SYMREGFLOW: SYMMETRY-REGULARIZED FLOW MATCHING FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-matching-based multi-view world models generate realistic videos, but are commonly restricted to fixed camera rigs. Extending them to continuously varying camera poses requires paired pose–video observations with dense pose coverage, which are costly to acquire. We introduce *SymRegFlow*, a symmetry-regularized flow-matching framework for multi-view-consistent video generation across continuous viewpoints without ground-truth novel-view RGB supervision. For each target pose, SymRegFlow geometrically warps source views into noisy anchors and combines masked dual-anchor supervision with cross-anchor denoising-output consistency to mitigate anchor-specific errors. Under an affine Gaussian surrogate, we prove that suitable consistency regularization recovers the clean-reference optimum at fixed noise levels, strictly outperforming single- and merged-anchor baselines. Experiments on Cosmos-Drive-Dreams and nuScenes demonstrate high-quality, multi-view-consistent autonomous-driving video generation: on nuScenes, SymRegFlow achieves the lowest FVD and FVMD among the evaluated baselines, reducing FVD by over 31% relative to the best baseline, and source-conditioned inference also attains the best FID and instance preservation.

1 INTRODUCTION

Video world models have become powerful generative simulators of dynamic environments. In autonomous driving, recent systems jointly synthesize the RGB streams of a surround-view camera rig, conditioned on maps, object layouts, and text (Gao et al., 2024a;b; Agarwal et al., 2025). Joint multi-view learning can implicitly capture spatial consistency across views. We refer to the desired property as video $\mathrm{SE}(3)$ symmetry: under a rigid camera-pose change $g \in \mathrm{SE}(3)$, the scene content and dynamics remain unchanged, while the generated RGB video transforms consistently with the induced reprojection on co-visible regions (Dupont et al., 2020; Xu et al., 2022). Because $\mathrm{SE}(3)$ is continuous, this consistency should extend across camera poses; current models, however, support only the finite viewpoints represented by their training rigs.

Continuous-view generation supports sensor simulation, closed-loop autonomy evaluation (Yang et al., 2023), training-data augmentation (Tonderski et al., 2024), and cinematographic camera control (He et al., 2025; Bai et al., 2025a). However, nuScenes and Waymo Open (Caesar et al., 2020; Sun et al., 2020) provide synchronized videos only from fixed calibrated cameras, leaving intermediate viewpoints and novel camera heights or baselines without paired targets. ReCamMaster and SynCamMaster obtain synchronized multi-camera training videos from Unreal Engine (Bai et al., 2025a;b), incurring high generation costs, a sim-to-real gap, and unmodeled simulator artifacts.

To avoid simulator-generated supervision, we use off-the-shelf 3D perception models to warp observed source views to sampled novel poses, producing noisy anchor videos affected by geometry errors, occlusions, and resampling. We introduce *SymRegFlow*, which combines masked dual-anchor supervision with cross-anchor denoising consistency under shared diffusion noise, while retaining observed-view supervision to preserve pretrained generative capabilities. This symmetry inductive bias reduces sensitivity to anchor-specific errors without requiring ground-truth novel-view videos.

SymRegFlow is a fine-tuning objective that keeps the flow-matching training and inference framework. Our implementation adds continuous-view conditioning, directed cross-view attention (DCVA), and source-latent replacement for source-conditioned inference. Our contributions are:

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.

- We propose SymRegFlow for cross-view-consistent video generation across continuous viewpoints without novel-view RGB supervision, using masked dual-anchor learning and cross-anchor denoising consistency.
- We prove that, under an affine Gaussian surrogate, suitable consistency regularization counteracts anchor-induced variance inflation and recovers the clean-reference optimum at each fixed noise level, strictly improving upon single- and merged-anchor training.
- We show that on the nuScenes midpoint-view benchmark, SymRegFlow achieves the lowest FVD and FVMD among the evaluated baselines (over 31% lower FVD than the best baseline), while source-conditioned inference also attains the best FID and instance preservation.

2 METHOD

2.1 PRELIMINARIES

Symmetry and equivariance. A symmetry is a group transformation that leaves an object or distribution unchanged. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, and let $\rho_{\mathcal{X}}(g)$ and $\rho_{\mathcal{Y}}(g)$ denote how a group element g acts on each space. A map $\mathcal{F}\colon \mathcal{X} \to \mathcal{Y}$ is equivariant if it commutes with these actions: $\mathcal{F}(\rho_{\mathcal{X}}(g)x) = \rho_{\mathcal{Y}}(g)\mathcal{F}(x)$. Invariance is the special case in which $\rho_{\mathcal{Y}}(g)$ is the identity (Bronstein et al., 2021). Thus, symmetry specifies the structure to preserve, while equivariance specifies how a model preserves it. For the video $\mathrm{SE}(3)$ symmetry defined in the introduction, $\rho_{\mathcal{X}}(g)$ changes the requested camera pose and $\rho_{\mathcal{Y}}(g)$ is the induced RGB reprojection. The scene content and dynamics remain invariant, but its observation changes; the generator should therefore be camera-pose equivariant on co-visible regions (Dupont et al., 2020; Xu et al., 2022).

Multi-view flow matching. Let $\mathbf{x} = (x^1, \ldots, x^V)$ be a clean, synchronized V-view video, where $x^v \in \mathbb{R}^{F\times H\times W\times 3}$. Here, F, H, and W denote the number of frames, frame height, and frame width, respectively. Let $\mathcal{C} = \{(\mathbf{T}^v, K^v)\}_{v=1}^V$ collect camera poses $\mathbf{T}^v \in \mathrm{SE}(3)$ and intrinsics K^v. The multimodal condition is $\mathbf{c} = (\mathcal{C}, c_{\text{text}}, c_{\text{map}}, c_{\text{box}}, c_{\text{act}}, \ldots)$, including any available text, maps, boxes, or actions. Given $\mathbf{x} \sim p_{\text{data}}(\cdot \mid \mathbf{c})$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and a noise level $t \sim p_{\text{train}}$ drawn from the training scheduler's distribution on $(0, 1)$, the forward path corrupts the video as $\mathbf{Z}_t = (1-t)\mathbf{x} + t\boldsymbol{\epsilon}$, with $\mathbf{Z}_t = (Z_t^1, \ldots, Z_t^V)$, $\mathbf{Z}_0 = \mathbf{x}$, and $\mathbf{Z}_1 = \boldsymbol{\epsilon}$. We denote the denoising output by $\widehat{\mathbf{Z}}_{0,\theta}$ and express velocity-matching supervision in clean-prediction space (Li & He, 2026):

$$\mathcal{L}_x = \mathbb{E}\left[\frac{1}{t^2}\left\|\widehat{\mathbf{Z}}_{0,\theta}(\mathbf{Z}_t, \mathbf{c}, t) - \mathbf{x}\right\|_2^2\right]. \tag{1}$$

Its component $\widehat{Z}^v_{0,\theta}$ is the denoising output for view v. The corresponding flow velocity is $\mathbf{u}_\theta = (\mathbf{Z}_t - \widehat{\mathbf{Z}}_{0,\theta})/t$ for $t > 0$ (Lipman et al., 2023). Starting from $\mathbf{Z}_1 \sim \mathcal{N}(0, I)$, we integrate $d\mathbf{Z}_t/dt = \mathbf{u}_\theta$ backward from $t = 1$ to 0.

Large video world models instead encode each view with a frozen video autoencoder $\mathcal{E}, \mathcal{D}$, yielding clean latents $X^v = \mathcal{E}(x^v)$ and the multi-view tuple $\mathbf{X} = (X^1, \ldots, X^V)$. The latent forward path is $\mathbf{Z}_t = (1-t)\mathbf{X} + t\boldsymbol{\epsilon}$, with $\mathbf{Z}_0 = \mathbf{X}$. Sampled clean latents $\widehat{X}^v = \widehat{Z}^v_0$ are decoded as $\widehat{x}^v = \mathcal{D}(\widehat{X}^v)$ (Rombach et al., 2022; Agarwal et al., 2025).

2.2 SYMREGFLOW FROM NOISY MULTI-VIEW ANCHORS

Problem setting. Let x^1 and x^2 denote synchronized ground-truth RGB videos collected at two available discrete camera views, with poses $\mathbf{T}^s \in \mathrm{SE}(3)$ and intrinsics K^s for $s \in \mathcal{S} = \{1, 2\}$. We assume a multi-view flow matching model pretrained on these views. Our goal is to learn the conditional generator $p(x^1, x^2, x^{\mathrm{n}} \mid \mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$ for a continuous novel view n, specified by camera pose $\mathbf{T}^{\mathrm{n}} \in \mathrm{SE}(3)$ and intrinsics K^{n}, by fine-tuning the pretrained model without ground-truth novel-view RGB x^{n}; the formulation generalizes naturally to more observed views. For this requested novel view, we encode $(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$ with a learned embedder e_ϕ, typically using the pretrained model's existing pose-conditioning mechanism. Here, ϕ denotes the embedder parameters, learned end-to-end under the SymRegFlow objective defined below.

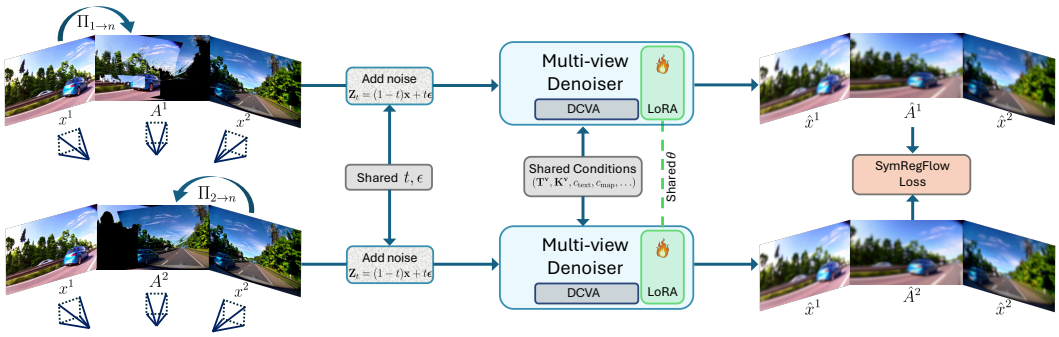


Figure 1: SymRegFlow pipeline. Two observed videos x^1, x^2 are independently warped to the same requested novel view, yielding noisy anchors A^1, A^2. The resulting multi-view pseudo-examples are corrupted with shared noise and fed into the same denoiser with directed cross-view attention (DCVA). VAE of latent flow-matching is omitted from the drawing for clarity.

