT	500	454	493	–	493	100	7
Best-of-4 SFT	500	454	493	–	493	100	7
Video-reward RL	500	454	493	–	493	100	7
Geometry-video RL	500	454	493	–	493	100	7
Asset-DPO	500	454	492	492	492	100	8
w/o hidden coverage	500	454	493	–	493	100	7
w/o normal consistency	500	454	492	–	492	100	8
w/o input penalty	500	454	492	–	492	100	8

I MATCHED INTERFACE AND REWARD CONTROLS

Receiver comparisons. All new controls preserve native conditioning from the actual input image. Fixed-4 changes both SS and Shape access and is a whole-interface reference. The other controls share the same trained all49 SS producer, while Shape branches start from the same backbone and receive 10,000 matched S2b updates. Shape-no-video removes video reading only from Shape. Global-4 uses equal ownership votes over SS grid addresses. Full-49 exposes all video patches to Shape; it provides the dense-reading reference.

Shape-selected-4 replaces SS source inheritance with current-Shape routing. At each positive Shape denoiser call, its reader first probes all 49 frames with the current queries and fixed nominal-camera cues, aggregates patch attention by frame, and retains up to four positive sources per query. It then rereads their patches, using the probe's source masses as priors; discarded mass joins null. Probe scoring and feature reading are both charged. This compares routing designs with their respective source choices and posteriors. The uniform-prior control instead keeps the SS source IDs and total retained mass a_r fixed, sets each selected prior to $a_r/|J_r|$, and leaves null unchanged. Empty rows keep their null-only behavior. No-bias reading changes only κ to zero. Adaptation budgets are matched, while each branch learns its own feature norms and attention distributions.

Table 12: Regional source controls with aggregate real/null prior mass held at 0.72/0.28. Matching this partition does not equate final Shape attention or injected residual norms.

Control	H-F↑	Retained real	Null	Input IoU↑
Global-4, retained mass matched	0.625	0.72	0.28	0.912
Region-4, retained mass matched	0.631	0.72	0.28	0.911
No spatial bias, real/null matched	0.628	0.72	0.28	0.910
Full regional interface	0.639	0.72	0.28	0.911

Under matched aggregate real/null prior mass, per-region source routing has higher H-F than a global source set. The matched partition does not hold final Shape attention or residual magnitude fixed. The no-bias partition control separates part of the real/null allocation change from spatial guidance.

Reward and structure comparisons. Reward removals set one coefficient in Equation 4 to zero, leaving the others, reference regularization, group size and training budget fixed. Global F-score remains present. The remaining coefficients retain their original values, so the removal contrast includes the changed reward balance. Table 4 uses common evaluation metrics and the same held-out structure cohort. The checklist separately tests required parts, connections and extra-part instances, alongside geometric and input-preservation metrics. Paired small-gap, duplicated-surface and orientation interventions remain separate diagnostics under the blinded checklist in Appendix G.

Cost scope. Receiver seconds in Table 2 start from an already generated video and include DINO encoding, SS sampling and memory construction, Shape routing/sampling, texture, decoding and export. Report mean per-object generation time and maximum allocated GPU memory on the same single H200-141GB target, with batch one, BF16 forwards, fixed output resolution and no cross-object feature reuse. Include failed attempts; exclude offline reference preparation, video generation and post-generation C8 evaluation, which have separate ledger entries. Shape-selected-4 includes its full-bank probes. Post-training GPU-hours include candidate video generation, complete receiver calls, reward/inspection work, offline selection pools and policy updates; shared S1/S2 adaptation, pilots and final evaluation are separately accounted. These durations are distinct from the reader-only MAC proxy in Figure 6.

Table 13: Incremental post-training GPU-hours by activity. The table excludes shared video/receiver adaptation and final test evaluation; training arms need not have equal cost per update.

Policy	Candidates	Scoring	Selection	Updates	Total
Continued SFT	0	0	0	148	148
Best-of-4 SFT	220	80	6	120	426
Video-reward RL	170	45	0	53	268
Geometry-video RL	280	170	0	102	552
Asset-DPO	420	250	20	100	790
Asset Flow-GRPO	380	320	0	164	864

The activity totals include each arm's own candidate generation, scoring, offline selection when applicable and parameter updates. Figure 10 uses update count as its horizontal coordinate; no cumulative GPU-hour trajectory is inferred from that axis.

