WORLDCEPTION: UNIFYING PERCEPTION AND GENERATION FOR WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual world modeling requires both inferring structure from observations and generating visual predictions under structural conditions. Although generative priors support visual perception, accurate perceptual readout alone does not establish that its representations can guide synthesis. A central challenge is to make dense perceptual fields and sparse states serve both purposes within a unified model. We introduce **WorldCeption**, a compact framework that unifies perception and conditional generation through a shared dense–sparse interface. Given RGB video, WorldCeption predicts dense perceptual representations and continuous or discrete sparse states; conversely, it synthesizes RGB video conditioned on dense representations or sparse states. Built upon a pretrained video transformer, the model combines shared attention with role-dependent feed-forward computation and task-dependent information access, accommodating perceptual readout and conditional synthesis within the same network. These complementary mappings can be composed, enabling estimated states to guide generation and perceptual representations to be translated through an RGB intermediate. Experiments show that estimated camera states retain reconstruction quality close to annotated conditions, while changing supplied states redirects recovered camera motion. Composed perceptual and generative paths further enable cross-representation completion without direct pairwise supervision. WorldCeption offers a compact architectural extension toward perception-centric world modeling through reusable visual representations. Code and models will be available upon acceptance.

1 INTRODUCTION

Visual world modeling connects observations with predictions of how a scene may evolve. Learned dynamics support behavior through imagined experience (Ha & Schmidhuber, 2018; Hafner et al., 2019), while generative simulators extend this direction toward controllable visual environments (Yang et al., 2023; Bruce et al., 2024). A complementary question concerns the explicit representations linking observation and synthesis. Perception recovers scene geometry, object structure, and camera motion; generation synthesizes how these properties appear in video. These capabilities are often developed and evaluated separately, yet their connection is essential: a depth field or camera trajectory should not only describe an observation, but also provide an explicit condition on what a model generates. This motivates a perception-centric question: *can a shared video model infer structured representations and reuse them as conditions for synthesis?*

Generative priors offer a foundation for this connection: diffusion-based perception transfers synthesis priors to geometry (Ke et al., 2024; Hu et al., 2024), and GenCeption (Wang et al., 2026) shows that such priors can support heterogeneous dense and sparse perception. However, accurate readout alone does not establish that the resulting representations can guide generation. Conversely, spatial and camera conditioning make supplied structure useful for synthesis (Zhang et al., 2023; He et al., 2024), but do not by themselves establish its estimation and reuse within the same model.

Generalist visual models already demonstrate shared interfaces across heterogeneous tasks and modalities (Lu et al., 2023; Bachmann et al., 2024). Our focus is therefore not bidirectionality alone, but a reusable dense–sparse interface within a pretrained video generator. Establishing this interface requires more than adding tasks: the model must accommodate both spatially dense fields and compact states, distinguish estimating an unknown state from using a supplied one, and reconcile

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.

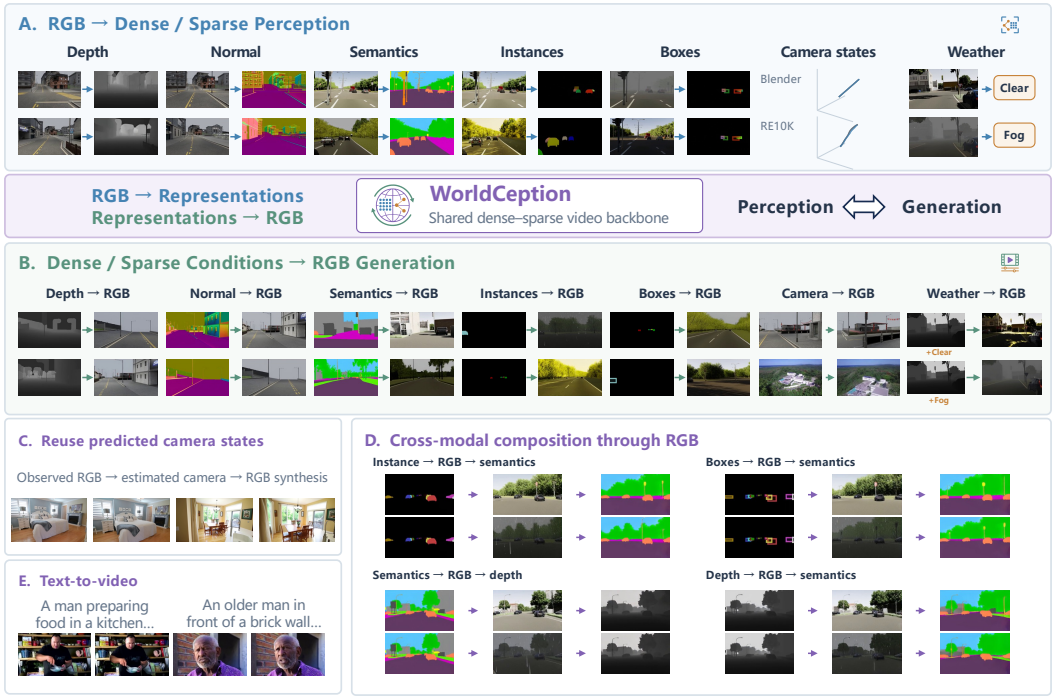


Figure 1: **WorldCeption connects perception and generation.** RGB video maps to dense fields and sparse states, which also condition RGB synthesis; trained directions compose through shared interfaces. Examples are experimental outputs; categorical states describe weather and illumination.

global perceptual evidence with protected generative conditions. The central challenge is therefore to connect *state estimation* and *condition use* within shared computation.

We introduce **WorldCeption**, a compact extension of a pretrained video transformer that unifies these directions through a dense–sparse interface (Figure 1). In the **perceptual direction**, RGB video is mapped to dense fields, such as depth, normals, and segmentation, or to sparse states. In the **generative direction**, these representations and available visual context condition RGB video synthesis. Sparse states may be *continuous*, describing camera motion, or *discrete*, describing weather and illumination categories. Supplied values act as controls, while learned queries request predictions. Native text-to-video synthesis remains part of the task family.

Shared attention connects dense video tokens and compact sparse tokens; *role-dependent feed-forward computation* distinguishes state readout from control, and *task-dependent attention visibility* reconciles perceptual evidence aggregation with protected conditioning. The resulting mappings can be composed: estimated camera states guide synthesis, and an RGB intermediate connects perceptual modalities without direct supervision of every conversion pair.

Our evaluation tests perceptual accuracy, same-checkpoint state reuse, and directional control through state intervention. Our contributions are:

- A bidirectional dense–sparse interface connecting video perception and conditional generation, with continuous and discrete sparse states.
- A shared video-transformer architecture with role-dependent FFNs and task-dependent visibility for perceptual queries and generative controls.
- An evaluation of individual and composed mappings, including estimated state reuse and cross-representation completion without direct pairwise supervision, supported by external comparisons and controlled ablations.

2 RELATED WORK

Generative priors and generalist perception. Generative pretraining has emerged as a foundation for visual perception. Marigold and GeoWizard adapt diffusion priors to depth and normal estimation (Ke et al., 2024; Fu et al., 2024), while GenPercept and Lotus explore deterministic or single-step dense prediction (Xu et al., 2024; He et al., 2025). DepthCrafter and NormalCrafter extend this paradigm to temporal geometry (Hu et al., 2024; Bin et al., 2025), and GenCeption unifies dense and sparse video perception within a pretrained generative backbone (Wang et al., 2026). In parallel, generalist models organize heterogeneous tasks through shared representations: Pix2Seq adopts a sequence interface, Painter and SegGPT use image-formatted predictions, and Unified-IO and 4M-21 accommodate multiple modalities (Chen et al., 2022; Wang et al., 2023b;c; Lu et al., 2023; Bachmann et al., 2024). Geometry-focused models, including Depth Anything V2, DUSt3R, and VGGT, provide strong references for perceptual accuracy (Yang et al., 2024; Wang et al., 2023a; 2025). WorldCeption connects perceptual estimation to conditional synthesis: dense fields and sparse states serve not only as predictions, but also as inputs to the generative process.

World models and conditional video generation. World models learn representations of environmental dynamics for prediction and control. World Models, PlaNet, and Dreamer operate through learned latent states, while IRIS models tokenized observations autoregressively (Ha & Schmidhuber, 2018; Hafner et al., 2018; 2019; 2023; Micheli et al., 2023). DINO-WM and V-JEPA 2 explore prediction and planning in pretrained feature spaces (Zhou et al., 2024; Assran et al., 2025). At the visual level, UniSim, Genie, DIAMOND, and GameNGen develop interactive environments through conditional generation (Yang et al., 2023; Bruce et al., 2024; Alonso et al., 2024; Valevski et al., 2025); Diffusion Forcing and History-Guided Video Diffusion study temporal conditioning, and Nano World Models explores compact future prediction (Chen et al., 2024; Song et al., 2025; Huang et al., 2026). GPDiT combines autoregressive latent-frame prediction with diffusion-based generation (Zhang et al., 2025). Related work expands the conditioning interface through spatial guidance and camera control (Zhang et al., 2023; He et al., 2024), perceptual modalities (Huang et al., 2025), and joint modeling of actions and visual outcomes (Shen et al., 2025; Ye et al., 2026; Kim et al., 2026). WorldCeption investigates a complementary connection: whether explicit representations estimated from observations can themselves guide synthesis. This connects perceptual readout and generative conditioning through compatible dense and sparse interfaces.

Shared architectures and specialized computation. Unifying heterogeneous capabilities requires balancing shared representations with specialized computation. Gato extends shared sequence modeling across perception and action, while Cosmos 3 brings understanding, generation, and action into a broader unified framework (Reed et al., 2022; NVIDIA et al., 2026). Adapters and LoRA demonstrate compact adaptation of pretrained models (Houlsby et al., 2019; Hu et al., 2021), and mixture-of-experts architectures allocate computation through learned routing (Shazeer et al., 2017; Fedus et al., 2021). WorldCeption instead specializes computation according to explicit token roles: dense representations, supplied sparse controls, and sparse prediction queries interact through shared attention but follow role-dependent feed-forward paths. Task-dependent attention visibility further reconciles perceptual evidence aggregation with conditional generation. Together, these mechanisms support bidirectional dense–sparse modeling within a shared video backbone.

3 METHOD

3.1 A BIDIRECTIONAL DENSE–SPARSE INTERFACE

WorldCeption builds on latent diffusion and video generative pretraining (Wan Team, 2025), using a shared transformer backbone (Peebles & Xie, 2023). Let O denote an RGB video, D^m a dense video representation of modality m, and S a sparse state representation, whose values may be continuous or discrete. The interface supports two directions:

$$\underbrace{O \to D^m, \quad O \to S,}_{\text{perception}} \qquad \underbrace{D^m \to O, \quad (C, S) \to O}_{\text{conditional generation}}. \tag{1}$$

Here C is available visual context: observed video frames for camera conditioning, or an aligned dense video condition for categorical appearance control. Mappings are task-selected, not simultaneous prediction of every modality. Dense fields preserve spatiotemporal layout; sparse states encode either

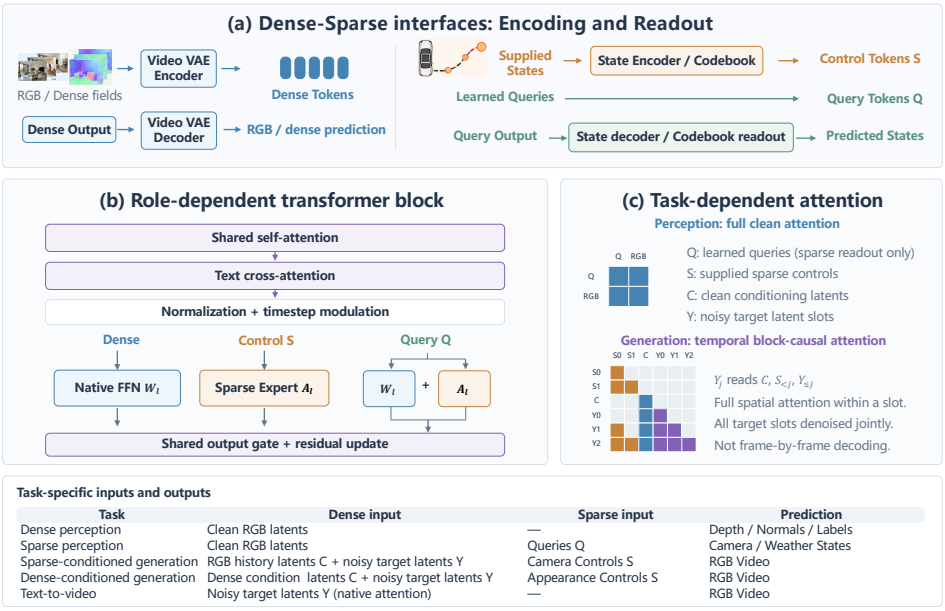


Figure 2: **WorldCeption's internal adaptation for perception and generation.** (a) Dense latents and sparse values/queries form a common interface. (b) Shared attention precedes role-dependent FFNs; control and query are alternative sparse roles. (c) Perception reads global evidence; camera-conditioned generation protects history and restricts temporal visibility. The block and temporal mask depict the principal continuous-state configuration.

continuous quantities, such as camera poses, or discrete categories, such as weather and illumination. Both state types support estimation and conditioning through the shared sparse-token interface, with type-specific encoding and readout.

