GEOTRANS: GEOMETRY-AWARE STATE TRANSPORT FOR TRAINING-FREE WORLD MODEL ACCELERATION

Anonymous authors
Paper under double-blind review

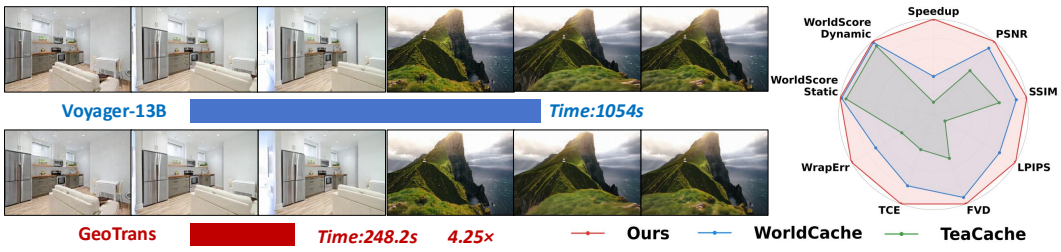


Figure 1: GeoTrans closely matches full-step Voyager-13B generation while substantially reducing denoising time. The normalized radar chart, where larger values indicate better performance, summarizes the speed–quality balance of GeoTrans against competitive methods.

ABSTRACT

Diffusion-based world models generate controllable videos through iterative denoising, but repeated transformer evaluations over long spatiotemporal sequences make inference expensive. Training-free caching exploits temporal redundancy, yet temporal prediction alone does not capture the shared state evolution of geometrically corresponding tokens. Our analysis reveals that geometric correspondence organizes coherent state changes while substantial residual variation remains, motivating geometry-aware latent state transport. We propose GeoTrans, a training-free acceleration framework that determines when to refresh anchors, how to construct geometric transport, and what state update to accept. GeoTrans adaptively refreshes anchors using temporal drift and a horizon limit, decomposes proposed displacements into geometric shared and residual components, and estimates their gains online in closed form from completed anchor transitions. Geometry defines the transition structure, while anchor transitions determine the accepted update, requiring no additional transformer evaluations or gradient computation for gain estimation. Experiments on Voyager and Aether achieve $4.25\times$ and $1.96\times$ acceleration, respectively, with improved generation fidelity over WorldCache. Controlled ablations further isolate the benefit of geometric state transport.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021) support high-fidelity image (Rombach et al., 2022) and video generation (Ho et al., 2022; Kong et al., 2024; Yang et al., 2025), with world models extending generation to scene simulation and exploration (Bruce et al., 2024; Duan et al., 2025). Camera-conditioned models such as HunyuanWorld-Voyager (Huang et al., 2025) use depth, camera rays, and scene layouts to generate sequences along specified viewpoint trajectories. Their billion-scale Diffusion Transformers (DiTs) (Peebles & Xie, 2023), however, repeatedly process long spatiotemporal sequences over dozens of denoising steps, making inference costly.

Acceleration methods reduce sampling steps or model-evaluation cost. Progressive and consistency distillation (Salimans & Ho, 2022; Song et al., 2023) require additional training, whereas DPM-Solver (Lu et al., 2022) accelerates sampling without retraining. Training-free caching exploits denoising redundancy across U-Nets (Ma et al., 2024b), DiTs (Selvaraju et al., 2024; Chen et al., 2024), and video models (Lv et al., 2025; Kahatapitiya et al., 2025). WorldCache (Feng et al., 2026) applies curvature-guided heterogeneous token prediction to diffusion world models.

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.

Temporal redundancy, however, does not specify how updates to geometrically corresponding tokens should be coordinated. World models provide scene correspondences through camera and geometry conditioning. Tokens associated with the same scene region may differ in appearance and feature values while sharing part of their denoising evolution. We argue that skipped-step inference in world models is better formulated as *geometry-aware latent state transport*, with scene correspondence defining how token updates are coupled. This formulation captures shared evolution while preserving token-specific variation and determining how much predicted displacement to accept.

Our analysis on Voyager and Aether (Sec. 3) shows: temporal proposals increasingly overestimate displacement magnitudes at longer horizons, while geometric correspondence reveals shared changes alongside substantial residual variation. These findings motivate bounded prediction horizons, geometric displacement decomposition, and anchor-based calibration of accepted updates.

We propose **GeoTrans** (**Geo**metry-Aware State **Trans**port), a training-free acceleration framework that models skipped-step inference as geometry-conditioned transitions between world-model inference states. GeoTrans addresses three questions: *when to refresh anchors, how to construct geometric transport, and what state update to accept*. Adaptive anchor refresh combines shadow drift with a horizon limit to bound prediction intervals. Geometric transport decomposes proposed displacements into orthogonal shared and residual components, coupling corresponding tokens while preserving token-specific variation. Anchor calibration estimates separate gains from completed full-anchor transitions in closed form, controlling the magnitude and direction of each component. Geometry thus defines the transition structure, while anchor transitions determine the accepted update. The pretrained model and sampler remain unchanged; gain estimation requires no offline fitting, additional transformer evaluations, or gradient computation.

We evaluate GeoTrans on HunyuanWorld-Voyager-13B and Aether-5B (Zhu et al., 2025) using WorldScore (Duan et al., 2025). GeoTrans achieves $4.25\times$ acceleration on Voyager with higher PSNR and WorldScore scores than WorldCache. On Aether, it improves generation fidelity and reconstruction accuracy at $1.96\times$ and $3.05\times$ acceleration, respectively. Controlled closed-loop ablations with matched FULL-step schedules isolate the effect of the state-transition operator.

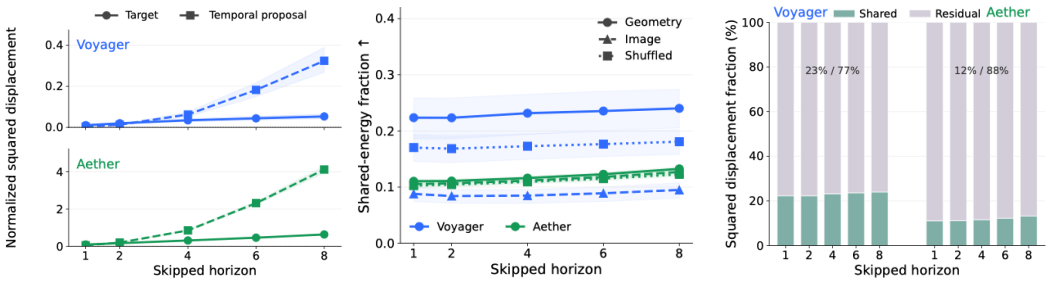
Our contributions are summarized as follows.

- We identify shared and residual structure in world-model state changes, motivating geometry-aware state transport that couples token updates through scene correspondence.
- We propose GeoTrans, a training-free framework combining adaptive anchor refresh, geometric displacement decomposition, and closed-form anchor calibration for skipped-state prediction.
- We demonstrate improved speed and fidelity over cache baselines on Voyager and Aether, with up to $4.25\times$ acceleration. Controlled ablations isolate the benefit of geometric transport.

2 Related Work

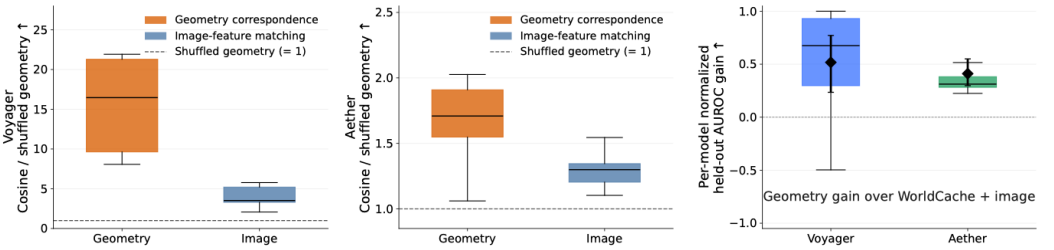
Diffusion world models. Diffusion models (Ho et al., 2020; Song et al., 2021) are widely used for image (Rombach et al., 2022) and video generation (Ho et al., 2022; Kong et al., 2024; Yang et al., 2025). Recent works extend them toward world simulation by conditioning on camera trajectories, depth, actions, or scene geometry (Bruce et al., 2024; OpenAI, 2024; Duan et al., 2025; Agarwal et al., 2025; Lu et al., 2024; Xiang et al., 2024). DiT-based systems (Peebles & Xie, 2023), such as Voyager (Huang et al., 2025) and Aether (Zhu et al., 2025), improve photorealistic and geometry-aware video generation, while interactive and game-oriented world models explore related simulation settings (Valevski et al., 2024; Alonso et al., 2024; Che et al., 2025; Bruce et al., 2024). However, their high inference cost limits practical deployment.