J CONTEXT FOR THE CONDITIONING INTERFACE

Table 14 compares the direct precedents' representations, geometric assumptions and update scopes. AO-Flow studies regional source retention from one SS trajectory to Shape under correlated generated inputs.

What is retained from each precedent. Stream3D motivates source indexing. AO-Flow extracts regional subsets from an existing SS call and trains the receiver; Stream3D uses a separate probe and training-free velocity fusion. Its evidence scores combine attention mass and spatial concentration, whereas our e_{rj} stores joint SS attention mass as a prior for Shape reading. Pixal3D motivates spatial reading. Our finite projection bias uses the registered $X_N = 0.9X_A$ contract as a soft prior: cells are addresses, and nominal cameras need not match the generated views. Ink3D motivates auxiliary supervision on confirmed GT surface relations; AO-Flow uses the optional distribution loss in Appendix F.3, while Ink3D uses geometry-first inference and texture processing.

Related conditioning mechanisms. UMI3D provides voxel routing, ROAR-3D trained token routing, and ReconViaGen global/local conditions. VRS relies on native special-token semantics. Vision–language and image-editing hidden-state interfaces provide further conditioning prece-

Table 14: Direct interface precedents and the regional evidence interface. Geometry available during training is distinguished from geometry required at inference.

Method	Representation	Selection or geometry	Injection and updates
Stream3D (Zhou et al., 2026)	Query–view evidence; top-D frame indices and scores.	Fixed-noise probe; query memories vote for a shared global top-K set.	Per-view velocity fusion with a frozen generator; probe work is additional.
Pixal3D (Li et al., 2026)	Projected image features in a spatial condition volume.	Camera-aligned 3D targets; multi-view extension assumes known cameras.	Spatial condition volume trained jointly with the 3D generator.
Ink3D (Han et al., 2026)	Geometry-first video features, position/normal maps and surface neighborhoods.	Existing geometry guides video and texture transfer; texture evaluation supplies GT geometry.	Geometry-guided attention, texture assignment and refinement for texturing.
UMI3D (Qu et al., 2026)	Native DINO patch and CLS/register tokens.	Per-voxel VRS from patch attention mass; hard view assignment and KNN refinement.	Training-free view-specific cross-attention inside the frozen 3D model.
ROAR-3D (Sun et al., 2026)	DINOv2 patches; mean-pooled view keys and 3D-token queries.	Trained hard view router; primary and auxiliary views have distinct roles.	Dual-stream attention and orientation perturbation; hard Gumbel routing during training.
ReconViaGen (Chang et al., 2025)	VGGT features resampled into global and per-view tokens.	Global SS conditions and local SLAT conditions; estimated cameras for refinement.	Per-view attention fusion plus rendering-aware velocity compensation.
Region-D	All 49 DINO patch grids; fixed 16^3 SS addresses; per-region source masses.	One actual SS call forms memory; requested 32^3 Shape support uses fixed lookup and nominal-camera soft bias.	Separate 128-wide readers; train SS, freeze producer, then train Shape; S3 updates video attention LoRA only.

dents (Ye et al., 2026; Chen et al., 2026). AO-Flow focuses on retaining regional source IDs and priors across SS and Shape.

Global-D, Region-D and Full-49 share SS, complete evidence, support and seed, with matched Shape adaptation. Global priors come from the full table; Full-49 retains all sources. Truncation changes content and prior mass, assigning discarded mass to null. A downward mass-matched diagnostic controls the latter; spatial bias and auxiliary supervision are tested separately. Auxiliary-without-injection distinguishes query regularization from content transfer.

Inherited video mechanisms and the legacy receiver. V23D supplies the orbit-conditioned GOA residual alongside video LoRA (Anonymous, 2026); π_0^{13} additionally requires contract-consistent S1 adaptation. Legacy V1.2 uses four-grid ACB and Qwen before OFA, whose global/local tokens condition SS/Shape. The new geometry path uses native input-image conditions and sequentially adapted readers. V1.2 is a whole-system baseline with its own OFA checkpoint and initialization.

K Concrete execution profile

K.1 Model identity and input processing

The new AO-512 receiver uses Wan2.2-Fun-A14B-InP, the TRELLIS.2-4B 512 branch and frozen DINOv3 ViT-L/16. Its video initializer retains the inherited attention/FFN adapters and GOA; its new geometry readers replace ACB/OFA/Qwen. V23D-native and the V1.2 receiver remain separate system configurations.