Representations as reusable interfaces. We distinguish three operational requirements: *estimability*, recovering a representation from observations; *condition use*, producing outputs that follow a supplied representation; and *reuse*, consuming an estimated representation through the same conditional interface. For a supported representation type r, write

$$\widehat{R}^r = P^r_\theta(O), \qquad \widetilde{O} = G^r_\theta(C, R^r, \epsilon), \qquad \widehat{O} = G^r_\theta(C, P^r_\theta(O), \epsilon). \tag{2}$$

Here P^r_θ and G^r_θ denote task-selected mappings with shared backbone parameters, and ϵ denotes sampling noise. These are evaluation criteria, not additional losses or a sufficient-state assumption. Condition use fixes C and ϵ; reuse replaces annotations with estimates. Sequential composition through RGB is distinct from simultaneous multi-condition generation.

Dense representations. A fixed codec converts each dense modality to a three-channel video, encoded by the pretrained video VAE E. Its temporal compression factor r_t determines the latent time grid. RGB, geometry, and rendered labels share the backbone's spatial tokenization and dense output head. Modality-specific codecs and validity conventions are specified in the appendix.

Continuous sparse states. A schema-specific encoder embeds numerical values into the transformer width. When those values are unknown, learned queries occupy the corresponding positions; a lightweight decoder reads their predictions from the final hidden states. For cameras, each six-dimensional state comprises translation and an axis–angle rotation relative to the initial camera frame. We standardize states using training statistics and concatenate each state with its temporal difference. Packing consecutive states aligns these vectors with the VAE time grid (Appendix A). Differences are computed in the standardized representation, not interpreted as physical velocities.

Discrete sparse states. Discrete states describe observed weather and illumination. An input category selects learned codebook embeddings; an output query is classified against the same prototypes. State values are discrete, while their embeddings are continuous. Together with numerical encoding, this supports both state types through a shared sparse interface (Appendix F).

3.2 ROLE-DEPENDENT FEED-FORWARD SPECIALIZATION

Within each transformer block, dense and sparse tokens interact through shared attention projections. Their feed-forward computation depends on whether a token encodes a known control or requests an unknown prediction. Let H_l denote the hidden sequence after self-attention and text cross-attention in block l. Native normalization and timestep modulation produce

$$U_l = (1 + \gamma_l(t)) \odot \mathrm{LN}(H_l) + \beta_l(t). \tag{3}$$

For a token u with role ρ, we compute

$$F_l(u;\rho) = \begin{cases} W_l(u), & \rho = \text{dense}, \\ A_l(u), & \rho = \text{control}, \\ W_l(u) + A_l(u), & \rho = \text{query}, \end{cases} \tag{4}$$

$$H_{l+1} = H_l + g_l(t) \odot F_l(U_l;\rho), \tag{5}$$

where W_l is the pretrained FFN and A_l a narrow, block-specific MLP. For sparse queries, the expert is parallel to the native FFN: both receive the same normalized, modulated input and contribute through the same gate. For controls, only the expert contributes to the FFN update; shared attention and the residual stream still connect controls to the backbone. The expert parameters within a block are shared across the two sparse roles. Routing is fixed by the known sequence contract, requiring neither a learned router nor a load-balancing objective.

Controls specify constraints; queries extract unknown structure. The expert specializes these roles while retaining shared visual interaction. Nonzero output initialization activates the control branch from the first update (Appendix A).

3.3 TASK-DEPENDENT INFORMATION FLOW

Perception integrates evidence across clean observations. Prediction must distinguish observed information from the noisy target being generated. We implement this distinction through attention visibility within the shared backbone.

For camera-conditioned generation, the sequence is $[S \mid C \mid Y_t]$: sparse controls, clean visual history, and noisy future. The directional visibility is

$$S_i \leftarrow S_{\le i}, \quad C \leftarrow C, \quad Y_j \leftarrow [S_{<j}, C, Y_{\le j}], \tag{6}$$

Here $S_0, \ldots, S_{K-1}$ encode successive state groups and $Y_0, \ldots, Y_K$ include a boundary latent Y_0 with no sparse input. Arrows specify visible keys; attention remains spatially unrestricted within each permitted latent slot. Equation 6 describes the history-isolated camera path. Dense aligned-condition synthesis uses its corresponding role mask; neither path lets noisy targets write into clean conditions. The temporal restriction concerns latent computation, not an autoregressive sampling procedure or physical causality of decoded pixels.

Why restrict attention? Later targets could relay later controls to earlier targets across blocks. The triangle in Equation 6 closes this indirect path. With history, text, and initial noise fixed independently of controls, token-local FFNs and normalization preserve the restriction by induction over blocks and sampling steps: Y_j depends only on $S_{<j}$ and preceding/current target slots. This is latent-prefix noninterference, not decoded-frame independence through a temporally coupled VAE. Sparse perception uses full attention so queries can read the complete clean video. Dense perception similarly aggregates global clean-video context. Where sparse tokens are present, their known input/output role selects the FFN route. Native text-to-video generation retains the original video-only path.

3.4 TASK-LOCAL SUPERVISION

Dense perception follows GenCeption's feed-forward parameterization (Wang et al., 2026). Given clean observations and instruction c_m, we train the native output as a negative target latent:

$$\hat{z}_m = -f_\theta(E(O), 0, c_m), \qquad \mathcal{L}_{\text{dense}} = \|\hat{z}_m - E(D^m)\|_2^2. \tag{7}$$

This is a supervised reuse of the native output head, rather than an identity of the general flow velocity. Sparse perception minimizes

$$\mathcal{L}_{\text{sparse}} = \|g_\phi(H_Q) - S\|_2^2, \tag{8}$$

with numerical targets expressed in their normalized schema. Conditional generation uses flow matching (Lipman et al., 2022), interpolating a clean target sequence Z with noise, $Y_t = (1 - \sigma_t)Z + \sigma_t\epsilon$, and minimizes

$$\mathcal{L}_{\text{gen}} = \|f_\theta(S, C, Y_t, t, c) - (\epsilon - Z)\|_2^2. \tag{9}$$

Sampled tasks supervise their designated outputs; categorical readout uses the tied-codebook objective in Appendix F. Shared parameters are optimized across sampled tasks without cycle-consistency or cross-task agreement losses; composition reuses these interfaces at inference.

Operational composition. State reuse instantiates Equation 2 with the first RGB frame as context. For dense representations, $\hat{D}^b = P_b(G_a(D^a, \epsilon))$ connects trained directions through an RGB latent without pairwise supervision. Since D^a can leave appearance unspecified, completion and constraint preservation are distinct from reconstruction of a unique paired scene.

4 Experiments

We evaluate WorldCeption as a generalist model for video perception and conditional generation, then examine state reuse, cross-representation composition, and the contribution of role-dependent computation.

4.1 Experimental setup

Model and data. WorldCeption adapts Wan2.1-T2V-1.3B (Wan Team, 2025). Training uses six data sources: RE10K for camera estimation and control; VKITTI2 for dense perception and conditional synthesis; Blender for geometry and camera tasks; TartanAir for depth and camera estimation; MVS-Synth for depth; and an OpenVid subset for native text-to-video replay (Zhou et al., 2018; Cabon et al., 2020; Wang et al., 2020; Huang et al., 2018; Nan et al., 2024). The adaptation corpus contains 7,967 distinct 81-frame RGB windows: 7,455 support perception and 6,796 support generation, with 6,284 shared between directions, comprising 497,379 distinct target frames. We render 428 Blender sequences (63,788 RGB frames) across 13 scene templates spanning urban driving, indoor scenes, and procedural environments. The model jointly learns perception and generation from pretrained Wan weights; the final training stage adds discrete-state readout with multitask replay. Dataset distributions, counting conventions, and training details appear in Appendices B and A. The final perception, generation, and weather evaluations share one set of weights without task-specific fine-tuning.

Evaluation protocols. Task-native perception evaluations contain 460 validation and 192 Blender training-set records (Appendix C); records can reuse RGB across tasks. Scores average within source and then across sources. External comparisons use the same final weights: 96 VKITTI2 depth windows, 64 RE10K camera windows, and five Sintel sequences per rendering pass. State reuse evaluates 32 RE10K sources under four conditions and two paired noise seeds. Protocols and uncertainty estimates appear in Appendix E.

4.2 Estimating dense structure and sparse state

Figures 3–4 illustrate dense scene perception and continuous camera translation estimation. On VKITTI2, WorldCeption obtains semantic mIoU of 0.542/0.426 on ID/scene-held-out clips, instance foreground IoU of 0.805/0.806, and rendered-box F1 of 0.456/0.277 (Table 20). These outputs are read directly from the shared model. Appendix P defines the foreground-overlap and rendered-box metrics. The ID/OOD contrast is task-dependent: foreground overlap remains similar, whereas semantic and box scores decline. Thus recovering object regions does not imply equally robust semantic labeling or localization. Continuous camera readout and discrete weather classification further test whether the shared interface extends beyond image-formatted outputs.

Comparison with external models. Table 1 compares outputs under common local protocols. Depth Anything V2-L and VGGT provide more accurate depth and camera estimates, respectively. Against released GenCeption, WorldCeption reduces Clean normal mean error but not median or threshold errors; Final also favors GenCeption. These comparisons distinguish stand-alone estimation accuracy from the ability to reuse estimates for synthesis. Additional datasets appear in Appendix C. Discrete

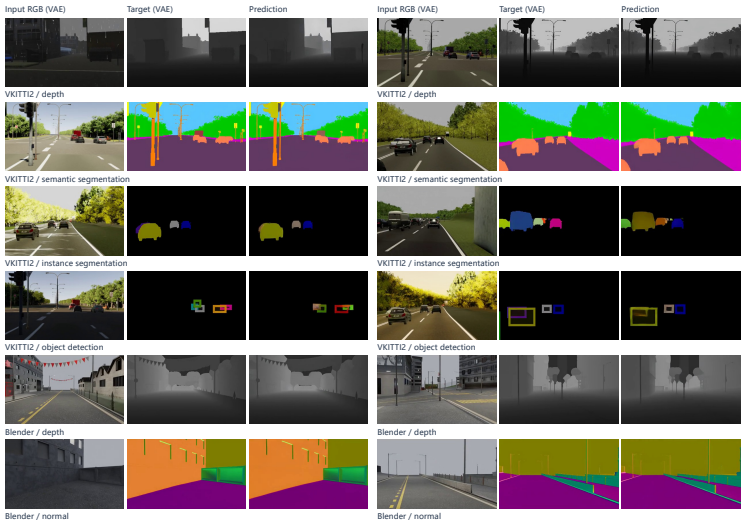


Figure 3: **Dense perception.** Qualitative examples on VKITTI2 and Blender (training clips), with two examples per task. Triplets: VAE-reconstructed RGB, VAE-reconstructed target, and prediction.

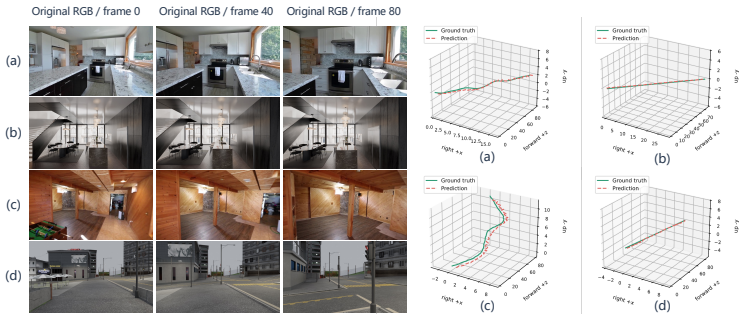


Figure 4: **Camera translation estimation.** Two examples each from RE10K and Blender pair three RGB frames with predicted/GT trajectories. Axes are independently scaled and clip-relative; splits and plotting conventions appear in Appendix N.

weather readout achieves 79.17% balanced accuracy (91.67% ID; 66.67% OOD), versus 16.67% with zero RGB (Appendix D).

4.3 Using dense and sparse representations as conditions

Figure 5 illustrates dense-conditioned synthesis and discrete weather control. Across two seeds, correct conditions yield lower paired-target latent error than wrong controls on all evaluated generation tasks. On RE10K, dynamic-history camera control gives 0.219 versus 0.624; on VKITTI2 ID, the range is 0.185–0.370 versus 0.728–0.892. These diagnostics measure condition sensitivity; full results and definitions appear in Appendix D. The controls test complementary uses of the interface: camera states specify viewpoint evolution, whereas weather categories change appearance with dense structure held fixed. For the latter, the wrong-control test changes the category, not the dense field; it therefore measures sparse-control use within dense-conditioned synthesis.

Figure 6 verifies that native iterative T2V sampling remains executable after unified post-training. An experimental Euler sampler attached to released GenCeption fails on these prompts.

4.4 Ablation Analysis

Estimated state as a generative condition. We fix the final generator, first frame, text, and noise, varying only the camera state: GT, WorldCeption-estimated, VGGT-estimated, or a fixed wrong-

Table 1: **External perception comparisons.** Source-balanced scores from common local protocols, not copied published benchmark results. Normals show mean angular error; full metrics appear in the appendix. Post-training data exclude foundation-model pretraining and use each work's reported unit; N/R means not reported as a single count. Our parameter count is measured; baseline sizes are reported by their authors.