Training-free acceleration for diffusion inference. Training-based acceleration reduces sampling cost through distillation, consistency training, or improved solvers (Salimans & Ho, 2022; Song et al., 2023; Lu et al., 2022), but requires additional optimization or training access. Training-free methods instead exploit redundancy in frozen denoising trajectories, including U-Net feature reuse (Ma et al., 2024b), DiT block or residual caching (Selvaraju et al., 2024; Chen et al., 2024; Liu et al., 2024), attention broadcasting (Zhao et al., 2025), video-DiT cache adaptation (Lv et al., 2025; Liu et al., 2025a; Kahatapitiya et al., 2025), and learned cache routing (Ma et al., 2024a). Most of



(a) Finite-horizon state displacement (b) Shared displacement by grouping (c) Shared and residual components

Figure 2: **State evolution and geometric displacement structure.** (a) Target and WorldCache proposal squared displacements, normalized by the anchor state's squared norm. Proposal magnitudes increasingly exceed their targets at longer horizons. (b) Target squared displacement captured by group means under geometry, image, and shuffled grouping. (c) Geometry-based shared and residual fractions: shared evolution accounts for only part of the state change.



(a) Displacement coherence in Voyager (b) Displacement coherence in Aether (c) Incremental predictive value of geometry

Figure 3: **Geometry provides additional transport information.** (a,b) Within-group displacement cosine similarity normalized by the same-scene shuffled baseline (dashed line at one). (c) Held-out AUROC gains from adding geometry to temporal and image cues for predicting high proposal error. Scene-level gains are averaged over $h = 4, 8$ and scaled by each model's maximum absolute gain. Zero indicates no improvement; scales differ across models.

these methods target generic image or video generation and do not explicitly exploit geometry- and camera-conditioned latent dynamics in diffusion world models.

World-model-specific cache acceleration. WorldCache (Feng et al., 2026) uses curvature-guided heterogeneous token prediction for diffusion world models, while HERO (Song et al., 2025) exploits layer-wise temporal stability through hierarchical extrapolation and refresh. These methods show that world-model acceleration benefits from structured temporal variation. GeoTrans shares the training-free setting, but shifts from cache reuse to geometry-aware latent state transport by explicitly modeling when to refresh anchors, how to construct transport using geometric correspondences, and what state update to accept through anchor calibration.

3 From Cache Reuse to Latent State Transport

We analyze full-step trajectories from Voyager and Aether to motivate geometry-aware state transport, using scene correspondence to structure state changes across skipped denoising steps.

Observation 1: Temporal proposals exhibit horizon-dependent mismatch. For an anchor at step a, the target displacement $Y_{a,h} = H_{a+h} - H_a$ increases with the skipped horizon h, while the temporal proposal increasingly overestimates its magnitude (Fig. 2(a)). This mismatch motivates anchor refresh to bound prediction horizons and calibration to control accepted updates.

Observation 2: Geometry reveals shared state evolution. Geometric groups exhibit higher displacement coherence than image-based and shuffled groups (Fig. 3(a,b)). On Voyager, geometric grouping also captures substantially more target squared displacement in the group-mean component than appearance-based matching (Fig. 2(b)). Geometry further provides incremental information for identifying large temporal-proposal errors beyond temporal and image cues (Fig. 3(c)). Together,

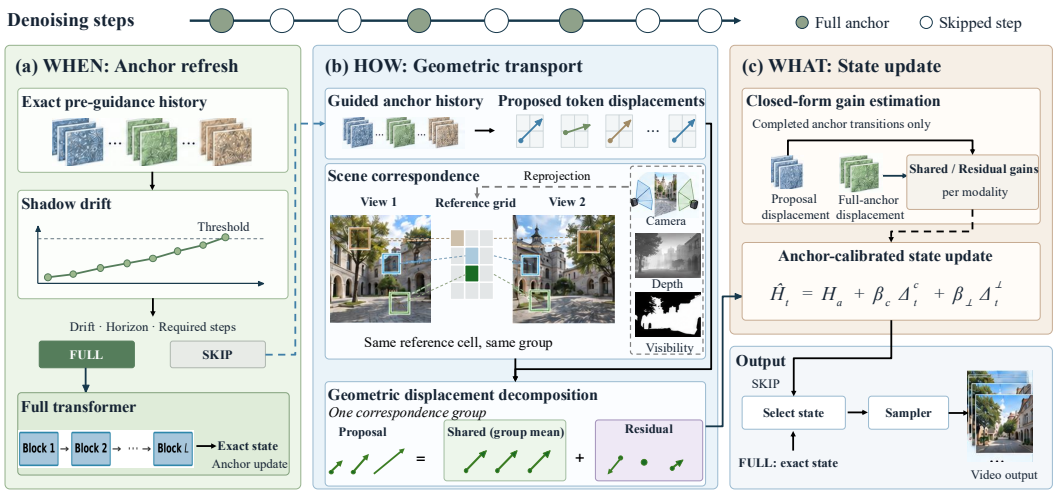


Figure 4: **Overview of GeoTrans. (a) Anchor refresh.** Accumulated shadow drift, a horizon limit, and required evaluations determine full transformer steps that refresh state anchors. **(b) Geometric transport.** Guided anchor history predicts token displacements. Camera geometry, depth, and visibility map corresponding tokens to reference-grid cells, defining groups whose mean displacements form the shared component; the remainder forms the residual. **(c) State update.** Matching proposed displacements to full-anchor transitions from previously completed anchor intervals yields shared and residual gains per modality online in closed form. At subsequent skipped steps, the weighted components update the latest anchor state for the original sampler, without additional transformer evaluations or gradient computation. Scene images and drift curves are schematic.

these diagnostics motivate using scene correspondence to organize the shared component of state transport.

Observation 3: Shared evolution coexists with substantial residual changes. At $h = 4$, geometric group means capture approximately 23% of target squared displacement on Voyager and 12% on Aether (Fig. 2(c)). The remaining state changes motivate retaining a residual component alongside the geometric shared component. Together with the temporal prediction mismatch, this motivates estimating separate gains from completed anchor transitions, allowing the two components to contribute differently to the accepted state update at skipped steps.

4 METHOD

As illustrated in Fig. 4, GeoTrans models skipped denoising steps as *geometry-aware state transport*. The framework determines *when* to observe a new anchor, *how* to construct transport using geometric correspondences, and *what* state update to accept through anchor calibration.

4.1 STATE TRANSPORT FORMULATION

Let t index denoising steps in inference order. We use the guided feature tensor at the model–sampler cache interface as the operational inference state. A full evaluation yields

$$H_t^{\mathrm{FULL}} = F_t^{\mathrm{u}} + s_{\mathrm{cfg}}\big(F_t^{\mathrm{c}} - F_t^{\mathrm{u}}\big) \in \mathbb{R}^{N\times C}, \tag{1}$$

where F_t^{u} and F_t^{c} are the unconditional and conditional features, s_{cfg} is the guidance scale, and N and C denote the token count and channel dimension.

A full evaluation yields the observed state H_t^{FULL}. For a skipped step, let $a < t$ be the latest full anchor and $\Delta_t = \widehat{H}_t^{\mathrm{prop}} - H_a$ the displacement proposed by a temporal predictor. GeoTrans submits

$$\widehat{H}_{t,i} = \begin{cases} H_{t,i}^{\mathrm{FULL}}, & t \in \mathcal{A}, \\ H_{a,i} + (B_{\sigma,t}\Delta_t)_i, & t \notin \mathcal{A},\ i \in \Omega, \\ \widehat{H}_{t,i}^{\mathrm{prop}}, & t \notin \mathcal{A},\ i \notin \Omega, \end{cases} \tag{2}$$

where $\mathcal{A}$ is the set of full-evaluation steps, Ω contains the tokens eligible for transport, and $B_{\sigma,t}$ is a transition operator conditioned on geometric correspondences σ. The following sections specify the anchor selection rule, the geometric structure of $B_{\sigma,t}$, and its anchor-calibrated gains.

4.2 WHEN: ADAPTIVE ANCHOR REFRESH

GeoTrans selects anchor refresh steps using a shadow predictor built from exact pre-guidance feature pairs $(F^{\mathrm{u}}, F^{\mathrm{c}})$ recorded at full evaluations. This shadow supplies the temporal drift statistic for refresh decisions, while guided anchor states separately provide the proposed displacement Δ_t.

Initialization and terminal steps, as well as steps without sufficient prediction history, are always evaluated in full. At any other step t, let $a < t$ be the latest full anchor. Let $e_s^{\mathrm{sh}} \geq 0$ denote the curvature-weighted change between consecutive shadow predictions on high-curvature tokens at a skipped step s. The accumulated drift available before the decision at t is

$$R_t^{\mathrm{sh}} = \sum_{a<s<t} e_s^{\mathrm{sh}}. \tag{3}$$

Outside the mandatory cases above, GeoTrans triggers a full evaluation when

$$(t - a > n_{\max}) \ \vee \ (R_t^{\mathrm{sh}} \geq \tau), \tag{4}$$

and skips the step otherwise. The horizon $n_{\max}$ bounds the interval without observation, while the threshold τ permits earlier refreshes in response to accumulated drift.