Geometry-derived noisy anchors. Because ground-truth RGB x^{n} is unavailable, we form a pseudo-target from each source using an efficient off-the-shelf 3D perception system such as ViPE (Huang et al., 2025). We use its estimated dense depth and camera geometry to lift the source RGB into a colored 3D point cloud, then project and render it at $(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$. Denoting this pipeline by $\Pi_{s\to\mathrm{n}}$, we obtain

$$(A^s, m^s) = \Pi_{s\to\mathrm{n}}(x^s), \qquad s \in \mathcal{S}. \tag{2}$$

Here, A^s is a partial novel-view anchor and m^s its visibility/confidence mask. Writing $M^s = \mathrm{diag}(m^s)$, its visible pixels satisfy $M^s A^s = M^s(x^{\mathrm{n}} + E^s)$, where E^s is the aggregate warping error. Geometry errors, scene dynamics, and occlusions make the anchors noisy and incomplete. In general, the magnitude of E^s grows with source–target viewpoint distance, as illustrated by the rotation-dependent $\mathrm{SE}(2)$ example in Fig. 2(a).

Symmetry-regularized fine-tuning. Dual-anchor fine-tuning constructs one pseudo-example $\widetilde{\mathbf{x}}^{(s)} = (x^1, A^s, x^2)$ for each source $s \in \mathcal{S}$. The anchors describe the same target view with different valid supports. Camera-pose equivariance motivates consistency on shared visible content; we extend this motivation to cross-anchor denoising agreement over regions supported by either anchor.

To implement this regularizer, we apply the forward corruption from Sec. 2.1 to each pseudo-example, yielding $\mathbf{Z}_t^{(s)}$, and predict its clean multi-view output under $\mathbf{c}^{\mathrm{n}} = (\mathbf{c}, e_\phi(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}}))$:

$$\widehat{\mathbf{Z}}_{0,\theta}^{(s)} = f_\theta\left(\mathbf{Z}_t^{(s)}, \mathbf{c}^{\mathrm{n}}, t\right), \qquad s \in \mathcal{S}. \tag{3}$$

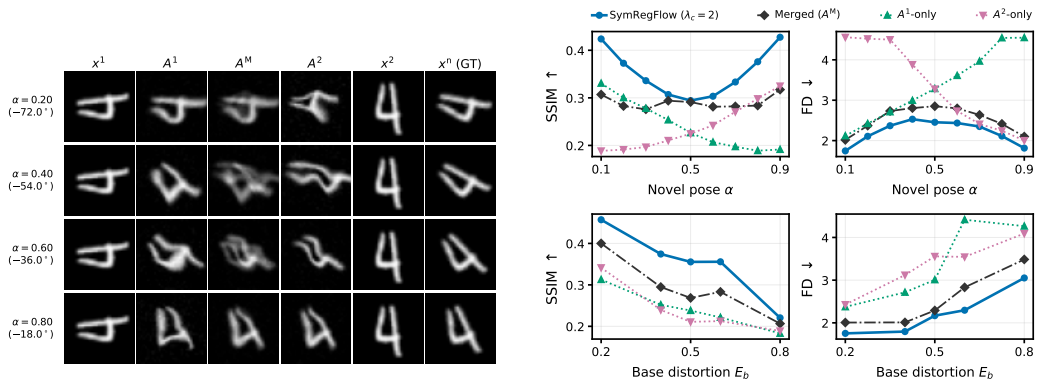
Writing $\widehat{Z}^{v,(s)} := \widehat{Z}_{0,\theta}^{v,(s)}$ and $M^{\cup} = \mathrm{diag}(\max(m^1, m^2))$, we define the cross-anchor consistency regularization loss over regions supported by either anchor:

$$\mathcal{L}_{\mathrm{cons}} = \left\| \widehat{Z}^{\mathrm{n},(1)} - \widehat{Z}^{\mathrm{n},(2)} \right\|_{M^{\cup}}^2. \tag{4}$$

The union mask includes locations supported by only one anchor, coupling the two predictions through the available anchor supervision. Locations unsupported by both anchors receive no consistency penalty. At each optimization step, the two forward passes share the sampled slot-wise Gaussian noise, noise level t, condition $\mathbf{c}^{\mathrm{n}}$, and observed-view inputs (Fig. 1).

To preserve the pretrained model's capabilities, we retain supervision on all observed views; to learn the novel-pose embedder e_ϕ, we supervise each novel-view output with its geometry-derived anchor:

$$\begin{aligned} \mathcal{L}_{\mathrm{src}} &= \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{S}} \left\| \widehat{Z}^{r,(s)} - x^r \right\|_2^2, \\ \mathcal{L}_{\mathrm{anc}} &= \sum_{s\in\mathcal{S}} \left\| \widehat{Z}^{\mathrm{n},(s)} - A^s \right\|_{M^s}^2. \end{aligned} \tag{5}$$



(a) (b)

Figure 2: SE(2) MNIST toy study. **(a)** Viewpoint-dependent noisy anchors A^1, A^2, their mask-weighted fusion A^{M}, and exact target x^{n}. **(b)** SSIM and FD versus novel pose α (top) and base distortion E_b (bottom), averaged over five distortion levels and nine poses, respectively. FD denotes pixel-space Fréchet distance on flattened 16×16 pixel features.

The complete objective is

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{t^2}\left(\mathcal{L}_{\mathrm{src}} + \mathcal{L}_{\mathrm{anc}}\right) + \lambda_c \mathcal{L}_{\mathrm{cons}}\right]. \tag{6}$$

We backpropagate the combined loss once through both branches, with λ_c weighting consistency. The supervision and consistency objectives extend to additional observed views. At inference, the novel camera is specified through $e_\phi(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$, without anchors or geometric warping.

Directed Cross-View Attention. To protect observed-view fidelity and multi-view consistency from noisy anchors, Directed Cross-View Attention (DCVA) restricts observed-view queries to observed-view tokens while allowing novel-view queries to attend to all tokens. Parameter-free masks enforce this observed-to-novel information flow in every full-attention layer. Sec. 3.3 evaluates its trade-off between view fidelity, cross-view agreement, and computational cost.

Appendix A gives the two-pass, three-view training procedure in Algorithm 1, including the video backbone's prediction parameterization. Algorithms 2 and 3 describe *source-cond.* inference (novel-view synthesis conditioned on observed source RGB) and *joint-generated* inference (joint three-view generation from noise), respectively.

2.3 SE(2) Toy Example

We illustrate SymRegFlow in an SE(2) MNIST setting with observed views x^1 and x^2 at $-90°$ and $0°$, respectively. A pose $\alpha \in (0, 1)$ specifies the novel-view angle $\vartheta_\alpha = -90°(1 - \alpha)$; the exact target x^{n} is withheld during fine-tuning and used only for validation. The distortion of anchors A^1 and A^2 increases with their respective source–target angular distances, simulating viewpoint-dependent warping errors (Fig. 2(a)). We pretrain a pixel-space multi-view flow-matching model on the observed views and fine-tune it with these anchors under Eq. (6). We compare against single-anchor baselines (A^1-only and A^2-only) and Merged anchor, which uses their mask-weighted fusion A^{M}. Appendix D details anchor construction and training.

Toy-example analysis. We evaluate nine withheld poses at five base distortion levels $E_b \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. All methods start from the same checkpoint, are fine-tuned without x^{n}, and are evaluated against exact rotations of held-out digits; Appendix D provides training and evaluation details and the consistency-weight sweep. SymRegFlow uses DCVA with $\lambda_c = 2$, selected by its mean exact-target pixel-space Fréchet distance (FD) over the pose–distortion grid. As shown in Fig. 2(b), SymRegFlow achieves both the lowest FD and the highest SSIM at every evaluated target pose and base distortion level.

Single-anchor performance improves with smaller rotations and cleaner anchors: A^1-only is strongest near $\alpha = 0$, and A^2-only near $\alpha = 1$. Overall, single-anchor fine-tuning performs worst due to source-specific warping bias, whereas using both anchors improves performance. SymRegFlow outperforms Merged anchor at every distortion level, with mean FD advantages of 0.254 at $E_b = 0.2$ and 0.434 at $E_b = 0.8$. Merged anchor remains competitive for 2D rotations, which omit the geometric, motion, and temporal misalignment of real-world video warps, but substantially underperforms SymRegFlow on videos (Sec. 3.3).

2.4 Why SymRegFlow Can Improve Noisy-Anchor Fine-Tuning

We analyze SymRegFlow's combination of dual-anchor supervision and consistency regularization in a scalar Gaussian surrogate at fixed condition $\mathbf{c}^{\mathrm{n}}$ and $t \in (0, 1)$, with unit masks and fixed source-view predictions. Let $x^{\mathrm{n}} \sim \mathcal{N}(\mu, q)$, $q > 0$, and $A^s = x^{\mathrm{n}} + E^s$. Assume $x^{\mathrm{n}}, E^1, E^2, \epsilon$ are mutually independent, with centered Gaussian errors E^1, E^2 of positive variance and $\epsilon \sim \mathcal{N}(0, 1)$ shared across branches. Restrict the novel-view clean predictor to $f_\theta(z, t) = \mu + \theta[z - (1 - t)\mu]$, with scalar slope θ and conditioning suppressed. This family contains the optimal clean Gaussian denoiser. On the clean reference path $Z_t = (1 - t)x^{\mathrm{n}} + t\epsilon$, define the excess flow-matching risk by subtracting the minimum clean-reference risk:

$$\mathcal{R}_t^{\mathrm{exc}}(\theta) := \frac{1}{t^2}\mathbb{E}_{x^{\mathrm{n}},\epsilon}\left[\left(f_\theta(Z_t, t) - x^{\mathrm{n}}\right)^2\right] - \frac{q}{(1-t)^2 q + t^2}. \tag{7}$$

Proposition 1 (Clean-reference improvement with SymRegFlow)**.** *Under this surrogate, for non-negative λ_c, let θ_{λ_c} minimize the expected SymRegFlow objective in Eq. (6) at the fixed time t, and let $\theta_1, \theta_2, \theta_{\mathrm{M}}$ be the corresponding unregularized optima for A^1-only, A^2-only, and any fixed convex merge A^{M}. Then*

$$\lambda_c^\star(t) = \frac{1}{q(1-t)^2}, \qquad \mathcal{R}_t^{\mathrm{exc}}\big(\theta_{\lambda_c^\star(t)}\big) = 0 < \min_{j \in \{1,2,\mathrm{M}\}} \mathcal{R}_t^{\mathrm{exc}}(\theta_j). \tag{8}$$

The strict improvement holds throughout a neighborhood of $\lambda_c^\star(t)$. The excess risk decreases up to $\lambda_c^\star(t)$ and increases thereafter.

Thus, dual-anchor supervision together with suitable consistency regularization recovers the clean-reference optimum in this surrogate. Appendix B gives the proof and numerical illustration; the guarantee is pointwise in t.