VKITTI2 depth	Params	Post-train data	Unit	AbsRel ↓	RMSE ↓	δ_1 ↑	Eval. frames
Depth Anything V2-L	335M	62M	images	**0.0570**	**4.1560**	**0.9628**	7,776
WorldCeption	1.44B	1.65K / 497K	videos / frames	0.1146	6.3225	0.8828	7,776
RE10K camera	**Params**	**Post-train data**	**Unit**	**Rot. (°)↓**	**Dir. (°)↓**	**Joint@30↑**	**AUC@30↑**
VGGT	1B	N/R	—	**0.0434**	**11.86**	**0.9433**	**0.7146**
WorldCeption	1.44B	1.65K / 497K	videos / frames	0.1641	23.10	0.7686	0.4103
Sintel normals	**Params**	**Post-train data**	**Unit**	**Clean (°)↓**		**Final (°)↓**	
GenCeption	1.3B	≥7.5K / ≥607K	videos / frames	32.46		**30.26**	
WorldCeption	1.44B	1.65K / 497K	videos / frames	**30.54**		33.67	

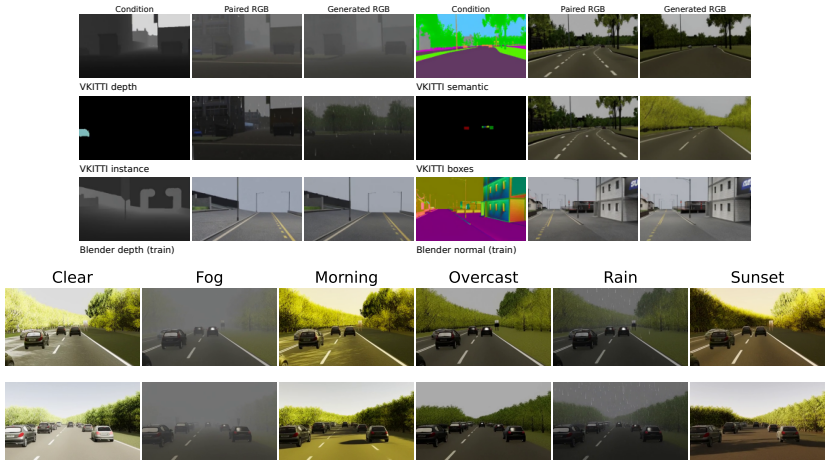


Figure 5: **Representation-conditioned generation.** Top: six reverse mappings (Blender uses training clips). Bottom: weather control, with fixed depth and six categories for each of two geometries.

source trajectory. Across 32 sources and two paired seeds, all 256 videos complete generation and motion recovery. Estimation sees the full observed clip; generation receives no future RGB or estimator features. This tests state-mediated reconstruction, not future-state forecasting. Frozen VGGT recovers generated motion for comparison with both the original GT and the supplied condition.

Estimated-state synthesis approaches GT-conditioned reconstruction (Table 2): PSNR changes by -0.116 dB and LPIPS by $+0.0013$, with paired source-bootstrap intervals containing zero. This supports a small observed reuse cost, not statistical equivalence. Wrong-source states increase direction error to original GT from $20.81°$ to $59.80°$; their error to the supplied state is $25.00°$. Thus the state affects generated motion rather than serving only as an auxiliary prediction. Better standalone camera accuracy does not automatically give better synthesis: VGGT states do not improve these reconstruction means. Appendix E reports uncertainty, state normalization, and recovery limitations.

Cross-representation completion through RGB. We test 12 directed pairs among depth, semantic, instance, and box representations without pairwise training: 96 conversions over eight camera–weather settings. Generated RGB supplies contextual cues absent from sparse annotations (Figure 7); Appendix J reports the direct baseline and input-constraint audit. Unlike state reuse, this connects different representation types through generated visual context. It completes unspecified content rather than recovering a unique scene from an incomplete representation.

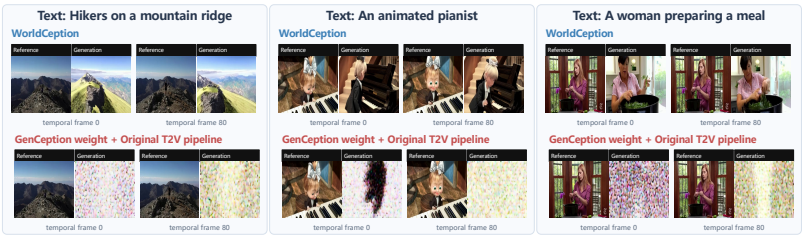


Figure 6: **Preserving text-to-video generation.** Pairs show caption-reference / generated frames (0 and 80), using text-only conditioning. At 40 steps, WorldCeption produces coherent videos; GenCeption with experimental Euler sampling yields noise-like outputs (Appendix R).

Table 2: **Camera-state reuse.** 32 sources, two seeds. Dir.-GT/Sup.: direction error (°) to GT/supplied state. Appearance excludes frame 1; real-video recovery is a reference.

Supplied state	PSNR↑	LPIPS↓	Dir.-GT↓	Dir.-Sup.↓	Joint@30↑
GT	16.588	0.2294	21.07	21.07	0.826
WorldCeption	16.471	0.2307	20.81	21.07	0.832
VGGT	16.217	0.2363	24.89	23.70	0.777
Wrong source	—	—	59.80	25.00	0.421
Real-video recovery	—	—	8.50	—	0.945

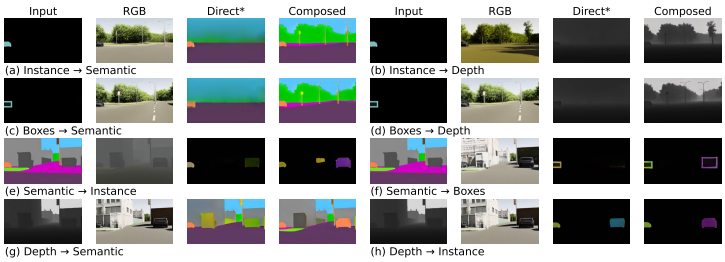


Figure 7: **Cross-representation completion through generated RGB.** Eight selected directions show the input, generated RGB, and perceptual readout; no pairwise mapping is trained.

Table 3: **Sparse FFN configurations on RE10K validation.** Diagnostic definitions are given in Appendix P.

Sparse route	State MSE ↓	Trans. RMSE ↓	Target gap ↑	Output change ↑
Shared only	**0.068**	5.510	0.001	0.005
Shared + expert	0.136	5.129	0.005	0.013
Expert only	0.266	**4.196**	**0.158**	0.317
Role-dependent	0.086	4.663	0.101	0.292

Feed-forward routing. We compare four FFN configurations under the same data and update budgets (Table 3). Shared-only routing reads out well but responds weakly to controls; role-dependent routing better balances both objectives. Expert-only routing yields the strongest control response but the largest state-readout MSE. Retaining the native FFN for queries while specializing controls addresses this trade-off; no configuration wins every diagnostic. Implementation details appear in Appendix N.

5 Conclusion

WorldCeption unifies video perception and generation through a compact dense–sparse interface that reads, reuses, and intervenes on perceptual states while preserving T2V sampling—a minimal perception-centric world model.

REPRODUCIBILITY STATEMENT

We specify the data units, split boundaries, representation codecs, task sampling, optimization budgets, and counterfactual evaluation protocol in the main text and appendix. The supplementary material will include configuration files, preprocessing contracts, checkpoint hashes, and commands for reproducing the reported experiments. We report source videos, windows, and training exposures separately to avoid treating overlapping clips as independent data.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising manuscript sections, improving language and organization, identifying and summarizing related literature, and preparing figures and tables. AI assistance also supported discussions of the conceptual framing, experimental design, evaluation protocols, code development, and interpretation of results, including qualitative inspection of model outputs. These tools served as research and writing aids rather than independent evidence for scientific claims. The authors retain responsibility for verifying AI-assisted text, references, code, and analyses, and for the accuracy and integrity of the final submission.

ETHICS STATEMENT

Generated videos and inferred states can be plausible while violating input constraints or physical reality. We report such failures and do not present this prototype as a validated controller for safety-critical deployment. Model and data use inherit the licensing and privacy constraints of the underlying pretrained systems and datasets.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for World Modeling: Visual Details Matter in Atari. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.12399.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, et al. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities. *arXiv preprint arXiv:2406.09406*, 2024. URL https://arxiv.org/abs/2406.09406.

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. NormalCrafter: Learning Temporally Consistent Normals from Video Diffusion Priors. *arXiv preprint arXiv:2504.11427*, 2025. URL https://arxiv.org/abs/2504.11427.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, et al. Genie: Generative Interactive Environments. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.15391.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *European Conference on Computer Vision*, pp. 611–625, 2012. URL https://files.is.tue.mpg.de/black/papers/ButlerECCV2012.pdf.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. URL https://arxiv.org/abs/2001.10773.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.01392.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. A Unified Sequence Interface for Vision Tasks. *arXiv preprint arXiv:2206.07669*, 2022. URL https://arxiv.org/abs/2206.07669.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv:2101.03961*, 2021. URL https://arxiv.org/abs/2101.03961.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, et al. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. *arXiv preprint arXiv:2403.12013*, 2024. URL https://arxiv.org/abs/2403.12013.

David Ha and Jürgen Schmidhuber. World Models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. *arXiv preprint arXiv:1811.04551*, 2018. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv preprint arXiv:1912.01603*, 2019. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. *arXiv preprint arXiv:2404.02101*, 2024. URL https://arxiv.org/abs/2404.02101.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, et al. Lotus: Diffusion-based Visual Foundation Model for High-quality Dense Prediction. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.18124.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. *arXiv preprint arXiv:1902.00751*, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. *arXiv preprint arXiv:2409.02095*, 2024. URL https://arxiv.org/abs/2409.02095.

Jiehui Huang, Yuechen Zhang, Xu He, Yuan Gao, Zhi Cen, Bin Xia, et al. UnityVideo: Unified Multi-Modal Multi-Task Learning for Enhancing World-Aware Video Generation. *arXiv preprint arXiv:2512.07831*, 2025. URL https://arxiv.org/abs/2512.07831.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning Multi-view Stereopsis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1804.00650.

Siqiao Huang, Partha Kaushik, Michael Chen, Hengkai Pan, Kaiwen Geng, Omar Chehab, Fernando Moreno-Pino, and Max Simchowitz. Nano World Models: A Minimalist Implementation of Future Video Prediction. *arXiv preprint arXiv:2605.23993*, 2026. URL https://arxiv.org/abs/2605.23993.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.02145.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, et al. Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning. *arXiv preprint arXiv:2601.16163*, 2026. URL https://arxiv.org/abs/2601.16163.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *arXiv preprint arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2206.08916.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are Sample-Efficient World Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.00588.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation. *arXiv preprint arXiv:2407.02371*, 2024. URL https://arxiv.org/abs/2407.02371.

NVIDIA et al. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, et al. A Generalist Agent. *arXiv preprint arXiv:2205.06175*, 2022. URL https://arxiv.org/abs/2205.06175.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538*, 2017. URL https://arxiv.org/abs/1701.06538.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. VideoVLA: Video Generators Can Be Generalizable Robot Manipulators. *arXiv preprint arXiv:2512.06963*, 2025. URL https://arxiv.org/abs/2512.06963.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-Guided Video Diffusion. In *Proceedings of the International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.06764.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion Models Are Real-Time Game Engines. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.14837.

Wan Team. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.11651.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, et al. Video Generation Models are General-Purpose Vision Learners. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.09024.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. *arXiv preprint arXiv:2312.14132*, 2023a. URL https://arxiv.org/abs/2312.14132.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, et al. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. URL https://arxiv.org/abs/2003.14338.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b. URL https://arxiv.org/abs/2212.02499.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Segmenting Everything In Context. *arXiv preprint arXiv:2304.03284*, 2023c. URL https://arxiv.org/abs/2304.03284.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What Matters When Repurposing Diffusion Models for General Dense Perception Tasks? *arXiv preprint arXiv:2403.06090*, 2024. URL https://arxiv.org/abs/2403.06090.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.09414.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators. *arXiv preprint arXiv:2310.06114*, 2023. URL https://arxiv.org/abs/2310.06114.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, et al. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*, 2023. URL https://arxiv.org/abs/2302.05543.

Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Bo Wang, Haoyang Huang, Jianlong Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 38, pp. 135524–135546, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c565514eb8cf8808d82a612c5f7a1920-Abstract-Conference.html.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. *arXiv preprint arXiv:2411.04983*, 2024. URL https://arxiv.org/abs/2411.04983.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. *arXiv preprint arXiv:1805.09817*, 2018. URL https://arxiv.org/abs/1805.09817.

APPENDIX

A REPRODUCIBILITY AND ADDITIONAL PROTOCOLS

Architecture. The backbone has 30 blocks, hidden width 1536, and 12 attention heads. Each block has an independent 1536–128–1536 GELU expert. Its nonzero output initialization is drawn using an isolated random stream. The frozen VAE uses strides $(4, 8, 8)$. An 81-frame video yields 21 latent temporal positions. Eighty boundary-relative six-dimensional poses are normalized using training statistics, augmented with increments, and packed into 20 sparse values of width 48. All backbone weights remain trainable; this is not a claim of parameter-efficient fine-tuning.

Geometry preprocessing. The accepted Blender geometry release audits 428 source videos and 63,788 native frames. Validity is recovered from raw geometric support, including regions incorrectly excluded by older masks. Scalar depth is resized with validity-weighted area interpolation before clip-median normalization and the mapping $0.25 \log(1 + d_{\text{rel}})$. A bounded nearest-valid halo (16 pixels at 256p) fills the immediate invalid boundary; remaining invalid regions use relative depth one. Filling is an encoding convention, never additional ground truth. Normals use validity-weighted vector resizing and unit renormalization. VAE reconstruction is evaluated separately to expose codec limitations. Older label releases are not matched baselines for data effects.