A full evaluation supplies the guided state for gain estimation and anchor renewal, appends the pre-guidance feature pair to the shadow history, and resets the accumulated drift. Refresh decisions use the shadow's temporal drift without directly using the accepted displacement $B_{\sigma,t}\Delta_t$ or the gain-fitting loss. The drift statistic and fixed settings are detailed in Appendix A.

4.3 HOW: STATE TRANSPORT IN GEOMETRIC COORDINATES

Geometrically corresponding tokens can share a denoising displacement component while retaining token-specific variation. For $m \in \{\mathrm{RGB}, \mathrm{depth}\}$, let $\Omega_m \subseteq \Omega$ denote its active tokens. Maps, projections, and gains are modality-specific; their modality indices are omitted below.

Temporal proposal and scene correspondence. We obtain the displacement Δ_t by temporal prediction from exact guided anchor states. Following depth-based reprojection (Tang et al., 2023; Lin et al., 2025), we use camera intrinsics, poses, and rendered depth to back-project token locations into 3D and reproject them onto a reference-view patch grid. Visibility restricts matches, yielding a map σ that assigns each valid token i to a cell $s = \sigma(i)$. This map is computed once per sequence. We retain groups $\mathcal{G}_s = \{i \in \Omega_m : \sigma(i) = s\}$ with $|\mathcal{G}_s| \geq 2$ and denote their union by $\mathcal{C}$. The temporal predictor and correspondence construction are detailed in Appendix A.

Geometric displacement decomposition. For a displacement field D, the shared component of group $\mathcal{G}_s$ is its least-squares consensus:

$$\overline{D}_s = \operatorname*{argmin}_{z \in \mathbb{R}^C} \sum_{i \in \mathcal{G}_s} \|D_i - z\|_2^2 = \frac{1}{|\mathcal{G}_s|} \sum_{i \in \mathcal{G}_s} D_i. \tag{5}$$

Assigning this consensus to each grouped token defines the projection

$$(P_\sigma D)_i = \begin{cases} \overline{D}_{\sigma(i)}, & i \in \mathcal{C}, \\ 0, & i \in \Omega_m \setminus \mathcal{C}. \end{cases} \tag{6}$$

For displacement fields U, V, define $\langle U, V \rangle = \mathrm{mean}_{i \in \Omega_m,\, 1 \leq \ell \leq C}(U_{i\ell} V_{i\ell})$ and $\|U\|^2 = \langle U, U \rangle$. Then P_σ orthogonally projects onto fields that are constant within each retained group and zero outside these groups, yielding

$$\Delta_t^{\mathrm{c}} = P_\sigma \Delta_t, \qquad \Delta_t^{\perp} = (I - P_\sigma)\Delta_t, \qquad \langle \Delta_t^{\mathrm{c}}, \Delta_t^{\perp} \rangle = 0, \tag{7}$$

where I is the identity operator. The shared component Δ_t^{c} captures group consensus; the residual $\Delta_t^{\perp}$ retains within-group variation and all displacements outside the retained groups.

Geometry-conditioned transition. We assign signed gains $\beta_{c,t}$ and $\beta_{\perp,t}$ to the shared and residual components, respectively:

$$B_{\sigma,t} = \beta_{c,t} P_\sigma + \beta_{\perp,t}(I - P_\sigma), \qquad \beta_{c,t}, \beta_{\perp,t} \in [-1, 1]. \tag{8}$$

Equivalently, the accepted displacement is

$$B_{\sigma,t}\Delta_t = \beta_{\perp,t}\Delta_t + (\beta_{c,t} - \beta_{\perp,t}) P_\sigma \Delta_t. \tag{9}$$

This update combines uniform displacement scaling with a geometry-dependent correction to the shared component. Orthogonality and bounded gains ensure $\|B_{\sigma,t} D\| \le \|D\|$.

4.4 WHAT: ANCHOR-CALIBRATED STATE UPDATE

GeoTrans determines the accepted state update by estimating shared and residual gains from transitions between full anchor states. The gains are obtained online in closed form and control the magnitude and direction of subsequent transport updates.

Calibration from full anchors. Let b be a new full anchor and $a < b$ its preceding full anchor. Before adding H_b^{FULL} to the history, we construct the temporal proposal at b using only earlier guided anchors. This yields the proposed and target displacements over the same anchor interval:

$$D_b = \widehat{H}_b^{\mathrm{prop}} - H_a, \qquad Y_b = H_b^{\mathrm{FULL}} - H_a, \qquad k_b = b - a. \tag{10}$$

For each modality, we fit $B = \beta_c P_\sigma + \beta_\perp (I - P_\sigma)$ by solving

$$(\beta_c^\star, \beta_\perp^\star) = \underset{\beta_c, \beta_\perp \in [-1,1]}{\operatorname{argmin}} \sum_j \frac{w_j}{k_j^2} \left[\|BD_j - Y_j\|^2 + \lambda \|BD_j\|^2 \right]. \tag{11}$$

Here j indexes completed anchor intervals with valid temporal proposals, k_j denotes their span, and $w_j \ge 0$ is an exponentially decaying recency weight. The factor k_j^{-2} normalizes both displacements by span, while a fixed $\lambda \ge 0$ controls the squared ℓ_2 penalty on accepted displacements. Each modality shares two gains across tokens, one for the shared component and one for the residual.

Closed-form solution. Decompose D_j and Y_j as in Eq. 7, with components D_j^r, Y_j^r for $r \in \{c, \perp\}$. Orthogonality separates Eq. 11 into two scalar convex quadratics with sufficient statistics

$$n_r = \sum_j \frac{w_j}{k_j^2} \langle D_j^r, Y_j^r \rangle, \qquad d_r = \sum_j \frac{w_j}{k_j^2} \|D_j^r\|^2. \tag{12}$$

The derivative is $2[(1+\lambda) d_r \beta_r - n_r]$. For $d_r > 0$, its unique constrained minimizer is

$$\beta_r^\star = \mathrm{clip}_{[-1,1]} \left(\frac{n_r}{(1+\lambda) d_r} \right), \qquad r \in \{c, \perp\}, \tag{13}$$

where $\mathrm{clip}_{[-1,1]}$ clips to $[-1, 1]$. Each gain is a regularized scalar regression coefficient mapping proposed displacements to their targets. For $d_r = 0$, we choose the minimum-magnitude solution $\beta_r = 0$. Recursive updates and fixed settings are detailed in Appendix A.

Accepted state update. At a skipped step t, the latest gains estimated from earlier anchors determine the active-token update by scaling the shared and residual displacements:

$$\widehat{H}_{t,i} = H_{a,i} + \beta_{c,t}(\Delta_t^c)_i + \beta_{\perp,t}(\Delta_t^\perp)_i, \qquad i \in \Omega_m. \tag{14}$$

The gain difference $\beta_{c,t} - \beta_{\perp,t}$ controls the geometric correction in Eq. 9. Tokens outside Ω retain the temporal proposal, as in Eq. 2. The resulting state passes through the existing model interface to the original sampler. GeoTrans remains training-free: gain estimation reuses scheduled full evaluations without offline fitting, additional transformer evaluations, or gradient computation.

Table 1: Image-to-world generation results on HunyuanWorld-Voyager-13B, averaged over the full 2,000-scene WorldScore benchmark. Bold indicates the best results among acceleration methods. Methods marked with * require CPU offloading.

Method	Benchmark Evaluation		Generation Fidelity			Acceleration	
	WorldScore Static↑	WorldScore Dynamic↑	PSNR↑	SSIM↑	LPIPS↓	Latency (s)↓	Speedup↑
HunyuanWorld-Voyager-13B (Huang et al., 2025) ($512 \times 768,\ \text{frames} = 49$)							
Voyager	66.28	46.40	∞	1.000	0.000	1053.7	1.00×
TaylorSeer*	62.46	43.72	18.32	0.615	0.293	1195.2	0.88×
HiCache*	63.80	44.66	18.56	0.623	0.281	1100.1	0.96×
TeaCache	60.88	42.61	16.25	0.565	0.372	311.5	3.38×
EasyCache	64.16	44.91	21.76	0.737	0.208	294.5	3.58×
HERO*	62.37	43.67	17.71	0.601	0.315	1100.0	0.96×
WorldCache	64.89	45.43	23.50	0.770	0.175	288.6	3.65×
GeoTrans	**65.12**	**45.59**	**23.83**	**0.791**	**0.166**	**248.2**	**4.25×**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models. We evaluate frozen HunyuanWorld-Voyager-13B (Huang et al., 2025) and Aether-5B (Zhu et al., 2025). Voyager uses MoGE depth (Wang et al., 2025), camera rays, and scene layouts for conditioning; Aether is a geometry-aware CogVideoX-based world model. We additionally evaluate the frozen LingBot-World-14B (Gao et al., 2026) for few-step causal generation.