2.5 Extension to Latent Flow Matching

The preceding formulation uses RGB anchors, whereas latent flow-matching models operate in autoencoder latent space. Let $\mathcal{E}$ and $\mathcal{D}$ be the frozen encoder and decoder, with $X^s = \mathcal{E}(x^s)$. We first introduce the induced latent-warp operator $\mathcal{W}_{s\to\mathrm{n}} := \mathcal{E} \circ \Pi_{s\to\mathrm{n}} \circ \mathcal{D}$. In practice, assuming near-perfect autoencoder reconstruction, we directly warp the source RGB before encoding:

$$A^s := \mathcal{E}(\Pi_{s\to\mathrm{n}}(x^s)) \approx \mathcal{W}_{s\to\mathrm{n}}(X^s). \tag{9}$$

Here, $\Pi_{s\to\mathrm{n}}$ is the pixel-space warping operation defined above, and A^s is the practical latent anchor used for fine-tuning. Downsampling the corresponding visibility mask to latent resolution yields M^s, and the latent anchor satisfies

$$M^s A^s = M^s(X^{\mathrm{n}} + E^s), \tag{10}$$

where E^s denotes the latent-space warping error. During fine-tuning, these latent anchors replace the RGB anchors in the paired pseudo-examples, using the three loss components in Eq. (6). The video implementation retains the backbone's supervised velocity losses, equivalent to t^{-2}-weighted clean-output losses, while evaluating consistency on clean predictions (Algorithm 1).

3 Experiments

Our video models are adapted from the Cosmos-Transfer2.5 multi-view flow-matching backbone (Ali et al., 2025), which we fine-tune on Cosmos-Drive-Dreams (Ren et al., 2025a) and nuScenes (Caesar et al., 2020) using 29-frame clips at 1280×720 resolution per view, following prior work (Guo et al., 2025). We report quantitative evaluation on nuScenes. Following the depth-based warping of GEN3C (Ren et al., 2025b), we unproject source frames into 3D using estimated depth and render them at target camera poses to obtain partial anchors and visibility masks.

3.1 Qualitative Generation on Cosmos-Drive-Dreams

Figure 3 shows SymRegFlow generating videos at continuously specified novel camera poses while maintaining visual consistency with the existing camera views. Each top-row novel view is generated in a separate inference run with the same random seed and scene conditioning. In the bottom two rows, under the same scene conditioning, seeds 0 and 42 produce different background trees and cloud patterns, with coherent appearance across views and frames. Unlike camera control by rendering a fixed 3D reconstruction, SymRegFlow supports both viewpoint variation and stochastic appearance diversity, which benefits multi-camera data generation for autonomous driving.



Figure 3: Qualitative results on Cosmos-Drive-Dreams. Top: existing and generated novel views at one timestamp. Bottom: novel-view frames and existing-view sequences for seeds 0 and 42.

3.2 Quantitative Comparison on nuScenes

Evaluation protocol. We train and evaluate on the same 30 validation scenes as DiST-4D (Guo et al., 2025). We report novel-view FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) for visual quality, Fréchet Video Motion Distance (FVMD) (Liu et al., 2024), VBench subject consistency, background consistency, and temporal flickering (Huang et al., 2024) for temporal quality, VBench-2.0 instance preservation (IP) (Zheng et al., 2025) for entity stability, and MEt3R (Asim et al., 2025) for cross-view consistency. The detailed protocol is provided in Appendix C.2.

We evaluate at the midpoint ($\alpha = 0.5$), a challenging interpolation pose that maximizes the distance to the nearest observed view (Sec. 2.3). Table 1 shows that both inference modes outperform all baselines in FVD and FVMD, indicating improved video quality and motion consistency. *Joint-generated* inference achieves the lowest FVD (158.62) and FVMD (2519.17) and the highest VBench temporal-flickering score (96.25), with subject and background consistency close to OmniRe. *Source-cond.* inference improves FID from 22.00 to 19.22, surpassing FreeVS (20.60), while retaining similar FVD (159.02) but yielding higher FVMD (2932.87) and a lower temporal-flickering score (95.82). *Source-cond.* inference also achieves the highest instance-preservation score (99.44), which is critical for autonomous driving. Because *joint-generated* content can differ from dataset observations under the same scene conditioning (Fig. 3), its MEt3R (0.2818) is computed against the generated existing views and measures internal cross-view agreement, whereas *source-cond.* MEt3R (0.2920) measures agreement with captured views. The latter trails FreeVS (0.2829), plausibly because our source conditioning relies on training-free latent replacement, whereas FreeVS injects reprojected source appearance through pseudo-images during training and inference (Wang et al., 2025).

Figure 4 highlights photometric inconsistencies in baseline outputs: color and brightness differences between nuScenes source cameras appear as left–right discontinuities in synthesized frames, resembling the fusion seam of merged-anchor fine-tuning (Sec. 3.3). Several baselines also distort

Table 1: Novel-view generation on all 30 nuScenes validation scenes at the interpolated midpoint ($\alpha = 0.5$) between two observed camera views. VBench scores are multiplied by 100. †MEt3R is computed using generated existing-view videos; other MEt3R scores use captured views. **Bold**: best; underline: second-best within each column, ranked at full precision.

Method	FID↓	FVD↓	FVMD↓	VBench Subj.↑	VBench Bg.↑	VBench Flicker.↑	VBench IP↑	MEt3R↓
DiST-4D (Guo et al., 2025)	34.59	231.95	5381.67	88.72	91.81	95.67	92.78	0.3362
OmniRe (Chen et al., 2025)	44.58	291.99	3104.05	**90.20**	**93.50**	95.65	96.11	0.3177
GEN3C (Ren et al., 2025b)	40.28	289.24	6144.97	88.39	92.66	95.34	37.78	0.3054
FreeVS (Wang et al., 2025)	20.60	258.53	4216.99	88.79	91.64	96.04	88.33	0.2829
Ours (*joint-generated*)	22.00	**158.62**	**2519.17**	90.13	93.32	**96.25**	96.11	**0.2818**†
Ours (*source-cond.*)	**19.22**	159.02	2932.87	90.17	93.32	95.82	**99.44**	0.2920

thin structures, such as street lamps. SymRegFlow produces more uniform appearance across each frame and better preserves these structures, consistent with its cross-anchor consistency regularization. Appendix E provides more comparisons.



Figure 4: Qualitative comparison of novel-view synthesis on nuScenes. The last row shows *source-cond.* SymRegFlow results for a fair comparison with the baselines.

3.3 Ablation Studies

Table 2 ablates anchor strategy, consistency weight, and DCVA for the midpoint novel pose. Each variant uses one LoRA adapter trained jointly on the same four nuScenes validation scenes and evaluated at step 2,048; Appendix C gives details and further analyses.

Model configurations. We compare three anchor baselines: A_L and A_R use only the left or right warped anchor, respectively, while A_M uses a mask-weighted pixel-space merge of both RGB anchors. All three use DCVA for fair comparison. The SymRegFlow dual-anchor sweep C_γ uses DCVA and $\lambda_c = \gamma$ for $\gamma \in \{0, 0.01, 0.05, 0.1, 0.5, 1, 2\}$. Across these weights, $C_{0.05}$ performs best, with the lowest existing- and novel-view FVD and MEt3R and the highest VBench subject and background consistency; we therefore select the optimal $\lambda_c^\star = 0.05$. To investigate the influence of DCVA, $C_{0.05}^{\mathrm{FA}}$ replaces directed attention with full attention at the same weight.

Effectiveness of dual-anchor fine-tuning. Dual-anchor fine-tuning improves novel-view generation even without consistency regularization: C_0 outperforms all three anchor baselines in novel-view FVD and both VBench consistency scores. For example, C_0 lowers novel-view FVD from 562.23 for the strongest single-anchor baseline A_L to 545.14, while $C_{0.05}$ lowers it from 704.32 for merged-anchor A_M to 498.94. Although A_M uses the same two anchors as SymRegFlow for training, it inherits artifacts from mismatched RGB merging, including misaligned overlaps and residual

Table 2: Ablation study on four nuScenes scenes. λ_c is the consistency regularization weight (–: not applicable). The superscript 'FA' in $C^{\mathrm{FA}}_{0.05}$ denotes full attention without DCVA. VBench scores are multiplied by 100. **Bold**: best; underline: second-best.

Model	Anchors			DCVA	λ_c	Existing view	Novel view			
	Left	Right	Merge			FVD↓	FVD↓	VBench Subj.↑	VBench Bg.↑	MEt3R↓
Anchor baselines										
A_L	✓			✓	–	238.14	562.23	87.86	90.36	0.2955
A_R		✓		✓	–	**198.86**	605.63	88.79	92.20	0.3512
A_M			✓	✓	–	213.97	704.32	87.71	91.78	0.2787
SymRegFlow variants (ours)										
C_0	✓	✓		✓	0	210.19	545.14	89.64	92.56	0.3380
$C_{0.01}$	✓	✓		✓	0.01	233.29	558.56	89.51	91.53	0.2832
$C_{0.05}$	✓	✓		✓	0.05	209.37	**498.94**	**89.73**	**92.75**	0.2660
$C_{0.1}$	✓	✓		✓	0.1	223.80	574.24	88.77	92.57	0.2935
$C_{0.5}$	✓	✓		✓	0.5	212.84	575.07	88.50	92.31	0.3032
C_1	✓	✓		✓	1	237.80	624.15	89.09	92.32	0.3293
C_2	✓	✓		✓	2	271.87	701.95	88.82	91.88	0.2969
$C^{\mathrm{FA}}_{0.05}$	✓	✓			0.05	261.40	586.53	88.93	92.62	**0.1996**

gray pixels where inpainting fails (Fig. 5). A_R nevertheless achieves the best existing-view FVD, plausibly because its single-anchor objective favors reconstructing existing views without competing consistency regularization; this advantage does not translate into stronger novel-view generation.

To better expose the biases in A_L and A_R, we conduct a controlled one-scene experiment. Both models tend to copy anchor-visible content when inpainting missing regions, duplicating construction workers and background buildings and violating camera-pose symmetry. In Fig. 5, queries on duplicated construction workers (red) attend most strongly to their source-supported counterparts (cyan), reaching $13.82\times$ and $12.22\times$ uniform attention, respectively. Cross-view and within-frame analyses further show that SymRegFlow reduces source-dependent attention biases in A_L and A_R (Appendix C.4). A_M avoids duplication but retains a conspicuous fusion seam and geometric discontinuity. SymRegFlow avoids both artifacts.

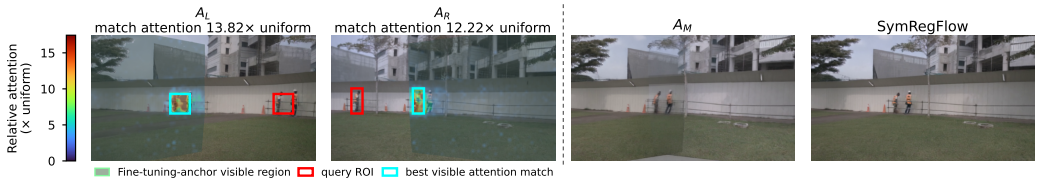


Figure 5: Examples of attention bias in anchor baselines. A_L and A_R copy workers from source-supported regions (cyan boxes) into missing regions (red boxes); green shading indicates anchor-visible regions. The colorbar shows attention strength as a multiple of uniform attention.