Final unified configuration. Training starts from pretrained Wan weights and comprises 15,500 updates followed by 2,000 updates incorporating categorical readout, at global batch 64. Initially, task sampling allocates 50% to controlled generation, 10% to native T2V replay, 20% to camera readout, and 20% to dense perception. The final stage allocates 25% to controlled-generation replay, 15% to paired weather generation, 10% to T2V, 20% to camera readout, 15% to dense perception, and 15% to weather readout. Camera readout uses RE10K, Blender, and TartanAir; external depth uses TartanAir and MVS-Synth. The six-source inventory and its cross-task overlap are detailed in Appendix B. The schedule contains 1.12 million task draws, not independent videos. Paired weather-generation draws use two target videos, so task draws are not identical to target-video passes. The OpenVid subset is used for native-generation replay, not full-dataset training. Evaluations from earlier training budgets retain their original provenance below until the final evaluation replaces them.

Earlier three-source evaluation configuration. The earlier reference is trained for 20,000 updates jointly on RE10K, VKITTI, and Blender. Historical depth/context diagnostics and the 96-conversion two-hop study use this configuration; the intermediate same-checkpoint state-reuse pilot is specified in Appendix O. The final-model external comparisons and expanded state-reuse evaluation are specified in Appendix E. Historical architecture tests and the following depth-expansion model are separate runs; they are not steps along a single training trajectory. This reference uses global batch 64, for 1.28 million sample exposures, not 1.28 million independent videos. RE10K/VKITTI and Blender each receive half of the exposures; generation and perception account for 60% and 40%. Its training inventory contains 3,916 RE10K windows per camera direction, 120 VKITTI windows per direction, 1,884 Blender windows per camera direction, and 2,248 Blender windows per dense direction. Directions often reuse the same underlying video. Available camera history is sampled 70% static and 30% dynamic, with 10% camera-condition dropout. VKITTI dense-to-RGB tasks supply the six weather category IDs through the existing sparse control codebook. Weather query readout is not part of this model's training schedule; accepting category inputs does not establish calibrated categorical perception.

Earlier depth-expansion checkpoint and data inventory. The additional model is trained for 5,000 updates. It starts from original Wan weights, independently of the principal model, with global batch 64. AdamW uses learning rates 5×10^{-5} for the backbone and 10^{-4} for the interfaces and experts, 250 warmup updates, zero weight decay, and gradient clipping at one. Camera-condition dropout is 10%. Of all sample exposures, TartanAir depth (Wang et al., 2020) receives 3.75% and MVS-Synth depth (Huang et al., 2018) 2.5%; trajectory tasks receive 7.5% RE10K and 6.25% Blender. Total generation/perception exposure is 60/40. The external datasets contribute 465/217 and 108/12 train/validation clips, respectively. Their held-out evaluation contains 28 clips in total and reports representation PSNR of 22.91 dB (TartanAir) and 24.69 dB (MVS-Synth). These are depth-codec image scores, not raw-depth accuracy.

Routing comparison. Shared-only, additive, expert-only, and role-dependent variants share the same random seed, the RE10K/VKITTI schedule, global batch 64, and 1000 updates. Expert-only and

role-dependent configurations use an isolated nonzero expert output initialization; shared-plus-expert uses zero output initialization. The evaluation manifest hashes, sampler settings, and evaluated update counts are checked for equality by the table export script. Initializations differ across some configurations, so the table is not an initialization-controlled causal test of the routing operation.

Data-composition comparison. The data-composition study compares three independently trained models. A-only and B-only use 1000 updates each; A+B uses 2000, all at global batch 64 with the same role-dependent model. Each active group receives 64,000 sample exposures. Both pure and mixed arms start from Wan. Exposure matching does not match total compute, learning-rate trajectory, or the number of parameter updates against either pure arm. B-only is evaluated on its training tasks and on common RE10K/VKITTI held-out tasks; the latter are cross-domain probes, not a fabricated Blender validation split.

Evaluation protocol. The standard pilot selects up to 16 samples per perception task and eight per generation task with a fixed sampling seed. Generation uses 40 denoising steps, guidance scale three, and shift three. Metrics are macro-aggregated by source where available. Static and dynamic history use separate conditional prediction settings. A wrong condition comes from a different clip; physical zero encodes a stationary trajectory, whereas null removes the condition. Only the correct condition has paired ground-truth future video. Latent reconstruction error under a wrong condition cannot determine whether its alternative motion is physically correct. Detection F1 is obtained from connected components in rendered boxes with IoU threshold .5; it must not be relabeled as average precision.

Trajectory objective diagnostics. The fitting diagnostic uses two independently sampled subsets. Each contains four distinct RE10K and four Blender training sources. Four objective variants per subset start from the extended 5k model and fit the same sample order for 200 updates with batch one. Backbone learning rate is 10^{-5} and interface/expert learning rate is 10^{-4}; all weights are trainable. Reweighted increments use full-training-set scale statistics, changing the effective objective weighting. Endpoint-only fitting supervises poses 4,8,...,80. The ratios in Table 19 divide the mean after-fit translation RMSE by the mean before-fit RMSE within each subset; they do not pool physical distances across datasets into a benchmark score. These seeds vary the fitted subset and are not independent full-training replications.

Evidence provenance and historical results. The manuscript tables are generated from frozen evaluation JSON files with SHA256 provenance in the supplementary evidence directory. Final perception, generation, weather classification, and expanded camera-state reuse share the final weights. The main-text T2V comparison also uses these weights. Historical ablations, weather-control galleries, and compositional demonstrations retain the provenance specified in Appendix R; their results are not pooled with final-model scores. Earlier training configurations are evaluated separately.

B ADAPTATION DATA INVENTORY AND MODEL OVERHEAD

Counting units. Table 4 describes the available adaptation corpus, excluding the backbone's pretraining data. A window is an ordered sequence of 81 RGB frames. We deduplicate identical windows across task directions and modalities; different overlapping windows remain distinct. Perception and generation columns therefore overlap and must not be added. Generation includes native text-to-video replay. A video used for both depth and normal prediction contributes one window, not two independent training videos. Source units differ across datasets: original videos for RE10K and OpenVid, rendered sequences for Blender, sequences for TartanAir and MVS-Synth, and scene–camera–weather streams for VKITTI2. They are not interchangeable with independent scenes.

The perception and generation inventories contain 603,855 and 550,476 frame occurrences, respectively. Their union contains 645,327 occurrences and 497,379 distinct target frames. Adding available visual history gives 704,979 distinct frames. RE10K frames are identified by source video and timestamp; other datasets use native frame identity. These counts describe the available corpus, not a claim that every frame was sampled during optimization. Weather classification and paired weather synthesis reuse the VKITTI2 footage: 120 readout windows and 600 ordered condition-pair records do not introduce additional RGB videos. TartanAir's 465 depth and 679 camera windows overlap, yielding 1,063 distinct windows.

Table 4: **Available training corpus.** P/G denote perception/generation windows. Union deduplicates across tasks. Native frames are deduplicated within source, including all available history; they are not sampler draws.

Dataset	Source units	P windows	G windows	Union	Native frames
RE10K	476	3,916	3,916	3,916	523,475
VKITTI2	48	120	120	120	9,720
Blender	428	2,248	2,248	2,248	63,788
TartanAir	80	1,063	0	1,063	57,776
MVS-Synth	108	108	0	108	8,748
OpenVid	512	0	512	512	41,472
Total	—	7,455	6,796	7,967	704,979

Self-rendered Blender data. We render RGB, depth, normals, and camera poses in Blender, using three groups of scene assets (Table 5). Multiple camera trajectories and temporal windows share scene geometry. Thus the 428 sequences are not 428 independent scenes. The urban and curated-asset sequences contain 161 frames each; the procedural sequences contain 81. Curated assets comprise a living room, a laundry room, a combined living/kitchen scene, and urban alleys. Procedural families comprise corridor, courtyard, human-close, human-full, living-room, park, street, and study scenes. Sampling weights within Blender are 40%, 40%, and 20% for the three groups, rather than proportional to window counts.

Table 5: **Blender corpus distribution.** Windows are deduplicated across tasks; each RGB frame may carry multiple aligned annotations.

Scene group	Templates	Sequences	RGB frames	Windows
Urban driving	1	284	45,724	1,704
Curated indoor/urban assets	4	80	12,880	480
Procedural scene families	8	64	5,184	64
Total	13	428	63,788	2,248

Scope of the data comparison. GenCeption reports 7,500 self-rendered human-centric videos in addition to public datasets (Wang et al., 2026); our self-rendered component contains 428 sequences. This compares rendered sequence counts, not total training data, scene diversity, or matched task accuracy. Our inventory quantifies the reuse of a compact adaptation corpus across perception and generation; it does not establish an overall data-efficiency ratio against GenCeption, whose aggregate public-data frame inventory is not specified.

Architectural overhead. The serialized video backbone contains 1,419.00 million floating-point parameter elements. The sparse experts add 11.85 million and the remaining interfaces add 5.31 million, for 17.16 million additional elements (1.21% of the backbone) and 1,436.16 million in total. This inventory includes retained interface tensors and excludes the VAE, text encoder, and optimizer state. It measures stored model size, not active computation or trainable parameter savings: the backbone is updated during adaptation.

C Completed final-model perception evaluations

All results in Table 6 and the updated Table 20 use the same final-model weights. The 652 task records comprise 460 validation and 192 training-set records; the latter evaluate three Blender tasks on 64 windows each. Distinct render sequences are not necessarily independent scene templates. VKITTI2 source aggregation groups camera streams, averaging weather/window records within each source. The two held-out camera streams belong to one unseen scene, not two independent unseen scenes. RE10K camera estimation covers 64 windows from 53 source videos; TartanAir depth and camera estimation each cover 64 windows from 32 sequences; MVS-Synth contributes 12 depth windows. Each VKITTI2 task covers 48 ID and 16 scene-held-out windows.

Table 6: **Final-model perception.** N counts task windows. Depth PSNR compares decoded model outputs with VAE-reconstructed targets in the three-channel representation. It is not a raw-depth benchmark score. Blender results measure training reconstruction.

Dataset	Task/split	N	Metric	Value
RE10K	Camera / val	64	Relative rotation error (°)	0.824
TartanAir	Camera / val	64	Relative rotation error (°)	6.889
TartanAir	Depth / val	64	Representation PSNR (dB)	24.719
MVS-Synth	Depth / val	12	Representation PSNR (dB)	24.933
VKITTI2	Depth / ID	48	Task-native AbsRel	0.210
VKITTI2	Depth / OOD	16	Task-native AbsRel	0.278
Blender	Depth / train	64	Relative-depth AbsRel	0.031
Blender	Normals / train	64	Mean angular error (°)	5.805
Blender	Camera / train	64	Relative rotation error (°)	0.147

Camera metric scope. Rotation error compares predicted and target orientations relative to the initial camera, not adjacent-frame rotations. An auxiliary translation direction cosine compares position vectors relative to that camera, excluding zero-length vectors: values are 0.9914 for RE10K, 0.8649 for TartanAir, and 0.9976 for Blender. These are not cosines of successive motion increments and must not be substituted for adjacent-pair direction error or joint AUC. Translation RMSE uses clip-scaled coordinates, not metres.

External depth comparison. Table 7 reports the final model on the expanded protocol and the smaller eight-source bridge. These raw-depth scores differ from the representation-space metrics above. Additional external comparisons appear in Appendix E.

Table 7: **Final-model common-protocol VKITTI2 depth.** Source-macro scores after per-image median alignment. Both methods have complete frame coverage. WorldCeption receives 81-frame video; DA-V2-L processes individual frames.

Set	Model	Frames	AbsRel ↓	RMSE ↓	δ_1 ↑
Expanded	WorldCeption	7,776	0.1146	6.3225	0.8828
	DA-V2-L	7,776	0.0570	4.1560	0.9628
Bridge	WorldCeption	648	0.1221	6.7127	0.8524
	DA-V2-L	648	0.0553	4.1062	0.9662

D FINAL-MODEL GENERATION AND DISCRETE-STATE EVALUATION

The final model is evaluated on 576 task windows per seed: 256 VKITTI2, 192 Blender training windows, and 128 RE10K/OpenVid windows. The latter contain 64 camera-control and 64 text-to-video records. Two seeds yield 1,152 task records; control variants reuse these windows rather than adding independent samples. These evaluations use the same weights as final perception and weather classification.

Condition sensitivity. Table 8 averages each seed's source-balanced scores with equal seed weights. The diagnostic is $\|\hat{z} - z_{\mathrm{GT}}\|_2^2/N$, averaged over the N scalar elements of the generated and target VAE latents. Correct supplies the paired control; wrong substitutes another camera trajectory or a different appearance category, as appropriate to the task; null invokes the evaluator's null control branch. Dense fields remain present when appearance controls are changed. Static and dynamic distinguish the stored visual-history inputs. All errors are measured against the original paired target, so wrong-control error does not measure fidelity to a counterfactual target.

Correct controls produce lower paired-target error than wrong controls in every listed group for both seeds. Null conditions remain competitive on Blender depth, normals, and dynamic camera control, and slightly outperform correct controls on OOD instance-conditioned generation. Thus the experiment demonstrates sensitivity to supplied conditions without establishing that conditioning always improves reconstruction. OOD errors and the null-control comparisons delimit that conclusion. Camera-state recovery against supplied trajectories is evaluated separately in Appendix E.

Table 8: **Final generation diagnostics.** Mean latent squared error over two seeds; lower is closer to the original paired target. Blender uses training windows. These scores are not perceptual video-quality metrics.