Benchmark. Main generation results use the full WorldScore benchmark (Duan et al., 2025). Latency–quality sweeps use a fixed 200-sample subset shared across methods and configurations.

Metrics. We report WorldScore-Static and WorldScore-Dynamic for world-generation performance. Following WorldCache (Feng et al., 2026), frame-level fidelity is measured against each model's full-step output using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). We additionally report FVD (Unterthiner et al., 2018) for video-level distributional fidelity, and TCE and Warp Error (Lai et al., 2018; Huang et al., 2024) for temporal consistency. For Aether reconstruction, we evaluate depth accuracy using AbsRel and threshold accuracies ($\delta < 1.25$ and $\delta < 1.25^2$), and camera-pose accuracy using ATE and translational/rotational RPE. Efficiency is measured by latency and speedup relative to the corresponding full-step model.

Baselines. We compare with representative training-free cache methods: Vanilla, TaylorSeer (Liu et al., 2025b), HiCache (Feng et al., 2025), TeaCache (Liu et al., 2025a), EasyCache (Zhou et al., 2025), HERO (Song et al., 2025), and WorldCache (Feng et al., 2026). For WorldCache, we evaluate $n_{\max} \in \{5, 6, 7, 8\}$ using the default settings.

GeoTrans configuration. GeoTrans combines a maximum skip horizon with accumulated shadow drift to trigger anchor refreshes, while retaining mandatory full evaluations. Shared and residual gains are estimated online from completed anchor transitions. The shadow-drift threshold is calibrated once on 10 pilot scenes from WorldScore and then fixed for all reported experiments; details are provided in Appendix A.4. All experiments run on a single NVIDIA A100 GPU.

5.2 MAIN RESULTS

Overall performance. Table 1 shows that GeoTrans achieves the best speed and generation fidelity among the acceleration methods on HunyuanWorld-Voyager. Compared with WorldCache, it increases speedup from $3.65\times$ to $4.25\times$ while improving PSNR from 23.50 to 23.83 dB. The improvements in SSIM and LPIPS indicate better preservation of structural and perceptual fidelity relative to full-step outputs. Both WorldScore scores also improve, extending the gains to benchmark-level world-generation performance. These results show that GeoTrans reduces inference latency while better preserving the outputs of the original model.

Table 2: Cross-model comparison with WorldCache on Aether-5B and LingBot-World-v2-14B-causal-fast. Speedups are relative to full-step inference for each task; ΔPSNR is relative to WorldCache. LingBot uses four denoising calls per causal chunk. Bold indicates the best results.

Aether-5B image-to-world generation (Zhu et al., 2025)

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	TCE ↓	WarpErr ↓	ΔPSNR (dB)	Speedup ↑
WorldCache	31.88	0.9374	0.0383	3.62	0.946	0.01887	—	1.68×
GeoTrans	**32.05**	**0.9390**	**0.0369**	**3.17**	**0.944**	**0.01885**	**+0.17**	**1.96×**

Aether-5B 3D reconstruction

Method	Abs Rel ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	ATE ↓	RPE trans ↓	RPE rot ↓	Lat(s) ↓	Speedup ↑
WorldCache	0.365	0.462	0.698	0.215	0.072	1.05	18.80	2.95×
GeoTrans	**0.343**	**0.501**	**0.735**	**0.186**	**0.068**	**0.84**	**18.20**	**3.05×**

LingBot-World-v2-14B-causal-fast (15% block-evaluation reduction) (Gao et al., 2026)

Method	Runtime (s) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
WorldCache	17.686	21.032	0.386	0.642
GeoTrans	**17.091**	**22.793**	**0.477**	**0.573**

Cross-model evaluation. On Aether-5B (Table 2), GeoTrans improves generation fidelity while increasing speedup from $1.68\times$ to $1.96\times$ over WorldCache. The gains extend from frame-level metrics to FVD, with slightly lower TCE and unchanged Warp Error at the reported precision. For 3D reconstruction, GeoTrans improves all reported depth and camera-pose metrics at a $3.05\times$ speedup. This extends the evaluation beyond visual fidelity to the accuracy of geometric outputs, supporting its applicability to both generation and reconstruction on a second world model. We additionally evaluate GeoTrans on LingBot-World-v2-14B-causal-fast, which uses four denoising calls per causal chunk. At the matched 15% block-evaluation reduction (Table 2), GeoTrans improves PSNR by 1.76 dB over WorldCache, with higher SSIM and lower LPIPS at lower runtime. These results provide supplementary evidence of improved fidelity under the few-step causal generation setting.

5.3 Visual Comparison

Fig. 5 compares GeoTrans with WorldCache using the first and final frames of each scene. The highlighted regions reveal differences in both appearance and scene structure. GeoTrans exhibits less shadow distortion in the indoor example and fewer unnatural color patches in the trees. In the road example, it better preserves structures that are missing in the WorldCache output. Across these indoor and outdoor examples, GeoTrans better retains local details across the displayed viewpoints, complementing the quantitative fidelity results with localized visual evidence.

Table 3: Anchor-refresh ablation on HunyuanWorld-Voyager. All variants use the same temporal predictor and geometric transport procedure, with identical $n_{\max}$ and τ wherever applicable. Adaptive variants retain mandatory full evaluations. Avg. FULL reports the realized number of full evaluations; the rows represent their natural operating points.

Refresh Policy	Refresh Rule	Avg. FULL ↓	PSNR ↑	Speedup ↑
Uniform	Periodic refresh	14.00	20.79	3.66×
Horizon only	$t - a > n_{\max}$	12.00	23.02	4.41×
Shadow drift only	$R_t^{\mathrm{sh}} \geq \tau$	13.17	23.30	3.91×
GeoTrans	$(\mathbf{t} - \mathbf{a} > \mathbf{n}_{\max}) \vee (\mathbf{R}_\mathbf{t}^{\mathrm{sh}} \geq \tau)$	12.42	**23.83**	**4.25×**

Table 4: Controlled closed-loop ablation of the state-transition operator on HunyuanWorld-Voyager. All variants use the same FULL-step schedule and temporal prediction rule, but evolve their own sampler trajectories. Subsequent anchors and proposals reflect the downstream effect of $B_{\sigma,t}$.

Transition Variant	$B_{\sigma,t}$	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	TCE ↓	WarpErr ↓
Raw temporal proposal	I	22.86	0.785	0.171	12.04	1.249	0.0283
Global calibrated gain	βI	23.21	0.787	0.168	11.38	1.231	0.0270
Shuffled correspondence	$\beta_c P_{\tilde{\sigma}} + \beta_\perp (I - P_{\tilde{\sigma}})$	23.01	0.776	0.169	11.88	1.241	0.0279
GeoTrans	$\beta_c \mathbf{P}_\sigma + \beta_\perp (\mathbf{I} - \mathbf{P}_\sigma)$	**23.83**	**0.791**	**0.166**	**11.36**	**1.228**	**0.0262**

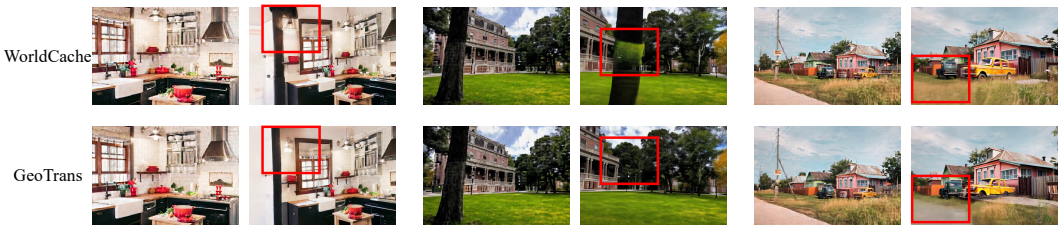


Figure 5: Qualitative comparison of WorldCache and GeoTrans on the first and final frames of each scene. Red boxes highlight local structure and appearance better preserved by GeoTrans.

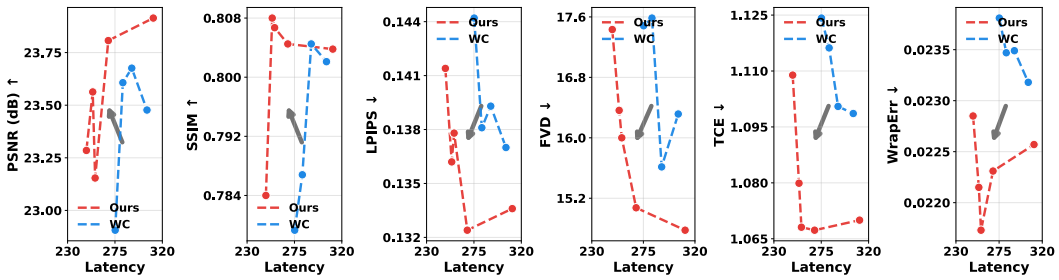


Figure 6: Latency–quality trade-offs between GeoTrans and WorldCache on HunyuanWorld-Voyager. Each point denotes a configuration evaluated on the same 200-samples WorldScore subset, rather than the full benchmark. Gray arrows indicate the preferred direction.