Effectiveness of λ_c. Increasing λ_c from 0 to 0.01 to 0.05 reduces MEt3R from 0.3380 to 0.2832 to 0.2660. Novel-view FVD improves from 545.14 to 498.94, despite a slight deterioration at 0.01; $\lambda_c = 0.05$ also maximizes both VBench consistency scores. Larger weights worsen FVD, from 574.24 at 0.1 to 701.95 at 2 (Table 2). This optimum qualitatively matches Proposition 1, whose excess risk decreases to $\lambda_c^\star(t)$ and increases thereafter. Appendix C.3 provides radar profiles.

Effectiveness of DCVA. Replacing DCVA with full attention increases existing-view FVD by 24.85% and worsens novel-view FVD, but lowers MEt3R by 24.96% with similar VBench scores. It also requires $1.47\times$ the backbone forward/backward time (Appendix C.3). DCVA thus trades some cross-view agreement for efficiency and view fidelity.

4 Related Work

4.1 Symmetry and Equivariance in World Models

Generative models commonly preserve symmetric target distributions through equivariant score or velocity fields (Köhler et al., 2020). Equivariant diffusion uses an $\mathrm{E}(3)$-equivariant denoiser, while equivariant flow-matching methods constrain continuous normalizing-flow vector fields or probability paths (Hoogeboom et al., 2022; Klein et al., 2023; Song et al., 2023). Flow Equivariant World Models instead build motion-aligned equivariance into a structured latent memory (Lillemark et al., 2026). More generally, these methods require symmetry-specific denoisers, vector fields, or latent-memory architectures. Architecture-agnostic alternatives avoid such redesign but rely, respectively, on group-averaged training targets or symmetry-structured noise during training and sampling (Tong et al., 2025; Liu & Vahdat, 2025). SymRegFlow introduces a symmetry inductive bias through its fine-tuning objective, aligning denoising outputs induced by independently warped, noisy anchors over the union of valid anchor supports while retaining the underlying flow-matching formulation.

4.2 Camera-Controlled Video World Models

Training-free camera control. Training-free methods keep a pretrained video generator frozen and enforce a target trajectory at inference through latent replacement, sampling-time guidance, or test-time optimization (Hou & Chen, 2025; Park et al., 2026; Zhou et al., 2025; Wang et al., 2026). They avoid paired training data but typically require modified sampling or trajectory-specific refinement. SymRegFlow instead learns camera control by fine-tuning and retains the standard sampler for arbitrary target poses; *source-cond.* inference additionally uses source-latent replacement.

Training-based camera control. MotionCtrl, CameraCtrl, 4DiM, and ReCamMaster amortize camera-controlled generation using costly pose-annotated or synchronized pose–video pairs, generally producing one target stream (Wang et al., 2024b; He et al., 2025; Watson et al., 2025; Bai et al., 2025a). ReCapture (Zhang et al., 2025) fits video-specific LoRAs to valid pixels of a noisy warped anchor for each requested trajectory, masking missing regions and relying on the pretrained prior to repair artifacts. It neither couples source-camera anchors nor enforces cross-view symmetry. SymRegFlow instead uses dataset-level fine-tuning and agreement among independently warped anchors to reduce view-specific errors.

Joint-view world models. Joint-view generators couple synchronized spatio-temporal camera tokens through self- or cross-view attention, conditioned on rig identities, maps, boxes, text, or actions. Examples include MagicDrive, Drive-WM, and DriveDreamer-2 for driving (Gao et al., 2024a; Wang et al., 2024a; Zhao et al., 2025), and Cosmos-Predict/Transfer2.5 for driving and multi-camera robotics (Ali et al., 2025). Their cross-camera consistency remains tied to training-rig slots and poses. SymRegFlow introduces a continuously pose-conditioned target slot while preserving the pretrained multi-view prior, without novel-view RGB supervision.

5 Conclusion and Discussion

SymRegFlow extends pretrained multi-view world models to cross-view-consistent video generation across continuous viewpoints without novel-view RGB supervision, combining masked dual-anchor supervision with cross-anchor denoising consistency to mitigate source-dependent biases. Under an affine Gaussian surrogate, suitable regularization recovers the clean-reference optimum at each noise level. On nuScenes, SymRegFlow achieves the lowest FVD and FVMD among evaluated baselines, with source-conditioned inference also attaining the best FID and instance preservation; ablations confirm improved novel-view quality and reduced source-dependent attention bias.

Limitations. SymRegFlow depends on anchor quality: depth, projection, and visibility errors can corrupt supervision, especially near occlusions or under large viewpoint changes. Dual-anchor consistency cannot correct shared systematic errors, and jointly unseen regions lack direct supervision. Better anchors and geometric-uncertainty modeling remain open directions.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist in editing and polishing the manuscript for clarity and readability, and to support the retrieval and discovery of relevant literature. We did not use generative AI tools for any tasks requiring mandatory disclosure, and the remaining disclosure categories are not applicable to this work. All AI-assisted text was reviewed and revised by the authors. References suggested through AI-assisted searches were independently checked against the original publications for relevance, accuracy, and bibliographic correctness. We take full responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

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.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6034–6044, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing multi-camera video generation from diverse viewpoints. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo De Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. OmniRe: Omni urban scene reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025.

Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, pp. 2761–2770. PMLR, 2020.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3D geometry control. In *The Twelfth International Conference on Learning Representations*, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, et al. DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27231–27241. IEEE, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Hansen Lillemark, Benhao Huang, Fangneng Zhan, Yilun Du, and T. Anderson Keller. Flow equivariant world models: Structured memory for dynamic environments. In *Forty-third International Conference on Machine Learning*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Chao Liu and Arash Vahdat. On equivariance and fast sampling in video diffusion models trained with warped noise. *arXiv preprint arXiv:2504.09789*, 2025.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024.

Jangho Park, Taesung Kwon, and Jong Chul Ye. Zero4D: Training-free 4D video generation from single video using off-the-shelf video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4045–4054, June 2026.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025a.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3D molecule generation. *Advances in Neural Information Processing Systems*, 36:549–568, 2023.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. NeuRAD: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, 2024.

Vinh Tong, Yun Ye, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Symmetry-preserving diffusion models via target symmetrization. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. FreeVS: Generative view synthesis on free driving trajectory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024a.

Yuzhu Wang, Xi Ye, Duo Su, Yangyang Xu, and Jun Zhu. h-control: Training-free camera control via block-conditional Gibbs refinement. arXiv preprint arXiv:2605.11871, 2026.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David Fleet. Controlling space and time with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yinshuang Xu, Jiahui Lei, and Kostas Daniilidis. Equivariant light field convolution and transformer. *arXiv preprint arXiv:2212.14871*, 2022.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. UniSim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2050–2062, 2025.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *CoRR*, abs/2503.21755, 2025.

Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-Reframe: Enabling camera control for video diffusion models without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12779–12789, 2025.

A Training and Inference Algorithms

Algorithms 1–3 use the notation of Sec. 2.2 in latent space, with source indices $\mathcal{S} = \{1, 2\}$ and slot order $(1, \mathrm{n}, 2)$. The shared condition $\mathbf{c}$ includes text, source-camera identities, and world-scenario controls encoded in the same slot order. We abbreviate $\widehat{\mathbf{Z}}_{0,\theta}^{(s)}$ as $\widehat{\mathbf{Z}}^{(s)}$. Losses use elementwise means; for a mask broadcast over latent channels,

$$\|R\|_M^2 := \frac{\sum_i m_i R_i^2}{\max(\sum_i m_i, 1)}, \qquad M = \mathrm{diag}(m). \tag{11}$$

Thus, an empty mask contributes zero loss.

Algorithm 1 SymRegFlow training

Require: Prepared clips with sources x^1, x^2, anchors and masks $(A^s, m^s) = \Pi_{s\to\mathrm{n}}(x^s)$, condition $\mathbf{c}$, and novel camera $(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$; pretrained θ; initialized e_ϕ; frozen autoencoder $(\mathcal{E}, \mathcal{D})$; weight λ_c; training noise law p_{train} and weighting $\omega(t)$ (see below).

Freeze base weights; train LoRA and e_ϕ; enable DCVA in both passes.
for each optimization step **do**
  Sample a prepared clip and its novel camera, anchors, masks, and condition.
  $X^s \leftarrow \mathcal{E}(x^s), \quad A^s \leftarrow \mathcal{E}(A^s), \qquad s \in \mathcal{S}$
  $m^s \leftarrow \mathbf{1}[\mathrm{Resize}_{\mathrm{latent}}(m^s) > 0.5], \quad M^s \leftarrow \mathrm{diag}(m^s), \qquad s \in \mathcal{S}$
  $M^{\cup} \leftarrow \mathrm{diag}(\max(m^1, m^2))$
  $\mathbf{c}^{\mathrm{n}} \leftarrow (\mathbf{c}, e_\phi(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}}))$
  Sample $t \sim p_{\mathrm{train}}$ and $\boldsymbol{\epsilon} = (\epsilon^1, \epsilon^{\mathrm{n}}, \epsilon^2) \sim \mathcal{N}(0, I)$ once.
  for $s \in \mathcal{S}$ **do**
    $\mathbf{Z}_t^{(s)} \leftarrow (1-t)(X^1, A^s, X^2) + t\boldsymbol{\epsilon}$
    $\widehat{\mathbf{Z}}^{(s)} \leftarrow f_\theta(\mathbf{Z}_t^{(s)}, \mathbf{c}^{\mathrm{n}}, t)$ {shared noise, time, and condition}
  end for
  $\mathcal{L}_{\mathrm{src}} \leftarrow \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{S}} \left\| \widehat{Z}^{r,(s)} - X^r \right\|_2^2$
  $\mathcal{L}_{\mathrm{anc}} \leftarrow \sum_{s\in\mathcal{S}} \left\| \widehat{Z}^{\mathrm{n},(s)} - A^s \right\|_{M^s}^2$
  $\mathcal{L}_{\mathrm{cons}} \leftarrow \left\| \widehat{Z}^{\mathrm{n},(1)} - \widehat{Z}^{\mathrm{n},(2)} \right\|_{M^{\cup}}^2$
  $\mathcal{L} \leftarrow \omega(t)(\mathcal{L}_{\mathrm{src}} + \mathcal{L}_{\mathrm{anc}}) + \lambda_c \mathcal{L}_{\mathrm{cons}}$
  Backpropagate $\mathcal{L}$ once through both passes; update only $(\theta_{\mathrm{LoRA}}, \phi)$.
end for
return $(\theta_{\mathrm{LoRA}}, \phi)$

Prediction parameterization. Setting $\omega(t) = t^{-2}$ gives Eq. (6). The video implementation retains the pretrained velocity parameterization $\mathbf{u}_\theta$, using $f_\theta(\mathbf{Z}_t, \mathbf{c}^{\mathrm{n}}, t) := \mathbf{Z}_t - t\mathbf{u}_\theta(\mathbf{Z}_t, \mathbf{c}^{\mathrm{n}}, t)$ for its clean prediction. Its supervised velocity losses are equivalent to $\omega(t) = t^{-2}$, since $\|\mathbf{u}_\theta - (\boldsymbol{\epsilon} - \mathbf{X})\|^2 = t^{-2}\|\widehat{\mathbf{Z}}_{0,\theta} - \mathbf{X}\|^2$. Consistency is computed on clean predictions without this factor. Here p_{train} is the training scheduler's distribution of normalized noise levels $t \in (0, 1)$; mask resizing is trilinear. Both passes reuse the same conditioning dropout.