Condition	Split	History	Correct	Null	Wrong
Blender camera	train	dynamic	0.127	0.123	0.291
Blender camera	train	static	0.147	0.248	0.370
Blender depth	train	—	0.196	0.182	0.480
Blender normals	train	—	0.188	0.170	0.385
RE10K camera	val	dynamic	0.219	0.318	0.624
RE10K camera	val	static	0.240	0.458	0.639
VKITTI2 depth	ID	—	0.185	0.445	0.728
VKITTI2 depth	OOD	—	0.362	0.522	0.959
VKITTI2 instances	ID	—	0.362	0.604	0.883
VKITTI2 instances	OOD	—	0.799	0.787	0.942
VKITTI2 boxes	ID	—	0.370	0.613	0.892
VKITTI2 boxes	OOD	—	0.586	0.813	1.090
VKITTI2 semantics	ID	—	0.193	0.450	0.740
VKITTI2 semantics	OOD	—	0.405	0.504	0.846

Discrete-state readout. Weather classification completes all 96 examples, with 48 ID and 48 OOD examples and six balanced categories. Overall accuracy and balanced accuracy are both 79.17%, with macro-F1 77.59%. The zero-RGB baseline reaches 16.67% balanced accuracy. This supports visually grounded categorical readout within the same model used for continuous states and generation.

Table 9: **Final weather classification.** All scores in percent.

Split	N	Accuracy	Balanced accuracy	Macro-F1
ID	48	91.67	91.67	90.74
OOD	48	66.67	66.67	58.33
All	96	79.17	79.17	77.59

The OOD errors concentrate in clear and morning categories, both confused with sunset; fog, overcast, rain, and sunset have full recall in this sample. The 16 geometry groups provide six weather variants each, with all six classified correctly for 25% of groups. These related variants should not be treated as independent unseen scenes.

E External comparisons and estimated-state reuse

Evaluation scope. All external perception rows in Table 1 and all generated videos in Table 2 use the same final weights. Comparisons use released Depth Anything V2-L, GenCeption 1.3B, and VGGT (Yang et al., 2024; Wang et al., 2026; 2025). Pretraining data, model size, and inference cost are not matched. VKITTI2 and RE10K are training domains for WorldCeption; these are capability comparisons, not zero-shot transfer claims. Exact public pretraining overlap is not established. Sintel is a five-sequence cross-dataset normal evaluation. Table 1 reports adaptation data separately from foundation-model pretraining. WorldCeption uses 1,652 source units and 497,379 distinct target frames across six sources. GenCeption's entry is a lower bound from its reported 7,500-video core synthetic collection (607,500 frames at 81 frames per video); additional task-specific data make the inventories non-identical. Depth Anything V2-L reports 335M parameters and training with 62M pseudo-labeled real images. VGGT reports a 1B model but not a single comparable training-sample count. WorldCeption's 1.44B entry uses the measured 1,436.16M stored parameters, including 17.16M interface parameters (Appendix B); the shared VAE and text encoder are excluded. The backbone's marketed 1.3B designation is not used as an exact parameter count.

Depth and pose. Depth covers 96 windows of 81 frames, yielding 7,776 successful frames per method, grouped into ten scene/camera streams from five base scenes. WorldCeption receives video; Depth Anything is applied framewise. Both use the same raw GT validity mask, spatial field of

view, and per-image median alignment. Scores average within source and then across sources. The eight-source bridge in Table 7 is a smaller evaluation, not an independent replication. Camera estimation covers 64 windows from 53 RE10K sources, with no failed windows. Both methods receive 81 real frames. Rotation and translation direction errors use adjacent poses; Joint@30 is the fraction whose two errors are both below 30°. Appendix G specifies the local AUC convention, which is not VGGT's published benchmark score.

Normals. The five Sintel sequences contribute 240 uniquely owned frames and 92,064,906 valid pixels per rendering pass. Overlapping windows do not duplicate frame ownership. Each sequence contributes equally to the score. Table 21 reports all angular thresholds, not only the favorable Clean mean. A lower mean and a higher median do not by themselves identify the underlying error distribution.

Paired state-reuse design. One window from each of 32 RE10K sources is generated under two paired noise seeds and four state providers: annotated poses, WorldCeption, VGGT, and a fixed wrong-source trajectory. All 256 generations and motion recoveries complete. State estimation observes the source clip, whereas generation receives only its first RGB frame repeated into the native 81-frame input, text, supplied state tokens, and noise. No future RGB or estimator features enter generation. Sampling uses 40 denoising steps, guidance three, and shift three. This is reconstruction through an explicit state interface, not prediction of an unknown future camera trajectory.

State coordinates and recovery. Camera-to-world poses are anchored at the initial camera. VGGT translations are normalized by their own median nonzero adjacent displacement, without fitting to test GT; WorldCeption predictions retain the training interface's native units. Wrong-source trajectories use the same valid state schema but are not separately magnitude-matched. Swapping them changes a camera condition, not an isolated physical variable. VGGT recovers motion at frames $0, 5, \ldots, 80$. Consecutive recovered poses are therefore five input frames apart; translation coverage is 100% for all four conditions.

Table 10: **Complete recovered-motion results.** Source-macro averages over 32 sources and two seeds. Errors are degrees; Joint denotes Joint@30. Real-video recovery applies the same procedure to observed target clips.

State provider	To original GT			To supplied state		
	Rot.↓	Dir.↓	Joint↑	Rot.↓	Dir.↓	Joint↑
GT	0.482	21.07	0.826	0.482	21.07	0.826
WorldCeption	0.498	20.81	0.832	0.499	21.07	0.820
VGGT	0.483	24.89	0.777	0.472	23.70	0.797
Wrong source	1.132	59.80	0.421	0.543	25.00	0.765
Real-video recovery	0.101	8.50	0.945	—	—	—

Appearance and uncertainty. PSNR and LPIPS-Alex exclude the supplied first frame. All 192 evaluations for the three reconstruction conditions complete; wrong-source videos have no paired counterfactual appearance target. We average seeds within source, then sources, and use 10,000 paired source-clustered bootstrap repetitions. Frames and seeds are not treated as independent source replicates.

Interpretation and scope. Estimated states are reusable conditions with a small measured reconstruction cost relative to annotations. Wrong-source motion is closer to the supplied trajectory than to original GT, supporting sensitivity to the state input. This does not establish physical causal identification, translation-magnitude accuracy, or arbitrary camera-control generalization. Real-video recovery is a reference, not a strict error floor; subtracting its error from generated-video error is not a valid error decomposition. VGGT is both a provider and recovery model, introducing a shared-model consistency bias. No direct paired superiority test between WorldCeption and VGGT providers is claimed. The experiment does not isolate joint training from state normalization. Blind-review media cover eight sources, two seeds, and four conditions (64 videos), but ratings are not available; no human preference score is reported.

Table 11: **Paired differences with 95% source-bootstrap intervals.** Motion is measured against original GT. Intervals containing zero do not establish equivalence.

Contrast	Metric	Difference	95% interval
WorldCeption − GT	PSNR (dB)	-0.116	$[-0.345,\ 0.084]$
	LPIPS	$+0.0013$	$[-0.0043,\ 0.0073]$
	Rotation (°)	$+0.016$	$[-0.004,\ 0.039]$
	Direction (°)	-0.258	$[-1.896,\ 1.649]$
	Joint@30	$+0.006$	$[-0.034,\ 0.035]$
WorldCeption − wrong	Rotation (°)	-0.633	$[-0.872,\ -0.421]$
	Direction (°)	-38.99	$[-53.64,\ -25.59]$
	Joint@30	$+0.411$	$[0.299,\ 0.522]$
VGGT − GT	PSNR (dB)	-0.371	$[-0.604,\ -0.140]$
	LPIPS	$+0.0069$	$[-0.0009,\ 0.0156]$
	Direction (°)	$+3.820$	$[1.121,\ 7.460]$
	Joint@30	-0.049	$[-0.096,\ -0.012]$

F CATEGORICAL SPARSE PERCEPTION AND CONTROL

A shared category interface. The discrete-state configuration recognizes the observed clip's weather and illumination and uses the same category vocabulary for synthesis. Its six classes are clear (the dataset's `clone`), fog, morning, overcast, rain, and sunset. Let E_k be the learned control-token embeddings for class k, and h_Q the hidden state of one query reading the clean RGB latent. Readout shares the input codebook:

$$p(k \mid O) = \mathrm{softmax}_k \left(\frac{\langle \mathrm{norm}(\mathrm{LN}(h_Q)), \mathrm{norm}(\mathrm{mean}(E_k)) \rangle}{\tau} \right), \qquad \tau = 0.1. \tag{10}$$

Training uses squared error against a six-dimensional one-hot target. There is no independent free classification decoder. At generation time, the selected category's embeddings form a known sparse prefix alongside an aligned dense condition and the noisy RGB target. One output query reads the category; four input tokens encode a supplied category. Thus a single query does not imply a single-token control prefix.

Historical reference configuration and evidence scope. The following independent reference predates the same-checkpoint consolidation in Appendix O. Its scores and gallery are retained as historical evidence, not the current main-text weather results. The categorical reference is trained independently for 25k updates from the pretrained video backbone, using native FFNs rather than the principal camera model's experts. It jointly trains dense perception, dense-conditioned RGB synthesis, weather recognition, and text-to-video replay. Recognition and weather control use the same categorical-model weights, but not the camera-model weights. Perception uses full clean attention. Dense generation uses a stable-condition mask: sparse controls read a causal sparse prefix, dense conditions read the controls and themselves, and targets read all three groups without writing into clean conditions. There is no camera-time prefix rule for these global categories. The earlier role-dependent model consumed weather IDs but omitted RGB-to-weather readout; the final training schedule adds this task. The historical reference demonstrates the tied category interface; it is not evidence that all reported capabilities have already been evaluated in one checkpoint.

Training and sampling. The task schedule allocates 40% of updates to native text-to-video replay, 40% to four controlled synthesis directions, and 20% to four dense perception directions and weather readout. A synthesis update uses two same-geometry, different-weather examples with shared timestep and noise, averaging their standard flow losses. The category is dropped on 15% of synthesis samples. Inference can then guide the prediction relative to its null-category counterpart. The archived six-weather gallery uses 30 steps and sparse guidance scale two; the separate control-matrix evaluation uses 40 steps and scale five. Gallery and matrix scores must not be conflated.

Recognition protocol. Each split contains 48 distinct geometry–weather clips: eight geometry windows with six appearances each. ID spans four training-domain scenes; the scene-held-out split

contains eight windows from the two Scene20 cameras, not eight independent scenes. Balanced accuracy is macro recall over the six categories; with balanced support it equals accuracy. The all-zero-latent reference keeps the instruction and timestep fixed. Weather recognition is an observed-state task, not meteorological forecasting.

Table 12: Categorical sparse readout using the generation codebook. Both splits contain 48 clips, eight per category.

Input / metric	ID	Scene-held-out
RGB: balanced accuracy (%) ↑	91.67	87.50
RGB: macro F1 ↑	0.907	0.855
Zero latent: balanced accuracy (%)	16.67	16.67



Figure 8: **Sparse weather recognition from RGB.** All six appearances of the first archived ID geometry, at its middle frame. Labels are recorded query predictions, not illustrative assignments. The same categorical model supplied the earlier weather gallery, not the consolidated-model outputs now shown in Figure 5.

G COMMON-PROTOCOL EVALUATION DETAILS

Depth. The crop-v2 protocol center-crops source RGB and raw labels to the same field of view before model-specific resizing, and restores predictions to the GT evaluation coordinates. Raw valid depths lie between 0.001 and 80 meters. The relative DA-V2-L output is converted to positive depth by reciprocal with a fixed 10^{-6} floor; both methods then receive per-image median scale alignment and range clipping. The baseline is the relative model, not the VKITTI-finetuned metric checkpoint. Scores average within source and then across sources: the main set has 7,776 frames and the bridge has 648. No comparison matches the models' pretraining data, parameter counts, or inference cost. The common test is not the published DA-V2 evaluation protocol.

Pose. All poses are converted to camera-to-world matrices and compared at common timestamps. The local adjacent-pair score computes the mean of $\max(0, 1 - \max(e_R, e_t)/30)$, where errors are in degrees. It is not labeled as VGGT's published ten-image AUC. Multiscale diagnostics use pairs $(i, i+k)$ for $k \in \{1, 4, 8, 16, 32\}$ and a separate first-to-all set. Nonzero GT translations define the eligible pair set; invalid predicted translations count as failures for threshold success, while conditional direction means are accompanied by coverage. Prediction coverage is one in the reported diagnostic results. The earlier adjacent scorer also excluded invalid predicted translations, but no such exclusions affect these clips. Depth prediction coverage is likewise one. This is a statement about valid outputs, not successful reconstruction of physical truth.

The following supplementary multiscale diagnostics use the earlier three-source model, not the final-model comparison. Relative rotations and directions are scale-invariant. Sim(3)-aligned camera center RMSE is 0.898 for ours and 0.920 for VGGT in GT trajectory units, not meters. First-to-all direction errors are $3.42°$ and $7.22°$, respectively, whereas rotations favor VGGT ($0.525°$ versus $0.229°$). We do not infer overall superiority from the favorable first-to-all statistic or the small aligned-RMSE difference.

Statistical scope. Frames, pose pairs, camera variants, and conversion edges share underlying sources and are not independent replications. VKITTI uses five base scenes; Scene20 contributes two cameras, not two unseen scenes. Final RE10K camera comparison uses 53 sources, and final state reuse uses 32 sources with two paired seeds. The earlier eight-source diagnostics and four-source

connection set are separately labeled historical experiments. We retain per-source results and avoid significance claims based on correlated frame counts.