5.4 Ablation Studies

Adaptive anchor refresh. Table 3 compares refresh policies using the same temporal predictor and geometric transport procedure. The combined rule achieves 23.83 dB PSNR at a $4.25\times$ speedup. Compared with horizon-only refresh, it increases the average number of FULL evaluations only slightly (12.42 vs. 12.00) while improving PSNR by 0.81 dB. Compared with drift-only refresh, it achieves both higher fidelity and higher speed. These results characterize the operating-point trade-off obtained by combining a bounded skip horizon with adaptive shadow-drift monitoring.

Geometric state transport and anchor calibration. Table 4 compares operators under a common FULL-step schedule and temporal prediction rule. GeoTrans gains 0.62 dB over global calibration and 0.82 dB over shuffled correspondence, showing that the improvement goes beyond uniform calibration or two-gain structure alone and supports the geometry-defined shared/residual assignment.

Latency–quality trade-off. We evaluate acceleration configurations on a fixed 200-sample WorldScore subset shared across methods. Fig. 6 shows more favorable latency–quality trade-offs for GeoTrans than WorldCache across the evaluated configurations. The gains extend beyond the main operating point, supporting the effectiveness of GeoTrans under different latency budgets.

Sensitivity analysis. We study $n_{\max}$, λ, and $c_{\min}$. Longer horizons trade fidelity for latency, while stricter correspondence filtering improves PSNR at higher latency. The default λ performs best among tested values; full sensitivity results are reported in Appendix F.

6 Conclusion

We presented GeoTrans, a training-free framework that formulates skipped denoising in diffusion world models as geometry-aware state transport. GeoTrans combines adaptive anchor refresh, correspondence-based decomposition into shared and residual displacements, and online closed-form calibration from completed anchor transitions. On HunyuanWorld-Voyager-13B and Aether-5B, GeoTrans improves inference efficiency and generation fidelity over the evaluated cache baselines, achieving up to $4.25\times$ and $1.95\times$ speedups. The benefits extend to depth and camera-pose accuracy in accelerated 3D reconstruction. These results support using scene correspondence to structure predicted state changes and anchor observations to calibrate their contributions, enabling efficient inference without retraining the world model or modifying the sampling procedure.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, including improving wording, grammar, clarity, and readability of the manuscript. All scientific content and AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sec. 4 provides the complete GeoTrans formulation, and Sec. 5.1 specifies the experimental setup, baselines, metrics, and runtime protocol. Appendix A details the implementation, backbone-specific state interfaces, and key hyperparameters, while Appendix B provides the complete online inference procedure. Mathematical derivations and additional experimental protocols, including controlled ablations, baseline configurations, hyperparameter sensitivity, and motivation diagnostics, are provided in the remaining appendices. GeoTrans is training-free and requires no model-weight updates.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *arXiv preprint arXiv:2405.12399*, 2024.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. GameGen-X: Interactive open-world game video generation. In *ICLR*, 2025.

Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Δ-DiT: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *ICCV*, 2025.

Liang Feng, Shikang Zheng, Jiacheng Liu, Yuqi Lin, Qinming Zhou, Peiliang Cai, Xinyu Wang, Junjie Chen, Chang Zou, Yue Ma, et al. Hicache: Training-free acceleration of diffusion models via hermite polynomial-based feature caching. *arXiv preprint arXiv:2508.16984*, 2025.

Weilun Feng, Guoxin Fan, Haotong Qin, Chuanguang Yang, Mingqiang Wu, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Dingrui Wang, et al. Worldcache: Accelerating world models for free via heterogeneous token caching. *arXiv preprint arXiv:2603.06331*, 2026.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. In *ICCV*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018.

Beibei Lin, Tingting Chen, and Robby Tan. Geocomplete: Geometry-aware diffusion for reference-driven image completion. In *NeurIPS*, 2025.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *CVPR*, 2025a.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *arXiv preprint arXiv:2404.02747*, 2024.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. In *ICCV*, 2025b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.

Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille, and Jieneng Chen. GenEx: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024.

Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K. Wong. Fastercache: Training-free video diffusion model acceleration with high quality. In *ICLR*, 2025.

Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. In *NeurIPS*, 2024a.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *CVPR*, 2024b.

OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. Technical report.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. Fora: Fast-forward caching in diffusion transformer acceleration. *arXiv preprint arXiv:2407.01425*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Quanjian Song, Xinyu Wang, Donghao Zhou, Jingyu Lin, Cunjian Chen, Yue Ma, and Xiu Li. HERO: Hierarchical extrapolation and refresh for efficient world models. *arXiv preprint arXiv:2508.17588*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In *ICLR*, 2025.

Xin Zhou, Dingkang Liang, Kaijin Chen, Tianrui Feng, Xiwu Chen, Hongkai Lin, Yikang Ding, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Less is enough: Training-free video diffusion acceleration via runtime-adaptive caching. *arXiv preprint arXiv:2507.02860*, 2025.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *ICCV*, 2025.

APPENDIX

A GEOTRANS IMPLEMENTATION DETAILS

This section specifies the temporal predictor, shadow refresh statistic, correspondence construction, and recursive gain estimation used by GeoTrans. The tensor layout and numerical defaults below describe the Voyager implementation; they are not assumptions about every evaluated backbone. We use denoising-step indices in inference order, independently of the sampler's physical noise or time parameter.

A.1 TEMPORAL PROPOSAL AND SHADOW REFRESH

Separate histories. GeoTrans maintains two histories at FULL evaluations: exact pre-guidance feature pairs $J_a = (F_a^{\mathrm{u}}, F_a^{\mathrm{c}})$ for the shadow refresh controller, and exact guided states H_a for temporal prediction and anchor calibration. Both histories are updated only when a FULL evaluation is performed. The shadow history is used to compute the accumulated drift R_t^{sh} for refresh decisions, whereas the guided-state history provides the temporal proposal $\widehat{H}_t^{\mathrm{prop}}$ and the completed anchor transitions used for gain estimation. Here, an exact anchor denotes the state obtained from a FULL model evaluation on the current sampler trajectory.

Curvature-based temporal prediction. Let $a_0 < a_1 < a$ be the latest three full anchors. For the guided history, define token velocities and curvature as

$$\begin{aligned} v_i^{\mathrm{cur}} &= \frac{H_{a,i} - H_{a_1,i}}{a - a_1}, & v_i^{\mathrm{prev}} &= \frac{H_{a_1,i} - H_{a_0,i}}{a_1 - a_0}, \\ \kappa_i &= \frac{\|(v_i^{\mathrm{cur}} - v_i^{\mathrm{prev}})/(a - a_1)\|_2}{\|v_i^{\mathrm{cur}}\|_2^2 + \epsilon}. \end{aligned} \tag{15}$$

Following curvature-based prediction (Feng et al., 2026), the 0.30 and 0.60 curvature quantiles partition stable, linear, and high-curvature tokens. With $k = t - a$, set $z_k = \min(k/n_{\max}, 1)$ and $\alpha_k = 3z_k^2 - 2z_k^3$. The temporal proposal is

$$\widehat{H}_{t,i}^{\mathrm{prop}} = \begin{cases} H_{a,i}, & \kappa_i < q_{0.30}, \\ H_{a,i} + k v_i^{\mathrm{cur}}, & q_{0.30} \le \kappa_i < q_{0.60}, \\ H_{a,i} + k[(1-\alpha_k) v_i^{\mathrm{cur}} + \alpha_k v_i^{\mathrm{prev}}], & \kappa_i \ge q_{0.60}. \end{cases} \tag{16}$$

Thus $n_{\max}$ enters both the refresh rule and the horizon-dependent velocity interpolation. The temporal predictor supplies the proposal; correspondence-based decomposition and anchor calibration determine the accepted update.

Shadow drift and its normalization. The shadow applies the same temporal rules to the pre-guidance history. Let $\mathcal{I}^{\mathrm{sh}}$ contain all branch/token positions in that history, and let $\mathcal{K}^{\mathrm{sh}} \subseteq \mathcal{I}^{\mathrm{sh}}$ be its high-curvature subset. Denote its high-curvature proposal at age k by $J_i^{\mathrm{h}}(k)$, with $J_i^{\mathrm{h}}(0) = J_{a,i}$. At a skipped step t, the increment is

$$e_t^{\mathrm{sh}} = \frac{1}{|\mathcal{I}^{\mathrm{sh}}|} \sum_{i \in \mathcal{I}^{\mathrm{sh}}} \mathbf{1}\{i \in \mathcal{K}^{\mathrm{sh}}\} \kappa_i^{\mathrm{sh}} \|J_i^{\mathrm{h}}(k) - J_i^{\mathrm{h}}(k-1)\|_2. \tag{17}$$

The mask selects contributions from high-curvature tokens, while the normalization is over the complete shadow token grid. Averaging only over the selected subset would change the units of the threshold. An empty selected set contributes zero.