DCVA execution. Source queries attend only to source keys; novel queries attend to all three views. The four-stream implementation used in the DCVA experiments shares source computation and isolates the two novel streams, giving the same two-pass objective and gradients.

Sampling conventions. Both modes use fine-tuned SymRegFlow models with DCVA, without anchors or geometric warping. Let $t_0 > \cdots > t_N = 0$ be the scheduler's decreasing noise levels. $\mathbf{u}_\theta^{\mathrm{CFG}}$ denotes classifier-free guided velocity; the backbone's guidance convention is $\mathbf{u}_{\mathrm{cond}} + w(\mathbf{u}_{\mathrm{cond}} - \mathbf{u}_{\mathrm{uncond}})$. $\mathrm{SchedulerStep}$ advances the joint latent state and its solver history using the pretrained numerical solver (UniPC in the video implementation).

Algorithm 2 *Source-cond.* inference: novel-view synthesis conditioned on observed sources

Require: Fine-tuned $(\theta_{\mathrm{LoRA}}, \phi)$; source videos x^1, x^2; condition $\mathbf{c}$; novel camera $(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$; noise schedule $\{t_k\}_{k=0}^N$; guidance w; replacement horizon $N_{\mathrm{rep}} \in \{0, \ldots, N\}$.

$X^s \leftarrow \mathcal{E}(x^s), \quad s \in \mathcal{S}; \qquad \mathbf{c}^{\mathrm{n}} \leftarrow (\mathbf{c}, e_\phi(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}}))$
Sample $\boldsymbol{\epsilon} = (\epsilon^1, \epsilon^{\mathrm{n}}, \epsilon^2) \sim \mathcal{N}(0, I)$; initialize $\mathbf{Z}_{t_0} \leftarrow \boldsymbol{\epsilon}$ and the scheduler.
for $k = 0, \ldots, N-1$ **do**
  $\mathbf{Z}_{\mathrm{in}} \leftarrow \mathbf{Z}_{t_k}$
  if $k < N_{\mathrm{rep}}$ **then**
    $\widetilde{Z}^s_{t_k} \leftarrow (1 - t_k) X^s + t_k \epsilon^s, \qquad s \in \mathcal{S}$
    $\mathbf{Z}_{\mathrm{in}} \leftarrow (\widetilde{Z}^1_{t_k}, Z^{\mathrm{n}}_{t_k}, \widetilde{Z}^2_{t_k})$ {replace source inputs only}
  end if
  $\widehat{\mathbf{u}} \leftarrow \mathbf{u}_\theta^{\mathrm{CFG}}(\mathbf{Z}_{\mathrm{in}}, \mathbf{c}^{\mathrm{n}}, t_k; w)$
  $\mathbf{Z}_{t_{k+1}} \leftarrow \mathrm{SchedulerStep}(\mathbf{Z}_{t_k}, \widehat{\mathbf{u}}, t_k \to t_{k+1})$
end for
return $\mathcal{D}(Z^{\mathrm{n}}_{t_N})$

The source noise ϵ^s is fixed throughout sampling. The horizon N_{rep} counts initial replacement steps: $N_{\mathrm{rep}} = 0$ disables replacement, while $N_{\mathrm{rep}} = N$ replaces source inputs at every step. Substitution changes only the denoiser input; the scheduler advances its maintained state $\mathbf{Z}_{t_k}$, which supplies all slots once $k \geq N_{\mathrm{rep}}$. The novel slot follows a single sampling trajectory throughout. The *source-cond.* results in Table 1 use $N_{\mathrm{rep}} = 30$ and $N = 35$ (Sec. C.1).

Algorithm 3 *Joint-generated* inference: joint three-view generation from noise

Require: Fine-tuned $(\theta_{\mathrm{LoRA}}, \phi)$; condition $\mathbf{c}$; novel camera $(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}})$; noise schedule $\{t_k\}_{k=0}^N$; guidance w.

$\mathbf{c}^{\mathrm{n}} \leftarrow (\mathbf{c}, e_\phi(\mathbf{T}^{\mathrm{n}}, K^{\mathrm{n}}))$
Initialize $\mathbf{Z}_{t_0} = (Z^1_{t_0}, Z^{\mathrm{n}}_{t_0}, Z^2_{t_0}) \sim \mathcal{N}(0, I)$ and the scheduler.
for $k = 0, \ldots, N-1$ **do**
  $\widehat{\mathbf{u}} \leftarrow \mathbf{u}_\theta^{\mathrm{CFG}}(\mathbf{Z}_{t_k}, \mathbf{c}^{\mathrm{n}}, t_k; w)$
  $\mathbf{Z}_{t_{k+1}} \leftarrow \mathrm{SchedulerStep}(\mathbf{Z}_{t_k}, \widehat{\mathbf{u}}, t_k \to t_{k+1})$
end for
return $(\mathcal{D}(Z^1_{t_N}), \mathcal{D}(Z^{\mathrm{n}}_{t_N}), \mathcal{D}(Z^2_{t_N}))$

Joint-generated inference uses no observed RGB inputs or source-slot substitution. Table 1 reports both inference modes.

B Proof and Numerical Illustration of Proposition 1

We analyze the scalar, fully visible reduction stated in Proposition 1, conditional on a fixed $\mathbf{c}^{\mathrm{n}}$ and $t \in (0,1)$. The source-view predictions are held fixed, so their loss is constant with respect to the scalar novel-view parameter. With unit masks, the union and intersection coincide; this analysis does not compare the two masking choices. All expectations below are over the clean sample, warping errors, and diffusion noise, or their relevant marginals. Write

$$(\sigma^s)^2 = \mathrm{Var}(E^s) > 0, \qquad \nu = \frac{(\sigma^1)^2 + (\sigma^2)^2}{2}, \qquad D(t) = (1-t)^2 q + t^2 = \mathrm{Var}(Z_t). \tag{12}$$

The two training inputs are $Z_t^{(s)} = (1-t)A^s + t\epsilon$, with the same ϵ in both branches. Removing the constant source term from Eq. (6) gives

$$\mathcal{J}_{\lambda_c,t}(\theta) = \frac{1}{t^2}\sum_{s=1}^{2}\mathbb{E}\Big[\big(f_\theta(Z_t^{(s)},t) - A^s\big)^2\Big] + \lambda_c \mathbb{E}\Big[\big(f_\theta(Z_t^{(1)},t) - f_\theta(Z_t^{(2)},t)\big)^2\Big]. \tag{13}$$

This retains both the sum of the two anchor losses and the t^{-2} weighting of the paper's objective; consistency is in clean-prediction space.

Proof of Proposition 1. First, the chosen affine intercept entails no restriction within the affine class. For a general predictor $\theta z + b$, each supervised residual has mean $b - [1-(1-t)\theta]\mu$. The intercept cancels from consistency, so minimizing over b gives $b = [1-(1-t)\theta]\mu$ and hence $f_\theta(z,t) = \mu + \theta[z - (1-t)\mu]$.

For the clean reference path, joint Gaussianity gives the optimal clean Gaussian denoiser and its conditional variance:

$$\mathbb{E}[x^{\mathrm{n}} \mid Z_t = z] = \mu + \theta_{\mathrm{cl}}[z - (1-t)\mu], \qquad \theta_{\mathrm{cl}} = \frac{(1-t)q}{D(t)}, \qquad \mathrm{Var}(x^{\mathrm{n}} \mid Z_t) = \frac{qt^2}{D(t)}. \tag{14}$$

Moreover, the target velocity is $\epsilon - x^{\mathrm{n}}$, and $(Z_t - f_\theta(Z_t,t))/t - (\epsilon - x^{\mathrm{n}}) = (x^{\mathrm{n}} - f_\theta(Z_t,t))/t$. Thus the weighted clean-prediction error is exactly the velocity-matching risk. Completing its square yields

$$\frac{1}{t^2}\mathbb{E}_{x^{\mathrm{n}},\epsilon}\Big[\big(f_\theta(Z_t,t) - x^{\mathrm{n}}\big)^2\Big] = \frac{D(t)\theta^2 - 2(1-t)q\theta + q}{t^2} = \frac{q}{D(t)} + \frac{D(t)}{t^2}(\theta - \theta_{\mathrm{cl}})^2. \tag{15}$$

In particular, $\mathcal{R}_t^{\mathrm{exc}}(\theta) = D(t)(\theta - \theta_{\mathrm{cl}})^2/t^2$.

For the training objective, independence and centering of the noises give

$$\begin{aligned}\sum_{s=1}^{2}\mathbb{E}\Big[\big(f_\theta(Z_t^{(s)},t) - A^s\big)^2\Big] &= 2\left[\big((1-t)^2(q+\nu) + t^2\big)\theta^2 - 2(1-t)(q+\nu)\theta + q + \nu\right], \\ \mathbb{E}\Big[\big(f_\theta(Z_t^{(1)},t) - f_\theta(Z_t^{(2)},t)\big)^2\Big] &= 2(1-t)^2\nu\theta^2.\end{aligned} \tag{16}$$

The second equality uses $Z_t^{(1)} - Z_t^{(2)} = (1-t)(E^1 - E^2)$: the shared clean sample and diffusion noise cancel. Consequently, defining

$$B_{\lambda_c}(t) = (1-t)^2(q+\nu) + t^2 + \lambda_c(1-t)^2\nu t^2, \tag{17}$$

we obtain

$$\mathcal{J}_{\lambda_c,t}(\theta) = \frac{2}{t^2}\left[B_{\lambda_c}(t)\theta^2 - 2(1-t)(q+\nu)\theta + q + \nu\right], \qquad \theta_{\lambda_c} = \frac{(1-t)(q+\nu)}{B_{\lambda_c}(t)}. \tag{18}$$

The minimizer is unique since the curvature is $4B_{\lambda_c}(t)/t^2 > 0$ for $\lambda_c \geq 0$. Substituting it into the clean excess risk gives

$$\mathcal{R}_t^{\mathrm{exc}}(\theta_{\lambda_c}) = \frac{(1-t)^2\nu^2t^2\big[1 - \lambda_c q(1-t)^2\big]^2}{D(t)B_{\lambda_c}(t)^2}. \tag{19}$$

At $\lambda_c = \lambda_c^\star(t) = 1/[q(1-t)^2]$, $B_{\lambda_c^\star}(t) = (q+\nu)D(t)/q$, so $\theta_{\lambda_c^\star(t)} = \theta_{\mathrm{cl}}$ and the excess risk is zero. Since $B_{\lambda_c}(t)$ strictly increases with λ_c, the optimizer θ_{λ_c} strictly decreases through θ_{cl} at $\lambda_c^\star(t)$. Equation (15) therefore implies that excess risk strictly decreases before this strength and strictly increases after it.