The VideoX-Fun and TRELLIS.2 code revisions begin `bce42b301f59` and `75fbf0183001`; their model revisions begin `2d27b954ae26` and `af44b45f2e35`. DINO's repository identity begins `ea8dc2863c51`. Full identifiers and source hashes accompany the execution profile. The new readers use separate parameter sets, native image conditions from actual I, and the address contract in Appendix F. The common initializer uses the orbit adaptation specified below.

Images share a deterministic float-sRGB canvas: resize the longer side to 512 pixels, round the shorter by $\max(1, \lfloor x + .5 \rfloor)$, and center on white, assigning odd padding right/bottom. Float Lanczos-3 uses half-pixel centers, downsampling antialiasing and replicated borders; clamp overshoot after resizing. Actual rounded dimensions define the pixel transform. The canvas transform is fixed across outputs. Depth retains camera-Z units and its separately defined interpolation.

Model input quantization. Inputs are checked for finite values. Corruption is applied on the float canvas, followed by one quantization in FP64 by $u = \lfloor 255\,\mathrm{clip}(I, 0, 1) + .5 \rfloor$. Wan hints and native image conditions use those same actual-input bytes. All 49 GT/generated frames use the same canvas and quantizer before frozen DINO. Its native processor consumes PIL RGB with its own rescaling/normalization, means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$; no further foreground crop is applied. Raw VAE output in $[-1, 1]$ receives $x/2 + .5$ once; Fun's already normalized $[0, 1]$ output does not. Feature-cache identity includes processed HWC/u8 bytes, shape and encoder/processor identity. Float evaluation targets are separate. The complete native DINO sequence has 1,024 spatial patches, one CLS and four storage tokens; only patches enter the video bank.

K.2 Coordinates, cameras, and sampling

For raw reference coordinates, define

$$X_A = R_z(-\alpha)(X_{\mathrm{raw}} - b)/L, \qquad X_N = \gamma X_A, \quad \gamma = .9. \tag{45}$$

Here b, L are reference bounding-box center and diagonal; α is the selected input azimuth. GT input/orbit images are rendered in X_A, whereas frozen 3D encoders receive $\mathcal{A}^{-1}(X_N)$ under a fixed native-axis adapter $\mathcal{A}$, preserving the target scale, orientation and center. Output geometry is mapped back by the fixed inverse and $1/\gamma$, without reference fitting. Native material attributes are queried before this world-coordinate conversion.

The nominal orbit has $a_j = 2\pi j/48$, elevation $15°$, radius 1.8 and perspective field of view $40°$. Positions use the declared input-aligned, z-up convention and OpenCV world-to-camera extrinsics. For native unit-grid center u, set $x_N = \mathcal{A}(u)$ before projection:

$$\pi_j(x_N) = \mathrm{inhom}(T_j K_j [R_j \mid t_j][x_N/\gamma; 1]), \tag{46}$$

where $[R_j \mid t_j]$ is 3×4 and T_j includes canvas, crop, padding and pixel centers. The closed endpoints share a nominal direction, not necessarily identical generated pixels. The new S1 orbit adaptation uses this same rendering contract, starting from inherited weights and producing the common π_0^{13}. V23D describes orthographic S1, perspective S2 pools and accelerated 1,024-pixel inference (Anonymous, 2026); the common initializer is adapted to the AO-512 perspective orbit defined here.

Supervised orbit-adaptation budget. The common initializer has one replica: 10,000 updates at effective batch eight, learning rate 10^{-5} and 500-step linear warmup, using the stated AdamW contract and inherited S1 flow objective/adapter scope. A preliminary execution check uses at most 500 additional updates and 20 fixed validation objects with one deployment draw each. It checks coordinates and conditioning with fixed learning rates, cameras and checkpoint choice; permitted adjustments are fixed-axis binding from known-coordinate fixtures and gradient-equivalent memory chunking. Final training resets weights/moments to the inherited initialization and uses update 10,000. Both the preliminary check and full adaptation are charged.

The positive video text is "A single static object. The camera makes one smooth complete orbit around the object. Preserve its identity, parts, materials and proportions. Constant illumination and a plain white background." The negative text is "camera shake, object motion, deformation, changing parts, duplicated parts, changing colors, flicker, text, watermark, moving background." These fixed prompts use no reference-only shape descriptions. The same observed image supplies both first and last hints, so the endpoint has no additional observation.