The decision at t uses only increments from earlier skipped steps, $R_t^{\mathrm{sh}} = \sum_{a<s<t} e_s^{\mathrm{sh}}$. The current increment is committed after choosing SKIP; FULL resets the accumulator. Shadow drift is a temporal proxy and does not require an unavailable full-state target at a skipped step. The controller does not directly read the gain-fitting loss or accepted displacement. Nevertheless, accepted updates affect later sampler inputs and hence future full anchors, so this separation does not imply that refresh schedules are invariant to changes in the transport operator.

Mandatory evaluations and horizon convention. For the 50-step Voyager setting, steps 0–4 and step 49 are mandatory FULL evaluations. Insufficient prediction history also requires FULL. Outside these cases, FULL is triggered when $(t - a > n_{\max}) \vee (R_t^{\mathrm{sh}} \geq \tau)$. With $n_{\max} = 6$ and no drift trigger, ages $1, \ldots, 6$ can be skipped and age 7 requires FULL. Thus the maximum number of consecutive skips and the spacing between full anchors differ by one.

A.2 Reference Correspondences

Depth-based projection. For a patch center $\widetilde{u}_i$ in homogeneous image coordinates at frame f, let d_i be the harmonic mean of its visible, valid rendered depths. With world-to-camera pose T_f and intrinsic matrix K_f, projection into reference frame 0 is

$$X_i^0 = T_0 T_f^{-1} \begin{bmatrix} d_i K_f^{-1} \widetilde{u}_i \\ 1 \end{bmatrix}, \qquad u_i^0 = \pi\big(K_0 [X_i^0]_{1:3}\big), \tag{18}$$

where π performs perspective division. This uses the supplied camera and rendered-depth conditions, rather than outputs from a future denoising step or the full-step reference video.

Patch-grid assignment. The geometry descriptors store horizontal and vertical projected displacements normalized by image width and height. Let these values be δ_i^x, δ_i^y after alignment to the latent token grid. For a token at patch row r_i and column c_i, its reference indices are

$$\widehat{r}_i = \mathrm{round}(r_i + 32\delta_i^y), \qquad \widehat{c}_i = \mathrm{round}(c_i + 48\delta_i^x), \qquad \sigma(i) = 48\widehat{r}_i + \widehat{c}_i. \tag{19}$$

Positive-depth and image-bound tests enter projection support; the rounded cell must also lie within the 32×48 patch grid. We require visible coverage at least $c_{\min} = 0.25$ and aligned projection-support value at least 0.5. Within each modality, a cell is retained only if it contains at least two active tokens. The map is computed once per sequence and held fixed during denoising. These patch-level correspondences provide shared coordinates, rather than exact pointwise surface identities.

Voyager interface and excluded regions. The guided interface tensor has shape $(13, 65, 48, 64)$: 13 latent frames, 32 RGB rows, one separator row, 32 depth rows, 48 columns, and 64 channels. Geometry descriptors are aligned from video frames to latent frames before grouping. The conditioning frame and separator row are outside Ω and retain the temporal proposal. RGB and depth groups and gains are computed separately. Coverage, the two projected-displacement coordinates, and projection support determine group membership.

Active tokens without a valid repeated reference-cell assignment are outside $\mathcal{C}$ but remain inside Ω. For such tokens, $P_\sigma \Delta_t = 0$, so their accepted displacement is $\beta_\perp \Delta_t$. They are not deleted or treated as full-evaluation tokens. In contrast, tokens outside Ω retain $\widehat{H}_t^{\mathrm{prop}}$ without either gain.

A.3 Causal Gain Updates

For each modality and $r \in \{\mathrm{c}, \perp\}$, initialize $n_r = d_r = \beta_r = 0$. At a full anchor b with three earlier full states available, form the proposal from the old guided history, then construct $D_b = \widehat{H}_b^{\mathrm{prop}} - H_a$, $Y_b = H_b^{\mathrm{FULL}} - H_a$, and $k_b = b - a$. Decompose both displacements using the fixed P_σ and update

$$\begin{aligned} n_r &\leftarrow \rho n_r + (1-\rho)\frac{\langle D_b^r, Y_b^r \rangle}{k_b^2}, \\ d_r &\leftarrow \rho d_r + (1-\rho)\frac{\|D_b^r\|^2}{k_b^2}, \\ \beta_r &\leftarrow \mathrm{clip}_{[-1,1]}\left(\frac{n_r}{\max\{(1+\lambda)d_r, 10^{-12}\}}\right). \end{aligned} \tag{20}$$

We use $\rho = 0.5$ and $\lambda = 0.1$. After q eligible observations, the implied recency weights are $w_j = (1-\rho)\rho^{q-j}$ for $j = 1, \ldots, q$. Both statistics share the same weights, so a common EMA bias-correction factor would cancel in their ratio when the numerical floor is inactive. No full-state target from a skipped step is used. The newly fitted gains apply to later skipped steps; the current FULL step submits its observed state unchanged.

The denominator floor is a numerical safeguard. Eq. 13 is the exact minimizer when that floor is inactive. If $d_r = 0$, all positively weighted proposal components are zero, hence $n_r = 0$ and the objective is independent of β_r. The selected zero gain is the minimum-magnitude solution. The initialization convention likewise yields anchor reuse on active tokens until an eligible calibration observation is available.

Span normalization and training-free calibration. The factor k_j^{-2} is equivalent to fitting D_j/k_j to Y_j/k_j. It prevents an interval from receiving quadratic weight solely because both displacements scale with its span; it does not assume that the underlying trajectory is linear. The objective contains no additional normalization by anchor-state energy. Gains are estimated from scheduled full evaluations, without a separate optimization loop or model-weight updates. This online coefficient estimation is distinct from selecting fixed controller hyperparameters.

A.4 Fixed Voyager Settings

Table 5 summarizes the default Voyager settings. The shadow-drift threshold is calibrated once using 10 pilot scenes sampled from WorldScore and is then fixed for all experiments. Calibration uses only the shadow-drift statistic under the horizon-triggered refresh policy and does not use generation-quality metrics or baseline comparisons. We set $\tau = 11.0$ from the median peak shadow drift observed on these pilot scenes. The resulting threshold is fixed before evaluating the full benchmark and is not adapted to individual test scenes. This calibration only determines the operating point of the temporal refresh controller and is separate from the online estimation of the shared and residual gains.

Table 5: Key GeoTrans hyperparameters for the Voyager experiments.

Component	Hyperparameter	Value
Refresh	Maximum skip horizon $n_{\max}$	6
	Shadow-drift threshold τ	11.0
Geometry	Minimum correspondence coverage $c_{\min}$	0.25
Calibration	Recency decay ρ	0.5
	Displacement regularization λ	0.1

B Online Inference Procedure

Algorithm 1 gives the execution order for the state interface in the main method. In particular, observing a new full anchor must not insert that anchor into the history used to predict its own target. Geometry remains fixed, while histories, drift, and gain statistics evolve online.

Implementation cost. The projection is implemented by indexed group sums, counts, and broadcasts; an $N \times N$ matrix is never materialized. For N active tokens, C channels, and K reference-grid cells, a projection costs $O(NC+KC)$ arithmetic with $O(N)$ group indices and $O(KC)$ group accumulators. Decomposition and gain fitting use tensor reductions of the same order, and each modality retains only two numerator/denominator pairs. Three exact observations suffice for each temporal history, so the stored history does not grow with the rollout length. These costs are additional to the frozen model, but require no extra transformer calls. Actual latency includes implementation-dependent memory traffic and synchronization; an operation-count reduction is not itself a wall-clock speedup.