For comparison, let $A^{\mathrm{M}} = \pi A^1 + (1-\pi)A^2$ for fixed $\pi \in [0,1]$, and denote the three anchor-error variances by

$$r_1 = (\sigma^1)^2, \qquad r_2 = (\sigma^2)^2, \qquad r_{\mathrm{M}} = \pi^2(\sigma^1)^2 + (1-\pi)^2(\sigma^2)^2. \tag{20}$$

Each r_j is positive. Minimizing the single-target weighted supervision for each $j \in \{1, 2, \mathrm{M}\}$ yields

$$\theta_j = \frac{(1-t)(q+r_j)}{(1-t)^2(q+r_j)+t^2}, \qquad \mathcal{R}_t^{\mathrm{exc}}(\theta_j) = \frac{(1-t)^2 r_j^2 t^2}{D(t)\left[(1-t)^2(q+r_j)+t^2\right]^2} > 0. \tag{21}$$

This proves Eq. (8). Finally, $\mathcal{R}_t^{\mathrm{exc}}(\theta_{\lambda_c})$ is continuous in $\lambda_c \geq 0$, and the minimum of the three baseline risks is strictly positive. The strict comparison therefore persists on an open interval containing $\lambda_c^\star(t)$. □

Interpretation and scope. SymRegFlow combines supervision from both noisy anchors with shared-noise consistency. Supervision alone fits the inflated anchor variance $q + \nu$; consistency penalizes sensitivity to anchor disagreement and, at the calibrated strength, recovers the clean-reference optimum. Stronger regularization further shrinks the denoiser's slope below the clean optimum, increasing excess risk. Zero excess risk means attaining the minimum total risk $q/D(t)$, not zero reconstruction error. The comparison concerns clean-reference evaluation, not the numerical values of different training objectives. It holds for the stated affine, independent-error surrogate with unit masks and fixed source predictions; it does not assert the same guarantee for the full neural network or one constant λ_c across all diffusion times.

Numerical illustration. To instantiate the viewpoint-dependent error scale from Sec. 2.3, set $\sigma^s = \kappa(E_b + d^s)$, with $d^1 = \alpha$, $d^2 = 1-\alpha$. Figure 6 evaluates Eqs. (19) and (21) using $\mu = 1.2$, $q = 1$, $t = 0.5$, $E_b = 0.4$, $\kappa = 0.6$, and merge weight $\pi = 1-\alpha$. At $\alpha = 0.5$ the single-anchor risks coincide; replacing α by $1-\alpha$ exchanges them and preserves both the merged and SymRegFlow curves. In all three cases $\lambda_c^\star = 4$, since the calibration depends on q, t rather than the anchor-error variances.

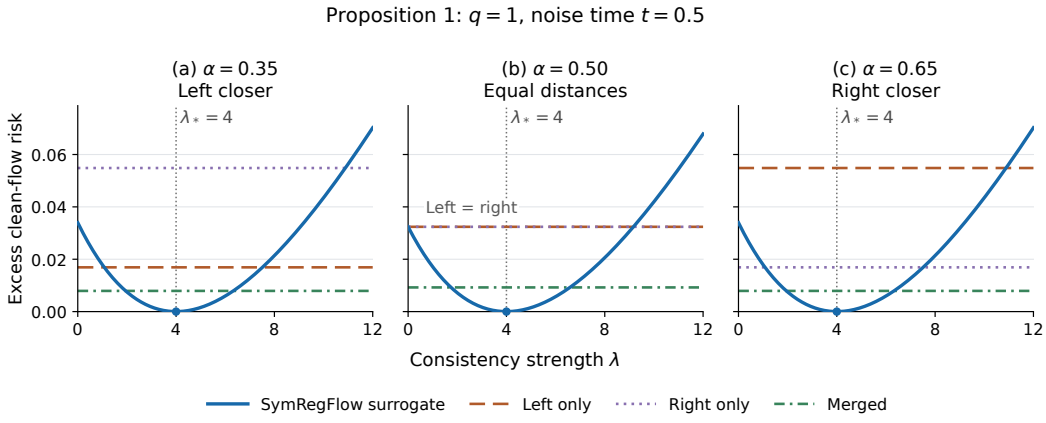


Figure 6: Analytic clean-reference excess risk for Proposition 1 at $\alpha \in \{0.35, 0.50, 0.65\}$. The horizontal baselines use the optimal unregularized predictors; the SymRegFlow curve uses the optimal predictor for each consistency strength (λ in the plot denotes λ_c). The midpoint makes left/right risks equal, and mirrored viewpoints exchange them. SymRegFlow has zero excess risk at $\lambda_c^\star = 4$ and strictly improves on all three baselines in a neighborhood of that strength. These are evaluations of the surrogate formulas, not additional MNIST measurements.

C Implementation Details and Additional Ablation Results

We first describe the shared implementation and study-specific evaluation protocols for the main video experiments and ablations, then present additional quantitative and attention-bias analyses.

C.1 Implementation Details

Backbone and adaptation. All video experiments use Cosmos-Transfer2.5 (Ali et al., 2025) with 29-frame clips at 1280×720 resolution per view. Fine-tuning uses rank-96 LoRA with LoRA alpha 12, gray fill for invalid warp regions, and the continuous-view conditioning described below. Dual-anchor variants regularize clean predictions; the default SymRegFlow configuration uses DCVA and $\lambda_c = 0.05$. For the quantitative nuScenes experiments and ablations, training uses 16 GPUs and a global batch size of 32. LoRA and view-embedding learning rates are 5×10^{-4} and 5×10^{-3}, respectively, with cosine decay, 25 warm-up steps, and weight decay 0.01. Reported quantitative checkpoints are evaluated at step $2{,}048$.

Novel-view pose conditioning. Cosmos-Transfer2.5 (Ali et al., 2025) conditions each view on a world-scenario video composed of projected lane/map polylines and 3D object bounding boxes, rather than captured RGB footage, to specify scene layout and viewpoint. During training, we render this representation at the novel camera poses and feed its encoded latents into SymRegFlow through the backbone's control branch. Both Cosmos-Drive-Dreams (Ren et al., 2025a) and nuScenes (Caesar et al., 2020) provide the map annotations, object geometry, and camera calibration required for this rendering.

To minimize modifications to the pretrained backbone, we retain its integer-indexed embeddings for existing rig cameras. Our autonomous-driving data synthesis setting requires novel views interpolated along the six-camera ring. We therefore parameterize the continuous view embedder by the endpoint camera identities and a scalar ring coordinate, with additional geometric conditioning supplied by the world-scenario renderings. The pretrained camera embeddings remain frozen; Fourier features of the global ring coordinate drive a zero-initialized, endpoint-vanishing MLP residual added to the linearly interpolated embeddings ($\alpha = 0.5$ at the midpoint; Sec. C.5).

This design specializes the general pose-conditioned formulation to the interpolation setting evaluated here. The formulation is compatible with arbitrary target-pose inputs: e_ϕ can instead encode camera extrinsic and intrinsic matrices, Plücker rays, or other pose representations and be trained jointly during fine-tuning.

Sampling. The nuScenes results in Tables 1 and 2 use 35 denoising steps, guidance scale 3, and seed 42. Table 1 reports both *joint-generated* inference, without observed-RGB source-slot replacement, and *source-cond.* inference, with replacement during the first 30 steps. Table 2 uses *joint-generated* inference only.

Distributional metrics. Novel-view FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) compare generated clips with unique captured source-view clips as *unpaired* distributions; captured midpoint RGB is unavailable, and world-scenario renderings are never used as RGB ground truth. FID uses 2,048-dimensional Inception features, and FVD uses 400-dimensional I3D features from 29-frame clips, pooling all scenes in each study.

Cross-view consistency. MEt3R (Asim et al., 2025) projects FeatUp-upsampled DINO features through MASt3R geometry and averages cosine distances over usable overlap and valid synchronized frame pairs. Images preserve aspect ratio with a maximum long edge of 512 pixels. Valid pairs require at least 5% raw overlap and 1% usable overlap after excluding relative depth disagreements above 0.10. The reference views and temporal inputs differ between studies, as specified below.

C.2 Evaluation Protocols

Thirty-scene comparison. Table 1 pools outputs from eight scene-group adapters covering 30 nuScenes scenes, with six midpoint camera pairs per scene and frames 0–115 at 10 fps. FVD uses

720 nonoverlapping 29-frame clips per distribution, starting at frames 0, 29, 58, and 87; FID pools their 20,880 frames per distribution. VBench (Huang et al., 2024) evaluates the 720 generated clips independently, using all 29 frames of each clip. FVMD (Liu et al., 2024) uses 180 sequences of 116 frames at 10 fps per distribution for all six methods. Our sequences concatenate four independently generated 29-frame clips; baselines retain the first 116 frames of each released video. We bilinearly resize frames to 256×256 and use PIPS2 to track 400 points with 16 iterations per 16-frame window. A stride of one frame yields 18,180 windows per distribution, including windows crossing generation joins. We concatenate each window's velocity and acceleration histograms and compute a single Fréchet distance between the generated novel-view and unpaired captured-source feature distributions, pooling all windows across scenes. Instance preservation (IP) (Zheng et al., 2025) uses these same 180 116-frame sequences per model and the official fine-tuned Qwen2.5-VL-3B-Instruct entity-anomaly detector. Each sequence is split into overlapping 2-second clips with a 1-second stride, retaining all 20 frames per clip. Following the official duration rule, the ten clips start at seconds 0–9 and cover frames 0–109, including all three generation joins; the final six frames are omitted. A sequence scores one only if every clip is classified as free of unnatural entity merging, duplication, or disappearance, and zero otherwise. We report 100 times the mean over all 180 sequences, with higher scores indicating better preservation. MEt3R compares novel views with synchronized existing-view videos: *generated* videos for the *joint-generated* result (†) and *captured* videos for all other rows. These settings measure internal cross-view consistency and agreement with observations, respectively. Each uses 41,760 candidate frame pairs per model, averaging its valid subset.