H INFORMATION-FLOW DESIGN AND HISTORICAL EVIDENCE

From stable conditions to role-dependent visibility. The initial dense conditional model separated clean conditions from noisy targets: $C \leftarrow C$ and $Y \leftarrow [C, Y]$. Sparse controls and queries were subsequently introduced into the same computation. A full-attention variant allowed all groups to interact, which also allowed clean conditions to depend on paired noisy targets during training. This is a possible shortcut in our setting, not a claim that bidirectional attention is invalid in every conditional diffusion model. Standard noisy-target dependence itself is part of flow training; the concern is whether a supposedly stable condition representation becomes target-dependent and substitutes for useful controls.

Table 13: Full-attention diagnostic variant, not the principal checkpoint. Mass is the measured softmax mass. Hidden change compares shuffled and correct paired targets at block input 29, holding the condition fixed.

Noise timestep	$C \leftarrow Y$ mass	$Y \leftarrow S$ mass	Condition relative L2
50	44.58%	0.0433%	0.3757
500	38.62%	0.0368%	0.1991
950	29.43%	0.0459%	0.0147

These measurements show that target perturbations can alter condition hidden states and that sparse inputs receive little target attention in this probe. Attention mass alone is not an attribution of causal importance. Later role-isolated training restored useful control behavior while also changing pairing, condition dropout, and guidance. We therefore use the probe as mechanistic motivation and do not present recovery scores as a mask-only effect. Historical weather readout, control, and text-to-video demonstrations belong to those earlier checkpoints, not to the principal 20k model.

Current temporal indexing and guarantee. The camera path has $K = 20$ sparse groups for 80 transitions and $K + 1 = 21$ VAE latent slots for 81 frames. With zero-based indexing, Y_0 sees no control, and Y_j sees $S_0, \ldots, S_{j-1}$ for $j > 0$. Every spatial patch inside an allowed slot remains visible. The guarantee in Equation 6 applies to this isolated-history path at the latent slot level. It assumes token-local state embeddings and fixed independent history, text, and noise; global control normalization that changes earlier tokens during an intervention would violate those assumptions. Dense aligned conditions use their task-specific connectivity and are not claimed to have the same temporal guarantee. The compiled block-mask implementation changes execution, not logical visibility.

I EARLIER REFERENCE-MODEL RECONSTRUCTION DIAGNOSTICS

This historical experiment uses the three-source reference. The current same-checkpoint intervention results are in Appendix O.

All conditions share the 20k generator, first-frame-only dense input, task prompt, and sample-specific noise. The first frame is repeated into the native 81-frame VAE input shape; no actual future RGB is provided. Supplied trajectories are encoded using the same sparse schema and training normalization. Our predictions already use its units; VGGT trajectories are anchored at their first pose and divided by their own median nonzero adjacent displacement, without fitting to test GT. Mismatched conditions cyclically exchange GT trajectories between the selected sources. Their motion magnitudes are not separately matched, so this test does not isolate path shape from amplitude. A supplied trajectory is available in full as a control even though the attention graph restricts its temporal influence.

The generator uses 40 flow steps, guidance three, and shift three. Lossless eight-bit output arrays supply pixel metrics and the camera evaluator; compressed display videos do not supply the scores. VGGT reconstruction uses frames $0, 5, \ldots, 80$. Consequently each successive recovered pair spans five original frames, not one. No separate matched 17-frame reconstruction of real videos is available

in this export, so recovered motion errors include evaluator uncertainty. The VGGT-provider row is auxiliary because the same model family measures generated motion. LPIPS and distribution-level video scores are not reported by this run.

Table 14: All four source-level connection results. PSNR excludes the first frame; direction errors are degrees. Source ordering matches Figure 9. Copy-first uses no generated motion.

Source	State	PSNR	Copy-first PSNR	To GT	To supplied
1	gt	18.86	17.30	12.25	12.25
2	gt	15.73	11.71	4.88	4.88
3	gt	19.99	14.29	6.82	6.82
4	gt	14.75	10.43	3.48	3.48
1	ours	21.41	17.30	8.97	12.26
2	ours	15.05	11.71	5.38	7.78
3	ours	20.96	14.29	6.85	7.05
4	ours	15.06	10.43	4.44	5.97
1	vggt	19.41	17.30	13.61	13.90
2	vggt	15.21	11.71	5.72	5.11
3	vggt	20.15	14.29	7.57	6.70
4	vggt	16.11	10.43	5.02	5.70
1	wrong	14.99	17.30	82.17	12.24
2	wrong	13.05	11.71	29.67	18.71
3	wrong	13.86	14.29	20.78	8.93
4	wrong	14.08	10.43	15.52	4.80

On all four sources, estimated-state generation improves PSNR and target direction error over mismatched-state generation. Wrong-state motion is closer to its supplied trajectory than to the true target on all four. The estimated-state mean target-direction advantage over GT is driven by one source; GT is better on the other three. No ranking of state providers is claimed from this small set. Nor does agreement in direction establish correct translation magnitude, object dynamics, or physical interaction.

J Cross-representation completion and constraint drift

Each of eight camera/weather groups contributes four annotation-conditioned RGB generations, each reused for three destination modalities: 32 intermediate generations and 96 conversions. All groups come from the same Scene01 time window. Two-hop reuses generated RGB latents directly for perception; direct uses annotation latents with the RGB perception instruction. GT RGB supplies a separate reference. Visual inspection focuses on scene completion, correspondence between generated RGB and the destination representation, and preservation of specified foreground objects and semantic regions.

The visual audit covers frame 40 and 80 of every conversion, with additional frame-zero inspection of the first 24 cases; all three-frame panels are available. It is a single non-blinded model-assisted review, not a human preference study or 96 independent win labels. Its clearest observations are more explicit background structure for sparse-to-dense completion, alongside missing or merged small instances and occasional semantic-region drift. Self-consistency with generated RGB is useful but does not independently validate geometry. We report qualitative observations rather than a subjective overall win rate or an accuracy ranking.

K Earlier-model depth representation and context diagnostics

Translation direction error falls from 21.53° at one-frame separation to 2.57° at 32 frames; VGGT gives 9.07° and 1.81°. Adjacent rotation error is 0.141° versus an identity baseline's 0.129°, whereas at 32 frames it is 0.485° versus 3.558°. The model captures longer-scale changes more reliably than small increments.

VKITTI uses its actual log-one-plus DepthCodec, not the different Blender geometry representation. No codec saturation is measured on these clips. Thus the resize-plus-codec score cannot be attributed

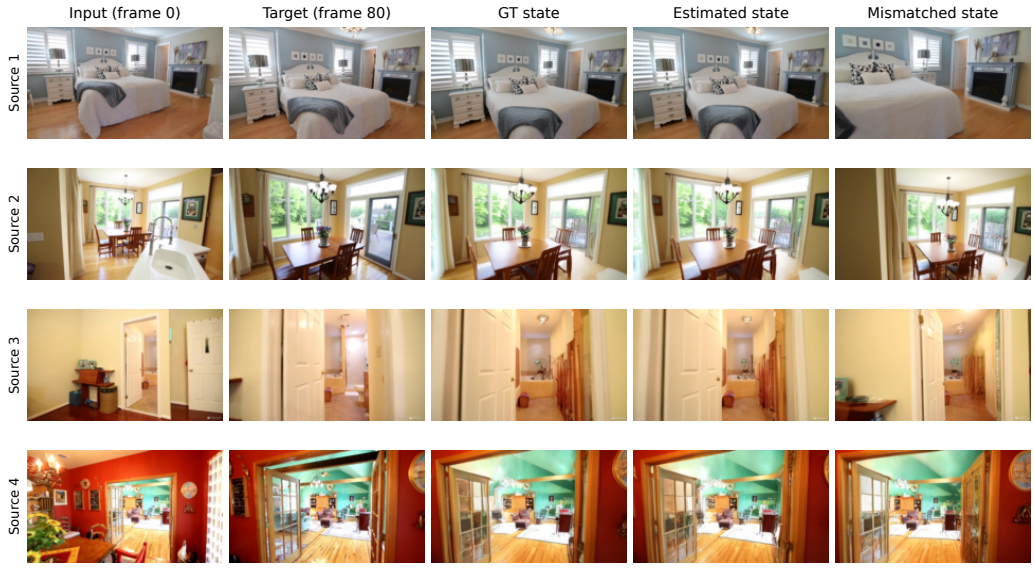


Figure 9: All four connection sources at frame 80, with their initial frame and true target. Only the trajectory source changes between generated columns; there is no outcome-based example selection. The figure is a fixed temporal cross-section, not evidence of flicker-free video.

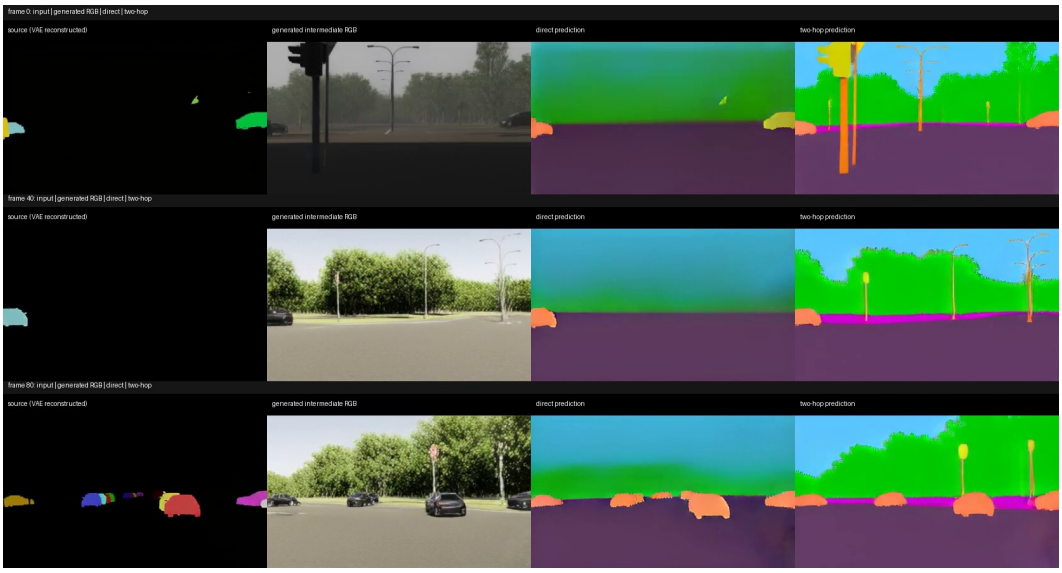


Figure 10: Instance-to-semantic completion (audit case 05). Columns show input, generated RGB, direct output, and two-hop output; rows show frames 0, 40, 80. This is a deliberately selected completion example, not a median-quality sample. Unspecified background becomes more explicit, without establishing that it reproduces the unique original scene.

Table 15: Depth diagnostic AbsRel on eight full clips. All/boundary/interior use per-frame median alignment; clip-scale uses one fitted scale per clip. The first two paths include downsampling and restoration to GT coordinates, not just an algebraic codec roundtrip.

Path	All	Boundary	Interior	Clip-scale
Resize + codec	0.043	0.079	0.035	0.041
Resize + codec + VAE	0.058	0.100	0.048	0.056
Model prediction	0.134	0.219	0.116	0.137

to clipping or to the scalar transform alone. A reconstruction reference is not a strict accuracy bound, and subtracting its error from model error is not a causal decomposition. Model AbsRel is 0.110, 0.112, and 0.197 over 0–10, 10–30, and 30–80m, respectively; every band uses the same full-image fitted scale.

Context variants keep frame 40 fixed. Reverse reorders all 81 RGB frames; shuffle permutes the other 80 with a fixed sample-specific seed; repeat uses only the center RGB frame. Every changed input is freshly VAE-encoded. Original-context predictions reuse the validated cache, whereas transformed inputs use the diagnostic loader, so small differences should not be overinterpreted without a fresh-original encoding control. Center AbsRel is 0.1134/0.1115/0.1123/0.1779 for original/repeat/reverse/shuffle. The shuffled input degrades all eight sources, but neither temporal integration benefits nor transformer-specific causality follow from this result alone.

L Data composition and additional checkpoint comparisons

Table 16: Adding Blender at matched original-data exposure. A is RE10K plus VKITTI; B is Blender. A+B has twice the total updates. Dense scores shown here are ID; OOD depth worsens from 0.327 to 0.421.

Validation metric	A only	A+B
Normalized state MSE ↓	0.201	0.040
Translation RMSE ↓	4.438	3.809
Depth AbsRel ↓	0.302	0.285
Semantic mIoU ↑	0.483	0.498
Instance foreground IoU ↑	0.831	0.818
Detection box F1@.5 ↑	0.396	0.368

Table 17: Cross-task results of the earlier configurations. Every row uses one checkpoint across its tasks; different budgets are not ranked as a controlled architecture comparison. B-only evaluation on real-domain tasks is a transfer probe.

		Dense perception: VKITTI ID					Sparse state: RE10K held-out		
Training configuration	Updates	AbsRel ↓	δ_1 ↑	mIoU ↑	fgIoU ↑	Box F1 ↑	Trans. RMSE ↓	Rot. (°) ↓	Dir. cos. ↑
A only	1k	0.302	0.586	0.483	0.831	0.396	4.438	2.752	0.971
B only (transfer probe)	1k	1.072	0.135	0.015	0.040	0.000	7.803	4.563	0.838
A+B	2k	0.285	0.589	0.498	0.818	0.368	3.809	1.842	0.986
Joint + external depth	5k	0.204	0.788	0.531	0.792	0.408	3.775	2.208	0.985

The original objective reduces fitted error more than increment reweighting or endpoint-only supervision. This shows learnability on fitted clips, not held-out improvement. It does not establish that query packing has no cost.