C Properties of Geometric State Transport

C.1 Orthogonal Shared and Residual Coordinates

For a retained group $\mathcal{G}_s$, the consensus objective $\sum_{i \in \mathcal{G}_s} \|D_i - z\|_2^2$ has Hessian $2|\mathcal{G}_s|I$, so its unique minimizer is the group mean. Applying the mean twice changes nothing. Moreover, for arbitrary

Algorithm 1 GeoTrans at the guided model–sampler interface

Require: Frozen model and sampler; T denoising steps; supplied geometry; $n_{\max}, \tau, c_{\min}, \rho, \lambda$
Build modality-specific correspondence projectors P_σ.
Initialize exact shadow/guided histories, $R^{\mathrm{sh}} = 0$, and all gain statistics and coefficients to zero.
for $t = 0, \ldots, T-1$ **do**
  Let a be the latest full anchor, if available.
  Set FULL if initialization, termination, or insufficient history requires it; otherwise use $(t - a > n_{\max}) \vee (R^{\mathrm{sh}} \geq \tau)$.
  if FULL **then**
    Initialize the current calibration proposal as unavailable.
    If guided history is sufficient, predict at t from the existing history and store $D_t = \widehat{H}_t^{\mathrm{prop}} - H_a$.
    Evaluate the full model and construct J_t and H_t^{FULL}.
    if D_t is available **then**
      Form $Y_t = H_t^{\mathrm{FULL}} - H_a$, decompose D_t, Y_t, and update gains using Eq. 20.
    end if
    Append the exact observations to their respective histories; set $a = t$, $R^{\mathrm{sh}} = 0$, and $\widehat{H}_t = H_t^{\mathrm{FULL}}$.
  else
    Form the guided proposal and $\Delta_t = \widehat{H}_t^{\mathrm{prop}} - H_a$.
    Set $\widehat{H}_t = \widehat{H}_t^{\mathrm{prop}}$ on excluded tokens.
    For each active modality, set $\widehat{H}_t = H_a + \beta_{\mathrm{c}} P_\sigma \Delta_t + \beta_\perp (I - P_\sigma)\Delta_t$.
    Compute e_t^{sh} from shadow history and commit $R^{\mathrm{sh}} \leftarrow R^{\mathrm{sh}} + e_t^{\mathrm{sh}}$.
  end if
  Pass $\widehat{H}_t$ through the original sampler interface.
end for
return the generated output using the original decoder.

fields U, V, each group contributes $|\mathcal{G}_s| \langle \overline{U}_s, \overline{V}_s \rangle$ to both $\langle P_\sigma U, V \rangle$ and $\langle U, P_\sigma V \rangle$, up to the common normalization of the field inner product. Unmatched tokens contribute zero. Consequently,

$$P_\sigma^2 = P_\sigma, \qquad P_\sigma^* = P_\sigma, \qquad \langle P_\sigma D, (I - P_\sigma) D \rangle = 0. \tag{21}$$

In particular, $\|D\|^2 = \|P_\sigma D\|^2 + \|(I - P_\sigma)D\|^2$. Shared and residual energy fractions therefore describe a decomposition of squared displacement, rather than competing estimates of prediction accuracy. A large residual fraction does not imply that the shared component is uninformative.

Group consistency without forcing identical updates. For $i \in \mathcal{G}_s$, the accepted displacement is

$$(BD)_i = \beta_{\mathrm{c}} \overline{D}_s + \beta_\perp (D_i - \overline{D}_s). \tag{22}$$

The group mean becomes $\beta_{\mathrm{c}} \overline{D}_s$, while the difference between two tokens in that group is $(BD)_i - (BD)_j = \beta_\perp (D_i - D_j)$. Thus geometry couples shared evolution without requiring corresponding tokens to have identical features or identical accepted updates. Residual variation is retained whenever its gain is nonzero.

C.2 Closed-Form Calibration and Scalar Baselines

Orthogonality separates the regularized objective as

$$\begin{aligned} \|BD_j - Y_j\|^2 &= \sum_{r \in \{\mathrm{c}, \perp\}} \|\beta_r D_j^r - Y_j^r\|^2, \\ \|BD_j\|^2 &= \sum_{r \in \{\mathrm{c}, \perp\}} \beta_r^2 \|D_j^r\|^2. \end{aligned} \tag{23}$$

For each component, the scalar objective is

$$\mathcal{L}_r(\beta_r) = (1 + \lambda) d_r \beta_r^2 - 2 n_r \beta_r + C_r. \tag{24}$$

When $d_r > 0$, its second derivative is positive and its unconstrained minimizer is $n_r/[(1+\lambda)d_r]$. Clipping to $[-1, 1]$ gives the exact constrained minimizer. This is a penalty on accepted *displacement energy*, not a token-level risk penalty or a penalty on a geometry-derived confidence score. A negative gain is possible when proposals and observed displacements have negative aggregate alignment; it reverses that component in feature space.

Relation to uniform calibration. If $\beta_{\mathrm{c}} = \beta_{\perp} = \beta$, then $B = \beta I$. Setting both gains to one recovers the raw proposal on active tokens; setting both to zero gives latest-anchor reuse. For a fixed geometry map, the two-gain operator family therefore contains global scalar calibration. In particular, its minimum regularized calibration objective cannot exceed the minimum obtained under the constraint $\beta_{\mathrm{c}} = \beta_{\perp}$.

More explicitly, when $d_{\mathrm{c}}, d_{\perp} > 0$ and both unconstrained optima $b_r = n_r/[(1+\lambda)d_r]$ lie in $[-1, 1]$, the common-gain optimum is their energy-weighted mean. The objective gap is

$$\min_{\beta} \mathcal{L}(\beta, \beta) - \min_{\beta_{\mathrm{c}}, \beta_{\perp}} \mathcal{L}(\beta_{\mathrm{c}}, \beta_{\perp}) = (1+\lambda)\frac{d_{\mathrm{c}} d_{\perp}}{d_{\mathrm{c}} + d_{\perp}}(b_{\mathrm{c}} - b_{\perp})^2. \tag{25}$$

This identifies when separate calibration adds capacity: the two components can prefer different gains when both carry proposal energy. The result establishes the additional calibration capacity of the two-component operator; its effect on the resulting rollout is evaluated empirically by the controlled operator ablation in Table 4.

C.3 Update Magnitude and Unsupported Tokens

For fixed gains and correspondences,

$$\begin{aligned} \|B_{\sigma,t} D\|^2 &= \beta_{\mathrm{c},t}^2 \|P_\sigma D\|^2 + \beta_{\perp,t}^2 \|(I - P_\sigma) D\|^2 \\ &\le \|P_\sigma D\|^2 + \|(I - P_\sigma) D\|^2 = \|D\|^2. \end{aligned} \tag{26}$$

The same bound applies separately to each active modality. It controls accepted-update magnitude relative to the proposal, not the error against the unknown full state or the stability of the entire nonlinear sampler rollout. These distinctions matter because smaller updates need not always be more accurate.

If no correspondence groups survive, $P_\sigma = 0$ and the operator reduces to residual scalar calibration. If the fitted gains coincide, the geometric correction vanishes regardless of the grouping. Otherwise, $(\beta_{\mathrm{c}} - \beta_{\perp}) P_\sigma D$ modifies only the shared part. This structure permits conservative fallback on unsupported regions without replacing every update by a group mean.

D Additional Interpretation of the Experiments

D.1 Anchor Refresh and Transition Calibration

Refresh policies. Table 3 reports realized, unmatched full-evaluation counts. The combined policy improves PSNR over horizon-only refresh by 0.81 dB, with average FULL count increasing from 12.00 to 12.42 and speedup changing from $4.41\times$ to $4.25\times$. It also improves both PSNR and speedup over drift-only refresh. These comparisons support the observed operating-point trade-off between the two triggers.

Uniform uses a fixed interval of five steps after warm-up, whereas horizon-only refresh uses $n_{\max} = 6$ with the strict condition $t - a > 6$. Their full-evaluation schedules are

$$\begin{aligned} \mathcal{A}_{\mathrm{uniform}} &= \{0, 1, 2, 3, 4, 9, 14, 19, 24, 29, 34, 39, 44, 49\}, \\ \mathcal{A}_{\mathrm{horizon}} &= \{0, 1, 2, 3, 4, 11, 18, 25, 32, 39, 46, 49\}, \end{aligned}$$

giving 14 and 12 FULL evaluations, respectively. Both retain the warm-up and terminal evaluations. The former permits four consecutive skips and the latter six; they are distinct refresh settings. Average FULL counts in the table describe the realized computation of each policy rather than a common target budget.

What the operator comparisons test. The raw proposal $B = I$ tests transport without gain calibration. The scalar operator $B = \beta I$ tests whether uniform correction alone explains the improvement. A shuffled projector retains two calibrated components but removes the original association between tokens and reference cells. The geometry operator tests whether the correspondence structure provides a useful choice of components. The reported PSNR increases from 22.86 to 23.21 dB with scalar calibration, and to 23.83 dB with geometric calibration. The latter also exceeds shuffled correspondence by 0.82 dB. The accompanying fidelity and temporal metrics in Table 4 also favor the geometric operator.

Matched structure in the shuffle control. The correspondence shuffle permutes reference-cell IDs among valid active tokens within each modality. It preserves their ID multiplicities and the valid-token set, then reapplies the minimum-group-size rule. It therefore preserves group sizes and the number of grouped tokens, while changing which tokens share each reference cell. Individual membership in a retained group can change when singleton IDs are also permuted. This control distinguishes geometric assignment from merely providing two gain coefficients or the same group-size distribution.

Controlled closed-loop protocol. All variants use the FULL step indices recorded from the full GeoTrans dynamic run and the same temporal prediction rule. Each variant then runs its own sampler trajectory, recomputing full anchor states and temporal proposals from that trajectory. Gains are fitted from the variant's completed anchor transitions. This controls observation timing and predictor design while measuring how the chosen transition operator affects the resulting generation, including its downstream effect on later states.