Four-scene ablation. Each variant in Table 2 uses one adapter trained jointly on scene-0012, scene-0013, scene-0014, and scene-0015 with a $100{,}000$-step optimization schedule and evaluated at step $2{,}048$. Evaluation uses the same scenes and four segment starts, yielding 96 novel clips (2,784 frames) and 192 generated endpoint clips per model. Novel-view FID/FVD use 96 unique captured source clips; existing-view metrics use paired captured endpoints. VBench uses the 96 separate 29-frame clips, rather than 116-frame sequences. MEt3R compares novel views with *generated* endpoints over 192 clip pairs and 5,568 synchronized frame pairs.

C.3 Additional Ablation Results

Variants and checkpoint selection. A_L and A_R use individual source anchors; A_M merges both RGB anchors using their visibility masks. All three use DCVA. The dual-anchor variants C_γ use four-stream DCVA with $\gamma \in \{0, 0.01, 0.05, 0.1, 0.5, 1, 2\}$. At $\lambda_c = 0.05$, $C^{\mathrm{FA}}_{0.05}$ replaces DCVA with full attention, changing both attention and training execution. All 11 reported variants are evaluated at step $2{,}048$; the configured $C_{0.001}$ run lacks outputs at this checkpoint and is omitted. The following analyses use the same four-scene measurements as Table 2.

Hybrid radar summary. Figure 7 summarizes the five metrics in Table 2: existing-view FVD; novel-view FVD, VBench subject and background consistency, and MEt3R. Ranks and min–max bounds use all $N = 11$ table variants, while the display contains the seven C_γ models in the consistency-weight sweep. For direction-aware rank r, $S_{\mathrm{rank}} = 100(N + 1 - r)/N$; if S_{mm} is the direction-aware min–max score on a 0–100 scale, the displayed radius is $S_{\mathrm{hyb}} = (S_{\mathrm{rank}} + S_{\mathrm{mm}})/2$. $C_{0.05}$ has the largest mean profile in the sweep, 92.4, followed by C_0 at 71.3 and $C_{0.1}$ at 61.2. This descriptive mean weights the five spokes equally and is not polygon area or a replacement for the individual metrics.

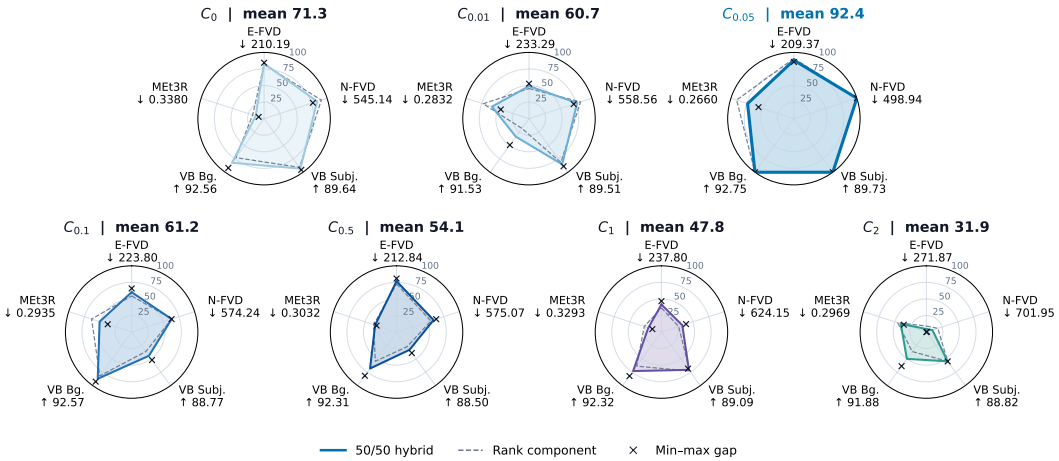


Figure 7: Consistency-weight profiles using the five metrics in Table 2, normalized against all 11 variants at iteration 2,048. Outward is better. Solid polygons show the 50/50 rank–min–max hybrid, dashed polygons its rank component, and crosses its min–max component. E/N denote existing/novel views; VB denotes VBench. Spoke values follow the table's precision and scaling (VBench $\times 100$); the mean is the arithmetic average of the five normalized spokes.

Computational-cost probe. On one NVIDIA RTX 5880 Ada GPU, we compare backbone costs using 28 DiT blocks, width 2,048, 16 heads, rank-96 LoRA, BF16, batch size two, and activation checkpointing at every block. Randomly initialized models receive $8 \times 16 \times 16$ tokens per view and retain both anchor branches until backward. DCVA shares source-view computation, whereas full attention recomputes it for each anchor. After one warm-up iteration, medians over three forward/backward runs are 3.07 and 4.51 seconds, respectively ($1.47\times$). This reduced-resolution probe excludes the control network, VAE, rotary embeddings, data loading, optimizer, and distributed communication; it is not an end-to-end training benchmark.

C.4 Attention-Bias Analysis

In a controlled one-scene diagnostic, each anchor strategy is trained for 300 iterations and sampled with identical noise, prompt, and world-state controls. These measurements characterize attention routing and are distinct from the four-scene benchmark in Table 2.

Cross-view routing. We measure the absolute route gap $G_{\mathrm{xv}} = |\bar{\ell}_{\mathrm{n}\to\mathrm{l}} - \bar{\ell}_{\mathrm{n}\to\mathrm{r}}|$, where $\bar{\ell}$ is the mean post-RoPE, pre-softmax QK logit over the captured denoising steps, DiT blocks, heads, and width, height, and time offset profiles. Table 3 reports gaps of 0.729 and 0.865 for A_L and A_R, respectively, with preferences aligned to their fine-tuning anchors. SymRegFlow reduces the gap to 0.307, a reduction of 57.8% and 64.5%, respectively.

Table 3: Attention bias in the controlled one-scene overfitting experiment. The cross-view columns report the absolute novel-to-left/right QK route gap G_{xv} (lower is more symmetric), averaged over three relative-offset axes. The within-frame columns report invisible-to-visible attention mass $P_{\mathrm{U}\to\mathrm{V}}$ under an identical paired visibility mask; Δ and the relative excess are single-anchor minus SymRegFlow.

Paired anchor	Cross-view route gap G_{xv}			Within-frame mass $P_{\mathrm{U}\to\mathrm{V}}$			
	Single	SymRegFlow	Reduction	Single	SymRegFlow	Δ	Relative excess
Left	0.729	0.307	57.8%	0.1745	0.1447	+0.0298	+20.6%
Right	0.865	0.307	64.5%	0.2091	0.1877	+0.0214	+11.4%

Figure 8 retains the relative key–query offset to show where the imbalance arises, reporting post-RoPE, pre-softmax QK compatibility along the latent width, height, and time axes. The left-only model raises the novel-to-left profile relative to the right-only model, while the right-only model raises the novel-to-right profile. SymRegFlow remains substantially closer across the two routes. These profiles average the same five denoising steps, all 28 DiT blocks, and all attention heads used in Table 3.

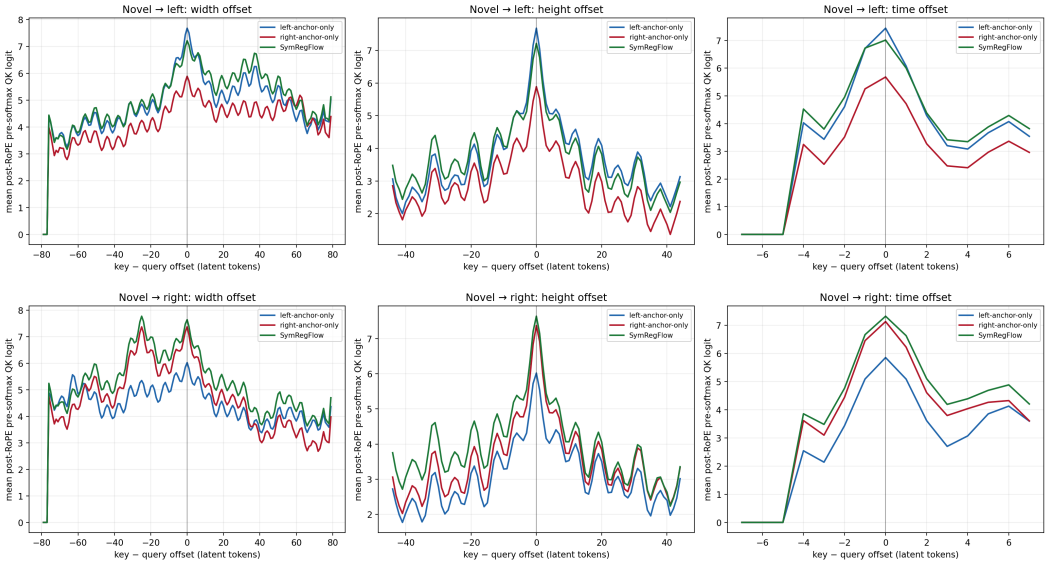


Figure 8: Cross-view relative-offset QK profiles for novel-to-left (top) and novel-to-right (bottom) routes. Curves are count-weighted means over valid token pairs and are further averaged over captured denoising steps, DiT blocks, and heads. Single-anchor adaptation preferentially raises compatibility with its fine-tuning source; SymRegFlow reduces this source-dependent route imbalance.

Figure 9 retains the denoising step at zero spatial offset. The signed gap reverses with the selected single anchor and remains source-aligned throughout denoising, whereas SymRegFlow stays much closer to zero. Because these measurements are attention diagnostics rather than controlled interventions on Q/K, they support a routing-bias explanation but are not by themselves a causal proof of the repeated-content artifact.

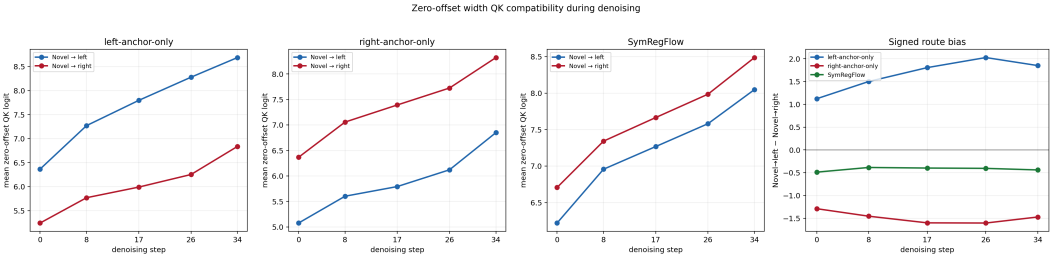


Figure 9: Zero-offset cross-view QK compatibility over denoising. The rightmost panel reports the signed route bias, novel-to-left minus novel-to-right. Its sign follows the fine-tuning anchor for both single-anchor models, while SymRegFlow has a smaller, stable residual imbalance.