The Fun side condition has four mask channels and 16 masked-video latent channels, separate from the sampled 16-channel state. VAE stride $(4, 8, 8)$ gives $16 \times 13 \times 64 \times 64$ sampled coordinates. The all-free density in Appendix F.5 requires the actual GOA wrapper to preserve them. Native time is

$1,000\tau$ with the public Fun expert boundary 900; the legacy reported boundary .358 is not equated without its time convention. The video flow convention remains $\epsilon - z_0$.

Deployment uses the 20-step Fun UniPC solver: shift 5, order 2, `bh2`, flow output in its x_0 form, lower-order final steps, terminal sigma zero and video CFG 5. Dynamic shifting, thresholding, TeaCache, CFG skipping, RIFLEx changes and latent-changing callbacks are disabled. Its once-shifted sigmas begin from .999, with integer model times. The scored stochastic rollout instead starts at continuous $\tau = 1$ and retains all positive-variance transitions. Sampler populations are evaluated separately.

SS, Shape and texture use 12-step Euler flow sampling with time rescale $(5, 3, 3)$ and $\sigma_{\min} = 10^{-5}$. New SS/Shape execute the positive condition directly (strength 1), without a negative pass or guidance rescale; texture retains its native strength-1 path. The geometry CFG diagnostic pairs a positive call with a zero native-image condition and a hard bypass of all video reading in the negative call. It retains latent, time, support and native positions but never reads the bank, memory indices or masses. Condition dropout is zero; the negative response is not assumed to be calibrated unconditional generation. The memory capture always belongs to an actual positive SS call.

Complete material-conditioning route. The fixed texture path independently consumes the actual task input, including any prescribed corruption. Native DINO supplies its full $1 \times 1029 \times 1024$ image condition; no video, region memory or Qwen feature enters this material condition. Shape SLat is restandardized with pinned shape mean/std and supplied as a 32-channel concatenated condition. Independent 32-channel texture noise on the same sparse coordinates completes the 64-channel input. Texture output has its separate normalization. Fixed shape decoding produces geometry/subdivision guides; fixed texture decoding maps six attributes by $x/2 + 1/2$ to RGB, metallic, roughness and alpha. Native postprocessing is shared across matched arms, without candidate-specific fitting. The downstream random stream is consumed in SS, Shape, texture order; memory adds no random draw. Material-input encoding and common inspection are explicit costs.

K.3 Record selection and random streams

Freeze identity-deduplicated inventories before outputs. Preserve the declared 10,000/500/500 Objaverse-derived train/validation/test split and 200-object external GSO inventory; four possible input azimuths are $0, 90, 180, 270°$. Split membership depends on reference identities, not method output. Input azimuth only defines the reference preparation above; it is not an extra pose observation supplied to the generator.

Use the fixed rejection-sampled hash integer routine for uniform training object/view selection, with a new V1.3 namespace and separate phase identifiers. S2a and S2b each prescribe 80,000 object records grouped in eights; their flow noises/times and S2b's SS sampling noise use independent streams. S3 prescribes 1,000 input records per replica and four videos per record. Repeated inputs remain records. GT/generated Shape continuations share input identities but compute memory from their own actual conditions. Memory reuse requires the complete identity key in Appendix F, not merely an object ID.

Before any campaign, materialize the complete canonical random-key registry and resolve uint32 collisions by the recorded sorted linear-probe rule. Paired controls share the intended stream keys; differing sparse supports need not share voxelwise noise realizations. Inventory, registry, profile, input and checkpoint identities travel with each output. Audit and deployment keys are separate; selection never uses a reference test mesh to choose a deployed sample.

K.4 Update scope and system references

S2a updates complete SS plus its reader; S2b freezes that producer and updates complete Shape plus its own reader. Frozen DINO/reference targets may be cached losslessly in their emitted dtype. Trainable Shape projections cannot be cached across updates. Non-reentrant activation checkpointing is a permitted implementation with separately verified gradient equivalence and recorded replay work. S3 updates existing attention LoRA in both video experts only; loaded alpha/rank multipliers are preserved, FFN LoRA remains active/frozen, and GOA never updates. The old policy/reference

retain the same full adapter composition. Controlled runs use a separable pre-Lightning initializer with the supervised adapters active.