M Additional perception evaluations and historical diagnostics

Released GenCeption and WorldCeption are evaluated on the same five Sintel sources (Butler et al., 2012), with 81 frames, 480×832 resolution, native 24 FPS, and unique-frame ownership masks. GenCeption uses its released $t = 49$ setting. Training data and compute are not matched. Clean and Final denote rendering passes.

Table 18: Representation fidelity of the extended 5k checkpoint, distinct from the principal 20k run. Reconstruction references expose codec loss; foreground overlap does not measure instance identity.

Model / reference	Split	AbsRel ↓	δ_1 ↑	mIoU ↑	fgIoU ↑	F1 ↑
Joint model	ID	0.204	0.788	0.531	0.792	0.408
VAE reconstruction reference	ID	0.076	0.959	0.744	0.968	0.469
Joint model	OOD	0.240	0.542	0.376	0.902	0.283
VAE reconstruction reference	OOD	0.051	0.978	0.609	0.977	0.402

Table 19: Training-clip fitting without changing the sparse representation. Values are after/before translation RMSE ratios. The two subsets are not independent full-training seeds.

Objective	Subset 1	Subset 2
Original objective	0.082	0.288
Absolute only	0.197	0.370
Reweighted increments	1.898	3.004
Slot endpoints only	0.971	1.038

For the earlier three-source model, on 1,000 KITTI images repeated to form 81-frame inputs and evaluated at the center with per-image median alignment, acquisition-date-macro AbsRel is 0.139, RMSE 5.293, and $\delta_1 = 0.810$. The five groups are acquisition dates, not independent scenes. Pooled-pixel AbsRel is 0.146. This is local relative-depth transfer, not a verified published KITTI comparison.

These historical 32-source tests and the historical four-source reconstruction experiment answer different questions; only the latter measures generated motion with an external estimator. They should not be merged into a 32-source motion accuracy claim. The final model's separate 32-source motion experiment is reported in Appendix E. The research comparison workbook retains other published reference values and planned evaluations outside this compiled manuscript; unmeasured cells are not presented as experimental results.

N Additional evaluation and ablation details

Trajectory visualization provenance. Figure 4 presents selected qualitative successes, not a representative sample for estimating aggregate accuracy. The four examples use final-model outputs: two RE10K validation clips and two Blender training clips. The three RGB frames are sampled at indices 0, 40, and 80 from the original videos, rather than VAE reconstructions. Trajectories show absolute clip-relative translations, with a shared origin prepended and coordinates displayed as $(x, z, -y)$. Coordinates are clip-scaled, not metric; independent axis ranges mean that projected aspect ratios should not be interpreted as physical distance ratios. No smoothing, fitted alignment, or cumulative integration is used.

Earlier reference-model state-mediated reconstruction. The four-source diagnostic fixes the generator, first video frame, task instruction, and sampling noise. Conditions are ground-truth trajectories, our estimates, VGGT estimates, and trajectories from a different source. Estimation observes the complete clip; the generator receives future information only through the explicit trajectory, not future RGB or estimator hidden states. Frozen VGGT reconstructs generated motion from 17 equally spaced frames. Direction error is measured separately against the real target and the supplied condition; full-video PSNR excludes the first frame. This isolates interface reuse rather than prediction of an unknown trajectory from past observations.

Composition units and interpretation. Figure 7 selects eight distinct conversion directions from the completed three-source-reference evaluation: instance-to-semantic, instance-to-depth, boxes-to-semantic, boxes-to-depth, semantic-to-instance, semantic-to-boxes, depth-to-semantic, and depth-to-instance. All panels use the fixed middle frame (40), with no enhancement of predictions. Selection prioritizes visible structure and task coverage, not a representative success rate. The instance-to-depth

Table 20: Final-model task-aware dense perception: 48 ID and 16 scene-held-out windows per task. This local task protocol is distinct from the crop-v2 median-aligned comparison in Table 7; its depth values must not be mixed into that ranking. Foreground IoU and box F1 are rendering-derived diagnostics.

Task	Metric	ID	Scene-held-out
Depth	AbsRel ↓	0.210	0.278
Semantic	mIoU ↑	0.542	0.426
Instance rendering	FG IoU ↑	0.805	0.806
Detection rendering	Box F1@.5 ↑	0.456	0.277

Table 21: Final-model matched Sintel normal evaluation. Angles are degrees and threshold accuracies percentages; these are not the published benchmark protocol. Ours has lower Clean mean error but worse median and threshold scores, and does not outperform GenCeption on Final.

Model	Pass	Mean ↓	Median ↓	$< 11.25^\circ$ ↑	$< 22.5^\circ$ ↑	$< 30^\circ$ ↑
GenCeption 1.3B	Clean	32.457	24.135	19.66	47.79	61.73
WorldCeption	Clean	30.544	25.795	15.09	44.77	60.27
GenCeption 1.3B	Final	30.260	25.605	16.84	43.47	60.22
WorldCeption	Final	33.669	29.445	11.28	37.94	54.12

example uses morning appearance and the semantic-to-instance example uses fog; other panels use the clear appearance of the same camera/window. Some share an RGB intermediate. Direct outputs are untrained-input references, not trained translation baselines. The categorical continuation subsequently completed a two-group composition evaluation, but these eight rendered panels retain the earlier model's provenance. Completion of a new metric run does not change their source.

The 96 conversions comprise eight camera/weather combinations of one Scene01 window and 12 directed representation pairs. Three destination modalities share each generated RGB intermediate, so these conversions are not independent generation trials. Direct annotation input to an RGB-trained perception path is an untrained-input reference, not a trained competing translator. Completion of unspecified content and preservation of supplied constraints are distinct criteria; agreement with the generated RGB does not independently establish fidelity to the original geometry.

Routing configurations and diagnostics. Shared-only, additive shared-plus-expert, expert-only, and role-dependent configurations use the same RE10K/VKITTI schedule, seed, global batch, and 1k updates. Expert-only and role-dependent configurations use nonzero expert output initialization; the additive baseline uses zero output initialization. Thus the study compares implemented configurations, not routing independently of initialization. Normalized state MSE averages squared errors over standardized poses and their increments. Translation RMSE uses decoded, clip-scaled translations, not meters. Definitions of the remaining diagnostics are given in Appendix P.

O Earlier same-checkpoint integration and intervention diagnostics

Model identity and consolidation. This historical integration pilot is separate from the final-model study in Appendix E. The evaluated consolidated model starts from the six-source configuration after 13,500 updates and receives 500 additional joint updates. It is not the subsequently trained 15,000-update base model. All consolidated-side classification, control, retention, composition, and camera-reuse stages strictly load identical weights, with no query or codebook reset. Training allocates 15% of samples to weather readout, 25% to same-geometry weather pairs, and 60% to replay of the existing task mixture. Replay includes camera estimation and control, dense perception and synthesis, and native T2V. The backbone learning rate is 5×10^{-6}, interface and sparse expert rates are 10^{-5}, and the designated sparse-query group uses 10^{-4}. The optimizer is restarted with 50 warmup updates. No model merging or task-specific checkpoint switching is used for these tests.

Table 22: Expanded control evaluation on 32 held-out RE10K sources for the earlier three-source model. Values are latent errors to the original paired target; wrong-condition outputs have no paired counterfactual GT.

History	Correct	Null	Wrong	Stationary
Static	0.213	0.442	0.619	0.476
Dynamic	0.204	0.316	0.621	0.442

Categorical readout and control. The 96 balanced clips comprise 48 training-domain and 48 Scene20 clips. Balanced accuracy is 100% and 79.17%, respectively; combined accuracy is 89.58% and macro-F1 is 0.8903. Zero RGB gives 16.67% balanced accuracy. The main out-of-scene confusion is clear weather with morning or sunset; clear-class recall is zero on Scene20. Twenty-eight control runs cover four geometries, each with six supplied classes and a null condition. Predicted categories are also reused through the same generation interface. Automated completion and self-consistency do not replace independent assessment of weather fidelity; the planned human review remains pending.

Camera-state reuse protocol. Four RE10K sources each use four conditions: GT, model-estimated, mismatched, and translation-reversed. All 16 runs complete, using one paired seed per source and identical recorded initial-noise hashes across conditions. Generation uses the first RGB frame repeated as its only visual history, 40 denoising steps, guidance three, and shift three. The estimator observes the complete clip, but its only communication to generation is the explicit state. No future RGB, future RGB latent, or estimator hidden state is supplied. Thus the experiment is state-mediated reconstruction plus controlled intervention, not prediction of an unobserved trajectory.

Measurement and aggregation. Frozen VGGT reconstructs poses at frames $0, 5, \ldots, 80$. Adjacent direction error compares successive sampled poses; first-to-all error compares every sampled pose to the initial pose. Per-source means are averaged equally. Translation coverage is 100% for all conditions. GT/estimated-state PSNR excludes the first frame; mismatched and reversed conditions have no paired counterfactual RGB target and are not assigned reconstruction PSNR. Real-video calibration gives mean adjacent rotation and translation errors of 0.089° and 5.923°. One of the four sources has 19.103° calibration translation error and is retained, not excluded.

Table 23: Generated motion relative to supplied states. Adjacent and first-to-all errors are source-macro means in degrees.

	Adjacent		First-to-all	
State	Rotation	Translation	Rotation	Translation
GT	0.573	14.810	1.885	6.774
Estimated	0.477	14.580	1.278	6.189
Mismatched	0.479	17.088	1.964	9.439
Translation reversed	0.417	21.100	1.455	12.650

Estimated states yield PSNR 16.074 dB versus 15.990 dB for GT conditions; the small difference does not establish superiority. For mismatched states, three of four sources are closer to supplied motion than original GT. Reversed translation provides a more discriminative intervention: every source moves toward the reversed direction, with source-average error to original GT of 158.900° (adjacent) and 167.350° (first-to-all). These are directional control results, not evidence of physical causal identification or accurate metric displacement.

Capability retention. Matched pre/post-consolidation checks use four samples per perception task, two per generation task, and six native T2V prompts; Blender reconstruction is explicitly evaluated on training clips. No uniform preservation claim is made: RE10K clip-scaled translation RMSE changes from 2.605 to 3.127 and TartanAir from 13.671 to 14.659, while Blender changes from 1.036 to 0.805. VKITTI training-domain depth AbsRel changes from 0.241 to 0.220, and out-of-scene depth from 0.211 to 0.228. Semantic mIoU is approximately stable (0.513 to 0.519 in-domain; 0.423

to 0.424 out-of-scene). These small tests establish coexistence of interfaces, not absence of forgetting. Native T2V is assessed qualitatively rather than by paired pixel error against a uniquely correct video.

Composition and historical illustrations. The consolidated weights also complete 24 directed conversions across two VKITTI camera/weather groups. These are not 24 independent scenes and self-consistency is not an independent geometry test. The eight rendered examples in Figure 7 retain their earlier three-source-reference provenance; metric-only completion of the new run does not turn those figures into consolidated-checkpoint outputs.

P METRIC DEFINITIONS AND AGGREGATION

Splits are not metrics. "Held-out" denotes exclusion from the relevant training split, not a scoring rule. Source-disjoint RE10K evaluation excludes original source videos, not merely different windows from the same videos. VKITTI ID denotes the training-scene domain and does not imply source-disjoint generalization; scene-held-out denotes an excluded base scene. OOD in the direct-conversion column instead means an annotation supplied to an RGB-trained input path. These uses are not one common generalization benchmark. Source-macro aggregation averages clips within each source, then weights sources equally. Routing and data-composition tables use this aggregation; other tests follow their separately specified image, pair, and task aggregation protocols.

Categorical state metrics. Weather prediction selects the largest codebook probability. For six classes, balanced accuracy is $\frac{1}{6}\sum_k TP_k/(TP_k+FN_k)$ and macro F1 is $\frac{1}{6}\sum_k 2TP_k/(2TP_k+FP_k+FN_k)$, assigning zero when a denominator vanishes. These standard classification metrics evaluate observed weather and illumination. They are distinct from continuous camera-state regression and from the visual controllability of generated video.

Normalized state MSE (internal diagnostic). Let $s_t \in \mathbb{R}^6$ concatenate boundary-relative translation and rotation vector, and $a_t = (s_t - \mu)/\sigma$ use coordinate-wise training statistics. Set $a_0 = (0-\mu)/\sigma$ at the boundary and $p_t = [a_t, a_t - a_{t-1}] \in \mathbb{R}^{12}$ for T subsequent poses. Four successive vectors are grouped into each sparse token. The score is

$$E_{\text{state}} = \frac{1}{12T}\sum_{t=1}^{T} \|\hat{p}_t - p_t\|_2^2. \tag{11}$$

This dimensionless diagnostic equally weights standardized absolute and increment coordinates. It is not physical trajectory distance or a standard pose benchmark. Lower is better; comparisons require identical representation and normalization statistics.

Translation and rotation readout diagnostics. Unpacking retains the absolute coordinates and reverses standardization. Translation RMSE is $[\sum_{t=1}^{T} \|\hat{u}_t - u_t\|_2^2/(3T)]^{1/2}$, not translation MSE. Its units are dataset-specific clip-normalized translation units, not meters; RE10K preprocessing divides translations by the median nonzero consecutive camera-center displacement in the clip (falling back to one for static clips). no fitted similarity transform is applied in the routing/data tables. Rotation error is the mean geodesic angle $\arccos((\text{tr}(\hat{R}_t^\top R_t)-1)/2)$ in degrees, clipping the cosine argument to $[-1,1]$. Direction cosine averages $\hat{u}_t^\top u_t/(\|\hat{u}_t\|\|u_t\|)$ over vectors whose two norms exceed 10^{-8}. It measures boundary-relative translations, not consecutive increments. Higher cosine and lower errors are better. The fitting ratio divides subset-mean post-fit translation RMSE by pre-fit RMSE; a value below one indicates improvement on fitted examples only.