Within-frame routing. Within the novel-view frame, we measure the same-frame softmax mass $P_{\mathrm{U}\to\mathrm{V}}$ from all unambiguously invisible (inpainted) query tokens U to anchor-visible keys V. Each comparison applies the same left- or right-anchor visibility mask to the single-anchor model and SymRegFlow. As reported in Table 3, A_L and A_R assign 20.6% and 11.4% more mass to the corresponding visible regions, respectively. Thus, the cross-view preference persists as biased routing from generated content toward source-supported content inside the novel view.

Local repeated-content diagnostic. Figure 5 uses generated frame 28, with both attention panels measured at denoising step 17 and DiT block 20. Red boxes mark queries in invisible (inpainted) regions; cyan boxes mark the strongest same-size, non-overlapping attention match restricted to the fine-tuning anchor's visible region (green). Heat maps show same-frame attention relative to uniform attention, only on visible support. A_M and SymRegFlow are shown as clean RGB controls without attention measurements; the dashed line separates them from the two attention-analyzed single-anchor panels.

C.5 Global Continuous Novel-View Encoding

Let $\mathbf{e}^k \in \mathbb{R}^D$ denote the frozen pretrained embedding of rig camera k, and assign each camera a fixed circular coordinate $\xi^k \in [0, 1)$. For the six-camera rig, front, front-left, rear-left, rear, rear-right, and front-right are assigned coordinates $0, \frac{1}{6}, \ldots, \frac{5}{6}$, respectively. For adjacent cameras a and b and interpolation factor $\alpha \in [0, 1]$, we define the shortest signed displacement and global novel-view coordinate as

$$\begin{aligned} \Delta^{a,b} &= \mathrm{wrap}(\xi^b - \xi^a), \\ \xi^{a,b}(\alpha) &= \left(\xi^a + \alpha\Delta^{a,b}\right) \bmod 1, \end{aligned} \tag{22}$$

where $\mathrm{wrap}(x) = ((x + \frac{1}{2}) \bmod 1) - \frac{1}{2}$. Thus, the interpolation factor is disambiguated by its position on the rig rather than represented in a pair-local space. We encode this circular coordinate using seam-continuous fixed harmonics,

$$\psi(\xi) = \left[\cos(2\pi k\xi), \sin(2\pi k\xi)\right]_{k=1}^{N_{\mathrm{harm}}}, \tag{23}$$

with $N_{\mathrm{harm}} = 4$ in our implementation.

The continuous novel-view embedding is anchored to the pretrained camera-embedding space:

$$\mathbf{e}^{a,b}(\alpha) = (1-\alpha)\mathbf{e}^a + \alpha\mathbf{e}^b + 4\alpha(1-\alpha)r_\phi\left(\psi(\xi^{a,b}(\alpha))\right), \tag{24}$$

where $r_\phi : \mathbb{R}^{2N_{\mathrm{harm}}} \to \mathbb{R}^D$ is a small MLP whose output layer is initialized to zero. The first two terms preserve the geometry of the pretrained embedding space, while the learned residual models smooth deviations along the continuous camera ring. The envelope $4\alpha(1-\alpha)$ vanishes at both endpoints, guaranteeing $\mathbf{e}^{a,b}(0) = \mathbf{e}^a$ and $\mathbf{e}^{a,b}(1) = \mathbf{e}^b$.

D Additional SE(2) Toy Study

Implementation. The base joint image transformer is trained for $8{,}000$ iterations on paired MNIST views at $-90°$ and $0°$, using a batch size of 512 and learning rate 2×10^{-4}. Following JiT (Li & He, 2026), the toy implementation replaces t by $\max(t, 0.05)$ when converting clean predictions or targets to velocities, giving the supervised weight $1/\max(t, 0.05)^2$. Every continuous-pose variant initializes from this same checkpoint and is fine-tuned in full for $2{,}000$ iterations, with batch size 256 and learning rate 2×10^{-4}. The target angle is sampled uniformly through $\alpha \sim \mathcal{U}(0, 1)$ and supplied through the learned scalar pose embedding $e_\phi(\alpha)$. All variants use the same optimizer settings, anchor generator, and source-preservation term. Denoising-output SymRegFlow uses Directed Cross-View Attention (DCVA), as defined in Sec. 2.2, and evaluates its two anchor-conditioned branches through the exactly equivalent four-stream factorization $[x^1, A^1, A^2, x^2]$. We sweep $\lambda_c \in \{0, 0.1, 0.25, 0.5, 1, 2, 4\}$ and select $\lambda_c = 2$ by the mean exact-target FD over the 45 pose–distortion cells.

Baseline construction. For each sampled pose, SymRegFlow uses the two pseudo-examples $\widetilde{\mathbf{x}}^{(s)} = (x^1, A^s, x^2)$ for the paired forward passes in Eq. (3). The A^1-only and A^2-only baselines each use the corresponding single pseudo-example; Merged anchor uses $(x^1, A^{\mathrm{M}}, x^2)$, where A^{M} is the mask-weighted fusion of both anchors.

Consistency regularization weight sensitivity. Figure 10 reports the complete sweep. Each thin curve averages the nine interior target poses at one fixed distortion base E_b; each thick curve averages all five bases, and therefore all 45 pose–distortion cells. On exact targets, PSNR and SSIM

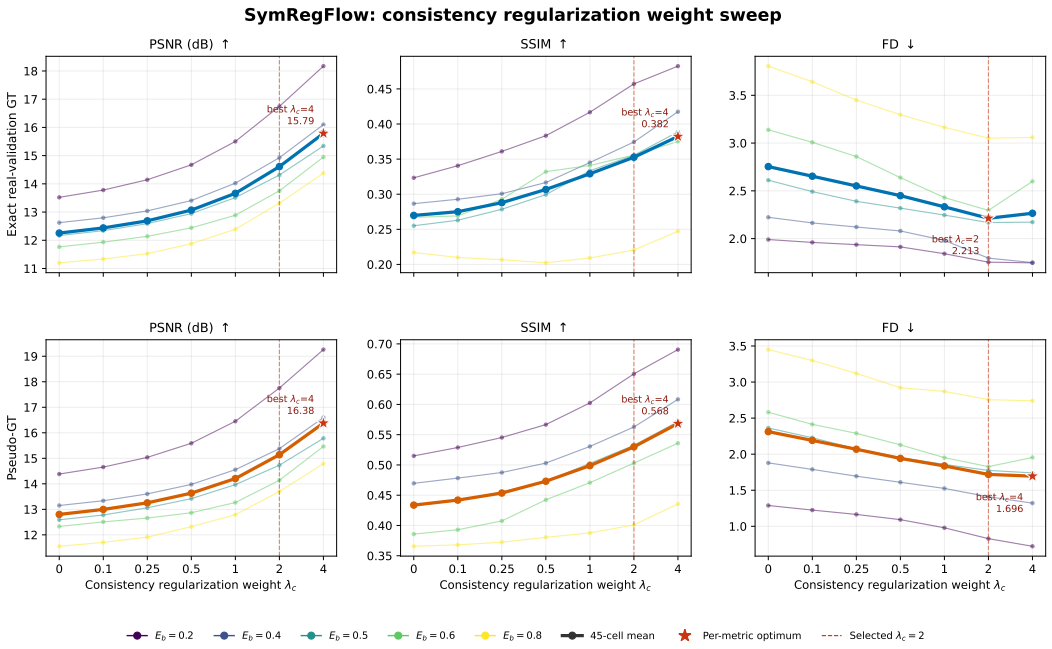


Figure 10: SymRegFlow consistency regularization weight sweep against exact held-out rotations (top) and pseudo targets (bottom). The dashed line indicates the selected $\lambda_c = 2$; stars mark the best sampled weight for each metric. FD denotes pixel-space Fréchet distance on 16×16 pixel features.

continue increasing through $\lambda_c = 4$, reaching 15.79 and 0.382, respectively. Exact-target FD instead falls from 2.754 at $\lambda_c = 0$ to its minimum of 2.213 at $\lambda_c = 2$, then rises to 2.265 at $\lambda_c = 4$. We therefore select $\lambda_c = 2$ using the exact-target FD criterion. All three pseudo-target metrics continue improving through $\lambda_c = 4$, consistent with stronger coupling fitting the imperfect anchors more closely; their divergence from exact-target FD motivates selection on withheld exact rotations.

Anchor corruption. The angular distances from x^1 and x^2 to the target at $\vartheta_\alpha = -90°(1-\alpha)$ are $90°\alpha$ and $90°(1-\alpha)$, respectively. Normalizing by $90°$ gives $d^1(\alpha) = \alpha$ and $d^2(\alpha) = 1-\alpha$. Let E_b denote the base distortion level. For $s \in \mathcal{S} = \{1, 2\}$, we set the scale of E^s to $E_b + d^s(\alpha)$, confidence to $w^s = 1 - d^s(\alpha)$, and loss mask to $m^s = w^s \bar{m}^s$, where $\bar{m}^s$ is the binary sampling-validity mask. The warp perturbs morphology, anisotropic scale and shear, angle, translation, a smooth deformation grid, and pixels with additive noise. Their magnitudes scale with $E_b + d^s(\alpha)$, using respective coefficients $3°$, 0.035, 0.055, and 0.040 for the last four perturbations. This construction is deliberately heterogeneous: the two anchors contain errors that are neither identical nor independent in every pixel, while their expected severity increases with source–target pose distance.

Evaluation protocol. We evaluate the nine interior poses $\alpha \in \{0.1, \ldots, 0.9\}$ at base distortion levels $E_b \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. Each pose–distortion cell contains 256 samples generated with a 50-step Heun solver. The same base Gaussian samples (seed 2027) are reused across methods and poses. PSNR and SSIM compare *source-cond.* predictions with exact rotations of held-out MNIST validation images. Pixel-space Fréchet distance (FD) compares *joint-generated* target-slot samples with the corresponding validation distribution using flattened, bilinearly downsampled 16×16 pixel features; no Inception features are used. Each method–distortion setting currently has one fine-tuning seed; variation across base distortion levels therefore measures robustness to anchor quality rather than optimization uncertainty.

E Additional NuScenes Qualitative Examples

Figures 11–14 provide additional qualitative comparisons of novel-view generation on nuScenes. Rows show DiST-4D, OmniRe, GEN3C, FreeVS, and SymRegFlow from top to bottom; columns show frames from each generated video.



Figure 11: Qualitative novel-view comparison on nuScenes scene-0012 between the front and front-right cameras.



Figure 12: Qualitative novel-view comparison on nuScenes scene-0013 between the front and front-right cameras.



Figure 13: Qualitative novel-view comparison on nuScenes scene-0014 between the front and front-right cameras.



Figure 14: Qualitative novel-view comparison on nuScenes scene-0014 between the back and back-left cameras.