Each new receiver begins from the same pinned public 512 geometry backbone within its comparison family; the shared-all49 controls reuse the exact trained SS producer. Historical V1.2 instead uses its inherited OFA/SS/Shape bundle and original guidance. Its full-system difference also includes the input condition, coordinate contract and initializer, making it a whole-system comparison.

Optimizer contract. Use AdamW, betas $(.9, .999)$, epsilon 10^{-8}, weight decay .01, bias correction and no AMSGrad. Decay covers all permitted trainable tensors. FP32 master parameters, accumulated gradients and moments accompany BF16 forwards. Clip the complete normalized gradient once at L2 norm 1 before one update. Clear gradients between effective batches; frozen or no-gradient parameters receive no artificial decay/update. S2 phase rates are 10^{-4} for readers and 10^{-5} for denoisers, scaled by $\min(u/500, 1)$ for phase update $u = 1, \ldots, 10,000$. Video-policy RL/SFT/DPO use constant 10^{-6}; Shape continuations use the matching S2 rates without warmup. Fresh arms/phases start with fresh moments; resumed runs restore all state. Skipped offline records advance the record counter only. Equal-reward RL groups may still receive reference regularization and ordinary decay.

Native system references retain their published pipelines and separately registered costs. Know3D/UniVerse3D use the neutral user request "Generate a 3D asset consistent with the input image" under their recorded internal templates. These prompt-adapted configurations are identified separately from native reproductions, with source, checkpoint and sampler identities recorded for each system.

L Evaluation records and renderer contract

L.1 Fixed observations and paired random streams

The evaluation unit is an object with one fixed observed image. For each eligible primary-test, GSO or validation object, select one of azimuths $0, 90, 180, 270°$ by the hash rule below, before generating outputs. Render that view with the camera and input-aligned transform of Appendix K.3; GSO inputs are these protocol renders, not photographs supplied by the dataset. The choice is unchanged across methods, training replicas, evaluation draws, corruptions and sampler diagnostics. Four draws therefore mean four realizations from the *same* input, not four observed azimuths. Input RGB, reference alpha, camera, canonical transform and their hashes belong to the object record.

The companion evaluation JSON specifies all record fields. Its hash keys are compact UTF-8 JSON arrays, with no ASCII escaping, under prefix `aoflow-eval-v13`. For input selection, hash the prefix, `input`, dataset namespace, object ID and counter; apply the full-digest rejection sampler of Appendix K.3 with $m = 4$. Method, replica and draw are absent from this key. Each output record additionally contains population, method/checkpoint and receiver identities, replica $r \in \{1, 2, 3\}$, draw $d \in \{0, 1, 2, 3\}$, condition, sampler and all component seeds. Fixed native systems reuse these evaluation panels with their fixed checkpoints.

Ordinary-evaluation seed keys contain dataset, object, replica, draw and component, omitting method, corruption and sampler to define paired streams. Video initial and transition noise have separate components. A downstream seed is consumed in SS–Shape–texture order without resets; different supports can consume different variates, so pairing is defined by the stream keys. Reference points share an object seed; candidate surfaces use common draw-level uniforms. Evaluation and training-reward point streams are separate. Deterministic rendering retains its provenance seed without consuming variates. Object-keyed corruptions are saved once per condition and reused across methods, replicas and draws. All uint32 lookups use the frozen registry below.

Freeze and resolve seeds before generation. Collect the complete ordinary-evaluation, reference/render and audit key set for the declared campaign. Canonical keys are compact UTF-8 JSON arrays of strings, integers and nested arrays; deduplicate identical keys and sort by their encoded bytes. For each distinct key, take the first four SHA-256 bytes as unsigned big-endian base b and assign $(b + c) \bmod 2^{32}$, using the smallest nonnegative c whose seed is still free. This deterministic linear probing resolves collisions without dropping objects; at most 2^{32} distinct keys are permitted.

Save the full key-to-seed mapping, counters and registry identity before generation. Every paired lookup reuses this table; unknown keys or changed scope require a newly frozen registry rather than an on-the-fly extension. Audit transition keys include the full parent video key, never its truncated/resolved seed. Input-view selection and direct full-hash corruption streams remain separate from uint32 seed allocation. Statistical aggregation remains at the object level.