D.2 Evaluation Sets, References, and Timing

Full-benchmark and subset results. The main generation comparison uses the WorldScore benchmark, whereas the latency–quality sweep uses a fixed 200-sample subset shared across methods and configurations. Comparisons within each panel therefore use the same evaluation set. Absolute scores from the subset need not match the full-benchmark table, and the two sets are not pooled into one empirical frontier. Frame-level fidelity measures agreement with the corresponding unaccelerated model output; it is distinct from WorldScore's benchmark-level evaluation and from depth/pose accuracy against a reconstruction reference.

Runtime and computation counts. Reported speedup is the full-step runtime divided by accelerated runtime under the corresponding task setting. CPU transfer overhead is included for the offloaded Voyager baselines, as stated in Table 1. Average FULL count and block-evaluation reduction characterize computational activity, but neither alone fixes wall-clock runtime. Predictor reductions, grouping, tensor transfers, and other fixed pipeline costs also contribute.

D.3 Cross-Model and Few-Step Evaluation

Aether. For Aether-5B, GeoTrans operates on the joint latent at the model–sampler interface. The latent contains RGB, depth, and raymap components on the same VAE frame grid, which are mapped by the patch embedding into three aligned token blocks. Depth and raymap provide the geometric information used to construct reference correspondences, with the same patch index representing the same spatial location across the three blocks. The resulting correspondence projector P_σ is applied to the RGB and depth state blocks, while the raymap is used only to define the geometry.

LingBot. LingBot-World-v2-14B-causal-fast evaluates four denoising calls in each causal chunk, followed by a zero-timestep KV update. We apply GeoTrans at an intermediate DiT state before the native prediction head, using the chunk-local state sequence as the prediction history. The first chunk is evaluated in full to initialize the online predictor; in each subsequent chunk, the first three calls are evaluated exactly and the fourth is predicted from the available chunk-local states. The prediction head and KV update remain unchanged and always execute in full. This implementation yields the reported 15% reduction in intermediate DiT evaluations without model-weight updates.

Table 6: Hyperparameter sensitivity on HunyuanWorld-Voyager. † marks the default value of the parameter being varied.

Parameter	Value	PSNR (dB) ↑	Latency (s) ↓
$n_{\max}$	4	24.807	267.5
	5	24.095	255.2
	$6^{\dagger}$	23.812	244.6
	7	23.179	243.2
	8	22.987	234.9
$c_{\min}$	0.10	23.507	237.5
	$0.25^{\dagger}$	23.846	245.2
	0.50	24.458	264.6
λ	0	23.310	236.5
	$0.1^{\dagger}$	23.807	245.2
	1	22.900	254.9

Comparison scope. WorldCache retains its original stable/linear/high-curvature token partition without LingBot-specific retuning. At the matched 15% intermediate-state evaluation reduction, GeoTrans improves PSNR by 1.761 dB, with higher SSIM and lower LPIPS at comparable runtime (Table 2). This result further evaluates GeoTrans under a few-step causal generation setting.

E WorldCache Baseline Scope

The WorldCache latency–quality comparison varies $n_{\max}$ over $\{5, 6, 7, 8\}$ with the remaining baseline settings fixed, as described in Section 5.1. These baseline configurations are separate from GeoTrans's shadow-controller and gain-calibration settings. The main results and the explicitly identified 200-sample sweep provide the empirical comparisons under their respective evaluation protocols.

F Hyperparameter Sensitivity

Table 6 reports PSNR and latency for three parameter sweeps on HunyuanWorld-Voyager: the maximum skip horizon $n_{\max}$, the minimum coverage threshold $c_{\min}$ used to construct correspondence groups, and the gain regularization coefficient λ. The sweeps were executed as independent runs. The marked default entries identify the default parameter values within each sweep, rather than repeated copies of one shared baseline measurement. Comparisons below concern the observed PSNR–latency trade-offs within each sweep.

Skip horizon. Increasing $n_{\max}$ from 4 to 8 reduces latency from 267.5 s to 234.9 s, while PSNR decreases from 24.807 to 22.987 dB. A larger horizon permits longer prediction intervals without a full evaluation, trading fidelity for speed. Because $n_{\max}$ also enters the temporal predictor, the sweep measures the combined effect of these changes. The default $n_{\max} = 6$ lies between the higher-fidelity, shorter-horizon settings and the faster, longer-horizon settings.

Correspondence filtering. The threshold $c_{\min}$ controls the minimum coverage required for tokens to participate in correspondence grouping. Higher thresholds yield higher PSNR across the tested values, suggesting that stricter filtering can improve the usefulness of the geometric shared component. Active tokens outside retained groups are not discarded: their proposed displacements remain in the residual component through $(I - P_\sigma)\Delta_t$. The threshold therefore controls where shared geometric correction applies, rather than whether these tokens receive an update. The higher PSNR at $c_{\min} = 0.5$ is accompanied by higher measured latency, so it does not dominate the default setting on both criteria.

Gain regularization. The coefficient λ penalizes the magnitude of the accepted displacement during anchor-based gain estimation. For fixed calibration statistics, increasing λ shrinks the fitted gains

toward zero, limiting the magnitude of the submitted update. The default $\lambda = 0.1$ achieves the highest PSNR among the tested values. Its advantage over $\lambda = 0$ is consistent with regularization limiting excessive updates, while the lower PSNR at $\lambda = 1$ is consistent with overly conservative transport. The tested values therefore support the default as a useful operating point, without implying that it maximizes PSNR for every scene or latency budget.

G Interpreting the Motivation Diagnostics

Appearance-based matching control. Image matching is used only as a diagnostic grouping baseline and is distinct from the geometric correspondence map P_σ used by GeoTrans. It operates on the input appearance over the same latent patch grid. Each patch is represented by its pooled RGB mean, within-patch color variance, and luminance-gradient magnitude. For Voyager, the descriptor additionally includes the patch visibility fraction; for Aether, the single condition image is broadcast over latent frames. Descriptors are standardized channel-wise, and each token is assigned to its nearest reference-frame token under Euclidean distance. Tokens assigned to the same reference token form a group, with groups containing fewer than two tokens excluded.

Coherence and shared displacement structure. For each grouping scheme, we measure the mean within-group pairwise cosine similarity of token displacements and the fraction of target squared displacement captured by the corresponding group means. The shuffled-geometry control preserves the grouping statistics while removing the original geometric association.

At horizon $h = 4$, the group-mean component captures 23.2% of the target squared displacement on Voyager for geometric grouping, compared with 8.5% for image matching and 17.3% for shuffled geometry. Geometric grouping also exhibits higher within-group directional coherence. Together, these diagnostics show that geometric correspondences organize more coherent shared displacement than appearance-based matching on Voyager.

Incremental predictive value of geometry. Fig. 3(c) analyzes whether geometric cues provide information beyond temporal and image-based cues for identifying large temporal-proposal errors. Under the same held-out analysis protocol, we measure the AUROC change after adding geometric cues and average the resulting gain over $h \in \{4, 8\}$.

For model m and scene s, let $\delta_{m,s}$ denote this AUROC improvement. For visualization, we report the normalized gain

$$\widetilde{\delta}_{m,s} = \frac{\delta_{m,s}}{\max_{s'} |\delta_{m,s'}|}. \tag{27}$$

This normalization preserves the sign and within-model ordering of the geometry gain. A positive value indicates that geometry provides additional predictive information beyond the temporal and image cues, while zero indicates no improvement. Because Voyager and Aether are normalized independently, the displayed magnitudes should not be interpreted as absolute AUROC comparisons across models.

H Scope and Discussion

What geometry contributes. GeoTrans uses correspondences to choose which displacement components share a calibration coefficient. It does not require the corresponding features themselves to coincide, nor does it replace every token's motion by its group mean. Equations 22 and 25 characterize this distinction algebraically. The scalar and shuffled controls isolate the contribution of the geometry-defined assignment from uniform calibration and from merely introducing two gain coefficients with a comparable grouping structure. The appearance-based grouping in Sec. G provides a complementary diagnostic comparison of how alternative groupings organize the underlying state displacements.

Approximate correspondence and conservative updates. Depth-based patch correspondences depend on the supplied geometry and discretization. Coverage and visibility filtering restrict their support, while the residual component retains active tokens without usable matches. The map remains fixed during a sequence, and gains adapt to completed anchor transitions. The norm bound

in Eq. 26 controls the accepted displacement magnitude; full anchor refresh supplies new observations as prediction horizons and shadow drift grow. These mechanisms address different parts of the transport problem and do not require an error target at a skipped step.

Evaluation target. Full-step-reference fidelity measures how well acceleration preserves the frozen model's output. Benchmark scores and reconstruction metrics provide complementary assessments of generated content and geometric outputs. The experiments support the reported operating points and controlled comparisons; the analytic properties characterize the transition operator rather than guaranteeing future rollout accuracy.