Latent condition diagnostics. For sampled latent sequence $\hat{z}_c$ under condition c and original paired target z^*, define $E_c = \|\hat{z}_c - z^*\|_2^2/N$, where N counts scalar latent entries. The control table reports E_c: lower paired-target error is better. Correct uses the paired trajectory, wrong another clip's trajectory, null removes the condition, and stationary supplies zero physical motion. Noise and other inputs are paired across conditions. Target gap is $E_{\text{wrong}} - E_{\text{correct}}$; output change is $\|\hat{z}_{\text{wrong}} - \hat{z}_{\text{correct}}\|_2^2/N$. All are in squared VAE-latent units. A positive gap favors the correct condition for the original target. Output change measures sensitivity and has no intrinsically desirable direction. Neither score establishes that a wrong-condition output is a physically correct alternative future.

Rendered-label diagnostics. Foreground IoU divides binary-mask intersection by union, pooling pixel counts across each clip. Prediction foreground has maximum RGB channel in $[0,1]$ at least

0.12; GT foreground is any nonblack pixel. It ignores instance identity. Rendered-box F1 extracts prediction connected components of at least 24 pixels and their enclosing boxes; GT boxes are recovered from rendered colors with the same minimum support. Greedy one-to-one IoU matching counts matches at IoU ≥ 0.5. Box precision P and recall R give $F1 = 2PR/(P + R)$; this is not detection AP. Boundary F1 in the conversion table is different: both prediction and GT masks are thresholded at 0.12 to exclude translucent box interiors. Mask edges use a one-pixel erosion difference and are matched within a two-pixel square tolerance, then edge precision and recall are combined. It is neither box F1 nor object recognition accuracy.

Standard metrics under specified adapters. Depth AbsRel is $\mathrm{mean}(|\hat{d} - d|/d)$ on valid pixels; RMSE is $\sqrt{\mathrm{mean}[(\hat{d} - d)^2]}$; δ_1 counts the fraction with $\max(\hat{d}/d, d/\hat{d}) < 1.25$. AbsRel and δ_1 are dimensionless; RMSE uses GT depth units after the stated scale alignment. Semantic mIoU averages class IoU after palette decoding, excluding classes with zero union. Normal errors are angles between unit vectors; threshold scores count valid pixels below each stated angle. PSNR is $10 \log_{10}(L^2/\mathrm{MSE})$ dB for image range L. Representation PSNR evaluates codec images, not raw geometry. Identical metric names do not make different crop, scale, validity, and aggregation protocols interchangeable.

Relative-pose and representation diagnostics. The common pose test measures relative-rotation geodesic error and relative translation direction error in degrees. Its local joint score averages $\max(0, 1 - \max(e_R, e_t)/30)$, the normalized area under the empirical joint success curve up to 30°, not a reproduced published VGGT benchmark. Generated-video "to GT" scores reference original motion; "to supplied" scores reference conditioning motion. Both use externally reconstructed motion. Codec and VAE reconstruction references measure representation distortion, not guaranteed achievable accuracy ceilings. Coverage counts numerically valid predictions, not accurate ones.

Depth-boundary pixels originate from valid horizontal/vertical neighbors whose log-depth difference exceeds 0.1, dilated with a 7×7 window. Interior pixels are the remaining valid pixels. Boundary and distance-band scores reuse the full-frame scale rather than fitting each subset separately. Clip-scale depth uses the median of valid per-frame median-ratio scales, then applies this single scale to every frame of that clip.

Attention diagnostics. Attention mass sums softmax probabilities over keys in the named group and averages over probed queries, heads, and layers. Condition relative L2 is $\|H_C^{\mathrm{perturbed}} - H_C\|_2 / \max(\|H_C\|_2, 10^{-12})$, with clean input fixed while paired targets change. These dimensionless scores describe internal sensitivity, not task accuracy or causal importance.

Q LIMITATIONS AND SCOPE

Routing and data comparisons use one training seed; some routing configurations differ in expert initialization, and the data mixture changes total compute. The 32-source reuse test measures state-mediated reconstruction, not unseen-domain transfer or open-loop future-state prediction. The single-scene composition audit is diagnostic rather than a broad generalization benchmark. The visual audit examines selected temporal cross-sections, is model-assisted and non-blinded, and is not a human preference study.

VGGT-based motion measurement is imperfect and shares a model family with one state provider. No independent ground-truth motion of generated scenes exists. Depth and camera specialists remain stronger on the principal local accuracy tests. Foreground overlap and rendered-box scores do not establish identity tracking or standard instance/detection AP.

The two-hop study evaluates visual completion rather than paired-label accuracy. The archived paired-label tests did not establish a consistent accuracy advantage; we do not claim one. Visual structure can become more explicit while small objects disappear or supplied semantic regions change. The failure example in Appendix J illustrates this distinction. Qualitative inspection does not establish a quantitative accuracy advantage.

Compactness refers to the architectural addition, not pretraining cost or frozen-backbone tuning. Capacity, data coverage, and optimization may all limit performance; their contributions have not been isolated by a scaling study. Larger backbones and broader data remain future tests. Consolidated weights support categorical readout and continuous camera-state reuse, but retention is assessed on

small subsets and is not uniform across metrics. External comparisons use final weights; historical ablations and qualitative galleries retain their original evaluation provenance. Human blind ratings are pending, and confidence intervals containing zero do not prove equivalence. The presented state-mediated reconstruction does not establish closed-loop interaction, planning, contact dynamics, or persistent physical simulation.

R ADDITIONAL VISUAL EVIDENCE AND SELECTION PROTOCOL

The galleries use recorded evaluation outputs, without new inference, image retouching, trajectory smoothing, or fitted pose alignment. The dense gallery selects the lowest-error saved depth/normal examples and highest-scoring saved semantic/instance/box examples for VKITTI and Blender. The reverse gallery selects the lowest paired-target latent error among saved videos of each task. These are explicitly successful-example selections, not estimates of average quality. Figure 3 uses final-model outputs on VKITTI2 and Blender, two examples per task. Standard video panels show frame 40; input and target columns are VAE reconstructions. The separate archived Sintel gallery uses fixed source scenes and raw target renderings.

The trajectory gallery contains two selected examples each from RE10K and Blender, with three original RGB frames and a 3D trajectory. Its split provenance and independently scaled coordinate axes are specified in Appendix N; these are final-model outputs. Figure 5 uses the earlier consolidated model described in Appendix O, showing six categories for each of two geometries. The weather recognition strip uses all six appearances of its first archived ID geometry at frame eight; prediction labels are copied from recorded query scores, with no re-inference. The recognition strip uses the historical categorical reference, whereas the consolidated model's joint readout/control scores are given separately in Appendix O. Neither gallery is newly generated from the final weights used in Table 1. Each configuration shares weights across its own perception and generation directions. Blender visualizations are labeled as training examples throughout.

Figure 6 selects three visibly coherent examples from 16 final-model validation generations under one frozen seed. Each panel uses saved frames 0 and 80 without retouching. For GenCeption, we reuse the exact cached text conditions and match resolution, 81-frame length, 40 steps, shift 3, and guidance 3. WorldCeption uses UniPC. GenCeption's public pipeline exposes no T2V sampler, so we apply an experimental rectified-flow Euler recovery to the released weights. The PyTorch and JAX implementations use framework-specific noise streams; this is a capability-recovery comparison rather than a distributional T2V score.

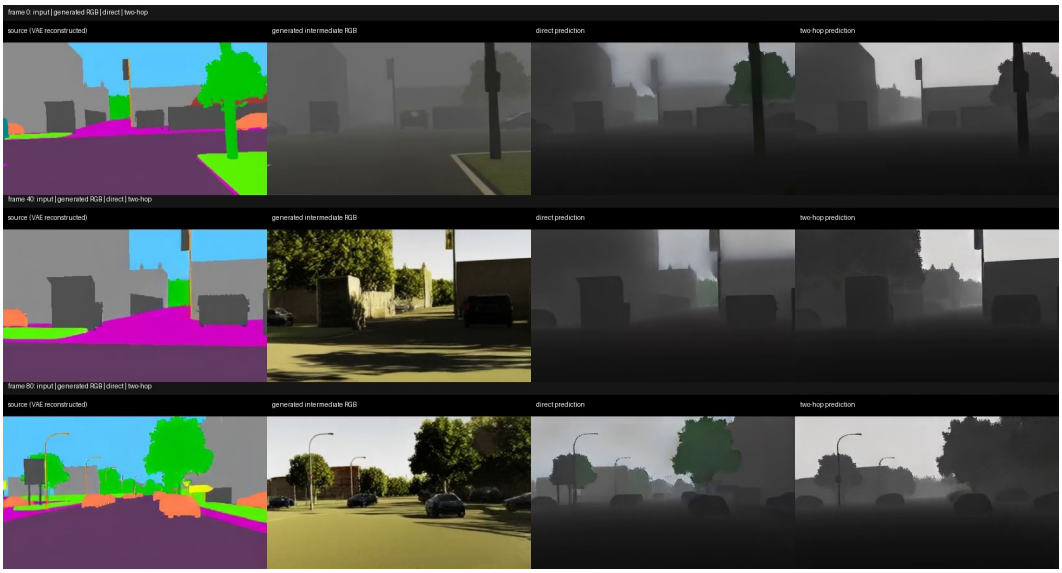


Figure 11: Semantic-to-depth constraint drift (audit case 33), with the same layout. At the middle frame, the generated intermediate changes a supplied building region into vegetation; the final depth inherits that change. A sharper or more depth-like output is not necessarily a more faithful one.

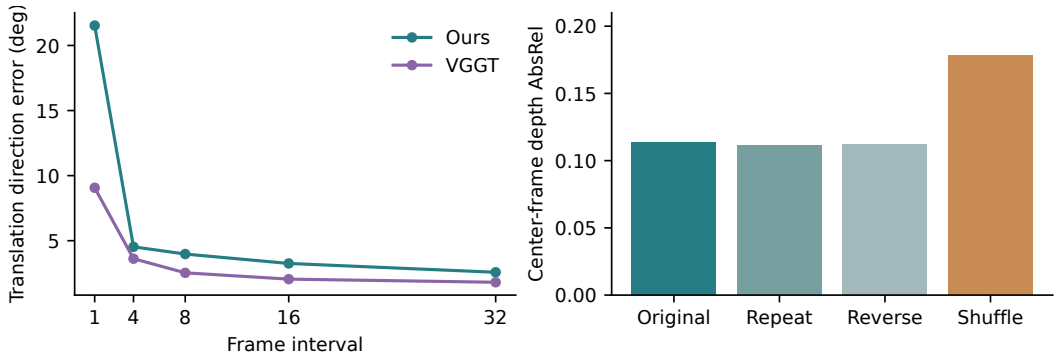


Figure 12: **Motion scale and temporal context.** Left: translation direction error across frame intervals. Right: center-frame depth under original, repeated, reversed, and shuffled context. Eight clips per test; correlated frames are not treated as independent samples.



Figure 13: **State reuse and translation intervention with identical weights.** All four frozen sources, without outcome-based selection. The initial observation is followed by generated frame 80 under GT, estimated, and reversed-translation conditions. Estimated-state and GT-conditioned outputs are similar, while reversed motion exposes different views. The figure includes distortions in newly exposed regions; directional response does not imply uniformly accurate scene reconstruction.

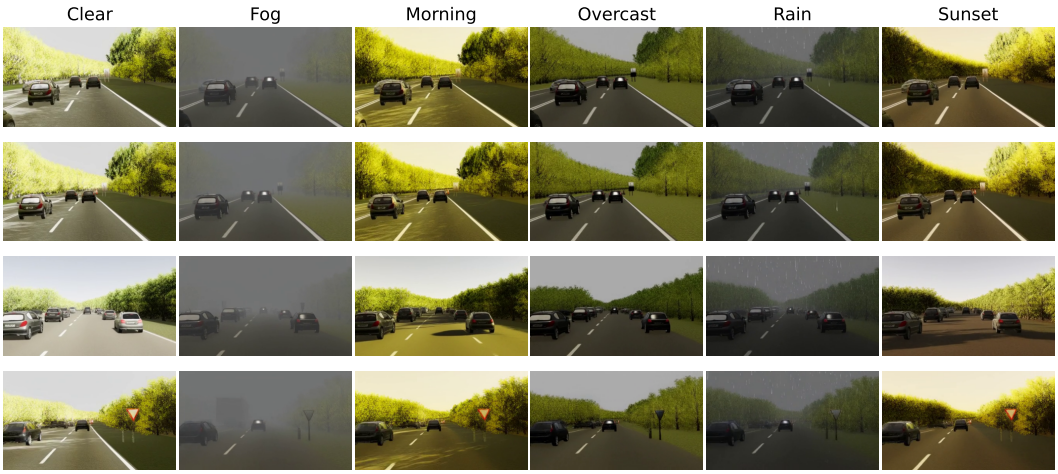


Figure 14: **Consolidated-model weather control.** All four frozen geometries at frame 40 under six supplied categories. Each row keeps its dense condition fixed. The first two rows are from Scene18 and the last two from Scene20. These cross-sections do not establish temporal fidelity or a human-rated success rate.



Figure 15: **Additional cross-dataset depth prediction on Sintel.** The same two scenes as the normal gallery. Target depth uses the evaluator's fixed relative-depth codec; predictions are displayed as saved, without per-image contrast enhancement or display-specific scale fitting.