Separate sampler populations. For controlled AO-Flow comparisons, `eval_det_v13` uses the 20-step deployment UniPC sampler; its initial latent remains random. `eval_stochastic_v13` uses the scored 20-transition policy on the same checkpoints, observations and paired initial/downstream streams. Each sampler population is reported separately with four draws. Native system references retain their own declared samplers, with unused streams marked inapplicable; equal seed indices do not equate latent tensors across different architectures. The four-candidate, 40-object reference-assisted audit is `audit_stochastic_v13`: it uses the stochastic policy, the prescribed first receiver replica and the registry-bound audit keys of Appendix G. Its seed indices are candidate indices, not deployment draw records; phases share their seeds and receive newly generated assets.

Unselected validation structure. The validation structure population uses the same 40 reference-selected validation objects, one fixed input, replica r1 and draw zero. For each Global-D or Region-D receiver, generate one initializer and one final-policy UniPC asset per object, without candidate selection or replacement. This is 80 attempts per receiver pair, 160 for the two primary pairs; it provides a single-replica validation diagnostic. Bind the same reference-only checklist and paired initial/downstream streams, retaining every failure.

Held-out test structure. The structure columns in Tables 3 and 4 use a distinct 100-object subset of the primary test split. Eligibility is determined from reference assets: part completeness, required relations and extra-part inspection must all be applicable. Sort eligible IDs by the SHA-256 digest of UTF-8 `aoflow-test-structure-v113/` followed by the object ID, break ties by ID, and take the first 100 before generating any method output. Save the full eligible inventory and chosen IDs separately from the 40-object validation audit. Reuse each method's ordinary replica-r1/draw-zero deployment asset, with no best-of-group selection. Two raters inspect anonymized assets in independently randomized orders; disagreements are adjudicated without revealing method identity. A pass requires all reference-defined part slots and required relations to be satisfied with no extra or duplicated part. Generation failures count as failures among all 100 objects. Annotation and infrastructure errors are resolved with their object records retained. The checklist is never used as a training reward or checkpoint selector. Agreement and per-object decisions accompany the structure records.

Common asset inspection. Every complete-asset consumer first checks the same C8 set: eight azimuths spaced by 45° from the input direction, including the input camera, under the prescribed elevation, resolution and 4×4 subpixel grid. Check every covered nearest-surface subpixel, including coverage below the final binary-mask threshold. The nominal raster budget is $8 \times 512^2 \times 16$; actual material-hit queries are counted separately. Inspection precedes rewards, metrics and annotation packets, and belongs to their computational account.

Geometry, material and execution assessments are recorded with input/asset, camera/check-set, profile, mapper and renderer identities. Candidate-material failures count as generation failures; mapper, I/O and reference-data errors are resolved before evaluation. Inspection covers C8 only. An expanded inspection set is evaluated on all paired outputs and reported separately.

Primary complete-asset utility retains zero bounded metrics for generation failures. Separately report geometry diagnostics whenever geometry is valid, including material-only failures, with geometry and material failure counts and their metric denominators. Gray geometry views are explicit diagnostics and never repair the full-asset status. Legal absent UV/PBR components use registered defaults; a broken declared dependency does not acquire that exemption.

L.2 IMAGE FORMATION AND PERTURBATIONS

Reference inputs and asset renders share deterministic opaque-surface rasterization, the prescribed cameras, unit-diagonal coordinates and a 512-pixel canvas. Triangle depth is positive camera z, with nearest-depth visibility, no back-face culling, near/far planes $.01/10$, and a fixed 4×4 subpixel grid per pixel. Coverage gives alpha; binary foreground is $\alpha \geq .5$. Reference visibility instead uses the existing 1024-pixel center-sampled depth buffer and tolerance .002, without supersampling or boundary exclusion. Ties choose the earlier stored triangle; no per-output recentering, exposure, camera or geometry fitting is permitted.

Native voxel attributes on the evaluation surface. Keep the decoder's vertices/faces and native postprocessing; add no evaluation remeshing, UV unwrap or texture bake. Geometry is rendered after its fixed native-to-AO transform. At each covered subpixel, perspective-correct barycentric weights interpolate the original native-space vertices of that visible triangle. Material queries use these native positions before any AO transform or γ scaling. Query voxel attributes at $g = (x_{\text{native}} - o)/s$, with stored origin o and voxel size s. Centers lie at $o+(k+.5)s$; the eight-neighbor cell starts at $\lfloor g-.5 \rfloor$. Drop absent/out-of-range neighbors and renormalize remaining trilinear weights, matching the pinned sparse sampler. No extra half-voxel offset is applied before its query.

A constant-one support channel shares this query; require its returned value to be at least $1 - 10^{-5}$. Thus genuine black RGB is distinct from missing support. A verified mapper's unsupported candidate fragment is an invalid-material asset failure under the existing failure rules. Coordinate/version/implementation mismatches are infrastructure errors; unsupported reference materials are data errors. Both are resolved before candidate evaluation. The companion profile fixes the sampler revision and diagnostic-channel packing.

Interpolate stored nonlinear RGB, clamp, and apply the exact sRGB inverse transfer once. This common evaluator convention replaces the native renderer's approximate $\mathrm{RGB}^{2.2}$ conversion. Use linear base color with shader $C\,[.6 + .4 \max(n^\top l, 0)]$, fixed world-space $l = (1, -1, 2)/\sqrt{6}$, unit interpolated normals oriented toward the camera, and no shadows, specular term, tone mapping or ambient occlusion. For UV/vertex-color assets, decode each sRGB texture texel before linear-domain bilinear filtering, then multiply the linear material factor and perspective-interpolated linear vertex color. Missing UVs use the linear texture mean; representation-specific defaults are fixed in the profile. These rules do not override voxel colors. Paired known-color-field fixtures quantify the difference between nonlinear-domain voxel interpolation and linear-domain UV interpolation. Metallic, roughness and material alpha do not alter opaque coverage. Composite over white and average subpixels in linear RGB, then apply sRGB transfer once. Gray audit views substitute constant base color only. The same material adapter serves asset-reward and evaluation renders; its work belongs to their render account. Store depth, alpha and float sRGB separately from quantized model inputs.

The six corruption conditions remain noise standard deviations $\{.03, .06\}$, blur $\sigma \in \{1, 2\}$ pixels, and foreground occlusion fractions $\{.05, .15\}$. They act on the clean 512-pixel sRGB input with foreground B from its saved alpha. Noise adds the same object-keyed standard-normal field at either strength inside B, followed by clipping to $[0, 1]$. Foreground blur uses the separable discrete Gaussian truncated at $\lceil 3\sigma \rceil$:

$$I_\sigma(p) = \frac{[G_\sigma * (BI)](p)}{[G_\sigma * B](p)} \quad (p \in B), \tag{47}$$

with zero extension, leaving pixels outside B unchanged. The denominator is positive inside B; this avoids background leakage into thin parts. For occlusion, choose one foreground center by an object-keyed uniform hash. Sort foreground pixels by squared distance to it, then row/column for ties, and replace the first $\lfloor f|B| + 1/2 \rfloor$ with white. This gives an exact-area disk clipped to the foreground; the two strengths share a center and nested masks. Empty foreground is a data error. Save the clean input, perturbed input and occluder mask; the reference geometry and camera are unchanged.

After corruption, the single quantizer in Appendix K produces model/cache bytes; clean and perturbed float targets retain separate identities. Primary corruption preservation uses the clean target and clean mask. The separately named actual-input diagnostic excludes the occluder in

both its scoring support and denominator. Evaluation LPIPS uses frozen AlexNet v0.1 with `spatial=True`, full-canvas white-composited sRGB mapped to $[-1, 1]$, no bounding-box crop, and a bilinear spatial-map resize with half-pixel centers and no corner alignment. Set pixels outside the target/render-mask union to white in both images, then average the resized map over that union. Empty support is an evaluator error; generated-asset failures retain the distinct metric rules in Appendix A.

L.3 Geometry-control scale and rotation

For geometry-video scoring, normalize every depth and camera center by one median s of finite positive foreground depths across all 13 frames. Empty scale support gives zero geometry terms; use $\bar{d} = 1$ after normalization and never normalize views independently. At 128 pixels use valid-weight bilinear depth resizing and the matching half-pixel intrinsic transform. Holes remain invalid, and the nearest-depth 3×3 reprojection splat assigns missing coverage error one.

For camera-to-world rotations A_i, accumulated rotation is

$$\Omega = \sum_{i=0}^{11} \arccos\left[\operatorname{clip}\left(\frac{\operatorname{tr}(A_{i+1}A_i^\top) - 1}{2}, -1, 1\right)\right]. \tag{48}$$

A full turn gives 2π, unlike endpoint rotation on a closed orbit. Translation length is $\sum_i \|\widetilde{c}_{i+1} - \widetilde{c}_i\|/s$ for original estimator centers. Angular smoothness uses principal SO(3) increments in this world frame; static sequences fail the motion guards.