SEEING AND IMAGINING AS ONE: UNIFIED VIDEO GENERATION AND 4D PERCEPTION FOR EMBODIED WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied world models should not only imagine plausible futures, but also perceive the geometric, semantic, and motion structure of the evolving world. We present GenSight, a unified embodied world model that brings future video generation and dense perception into a single pretrained video diffusion transformer. GenSight uses a unified spatiotemporal representation for video, depth, embodiment segmentation, and first-frame-relative 2D flow. Built on a shared DiT backbone, GenSight uses lightweight task modulation to predict all modalities within the same architecture. All tasks are jointly optimized within a unified velocity prediction space. At inference, GenSight supports two complementary modes: multi-step future video generation with aligned dense predictions and single-step dense perception from an observed video. By unifying modality representation, model architecture, and prediction space, GenSight avoids separate modality-specific pipelines and couples appearance, geometry, semantics, and motion in a shared representation. It promotes cross-modal consistency and allows dense perception to benefit from the spatiotemporal priors learned through video generation. Trained on embodied manipulation data spanning multiple robot embodiments and human-hand interactions, GenSight demonstrates strong performance in both video generation and dense 4D perception, highlighting the effectiveness of unifying seeing and imagining for embodied world modeling.

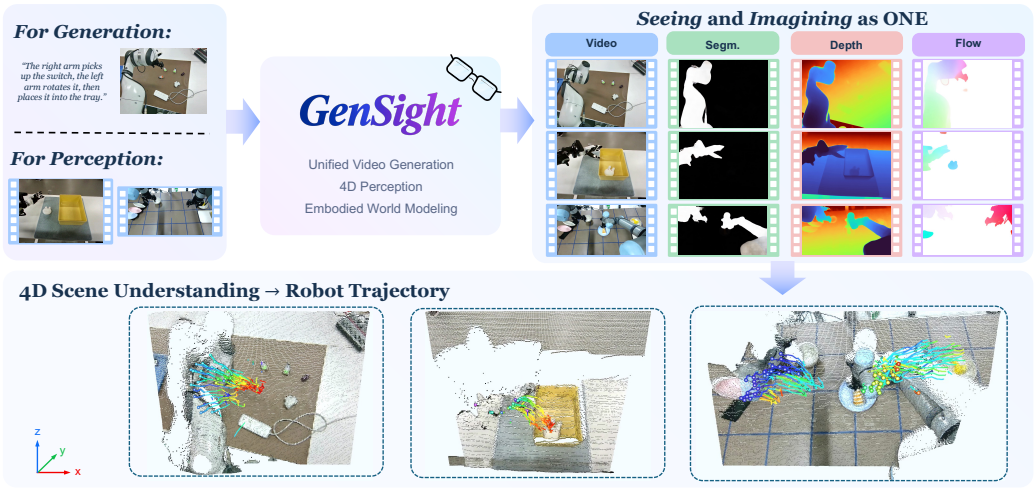


Figure 1: **GenSight unifies seeing and imagining for embodied world modeling.** Given an initial image and a text prompt, GenSight generates future video together with aligned embodiment segmentation, depth, and first-frame-relative 2D flow. Given an observed video, the same model predicts each dense modality in a single step. Together, these outputs characterize appearance, semantics, geometry, and motion. The bottom panel illustrates 4D scene understanding through scene geometry and robot motion trajectories over time.

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.

1 INTRODUCTION

Large-scale video generators have learned spatiotemporal priors (King et al., 2026; Wang et al., 2026b) and can synthesize realistic future video from images and language (Wan Team, 2025; NVIDIA, 2026). However, for embodied interaction, RGB video is only one view of the predicted world and does not directly provide structured state estimates needed to understand the environment, analyze agent-environment dynamics, or support reasoning (Guo et al., 2026a; Zhao et al., 2026). Combining these capabilities allows an embodied world model to *see* the structured state and *imagine* how it may evolve, providing a unified basis for understanding embodied dynamics.

Despite recent efforts to connect generation and perception, several limitations remain in embodied world modeling. First, future generation and dense perception are commonly developed as separate systems: world models generate RGB futures, while independent estimators recover their physical and semantic structure (Chen et al., 2025a; Yang et al., 2024a; Ravi et al., 2025; Teed & Deng, 2020). This generate-then-estimate pipeline can propagate visual artifacts into perceptual outputs and offers no mechanism for cross-modal constraints during training. Second, attempts to incorporate structured signals commonly rely on modality-specific branches or heads (Guo et al., 2026a; Yuan et al., 2026), or parallel-stream architectures (Zhao et al., 2026; Chen et al., 2026), increasing architectural fragmentation. Moreover, these cues are used mainly as auxiliary supervision, rather than enabling the same model to densely perceive observed interactions. Third, although generative models have been adapted for dense perception (Wang et al., 2026b; Guo et al., 2026b; Zhang et al., 2026), jointly supporting future video generation and dense perception through a fully shared model with a common prediction formulation remains underexplored. This raises a central question: can a single embodied world model both imagine futures and perceive interactions?

To this end, we introduce GenSight, a unified embodied world model built on a pretrained Video DiT. GenSight unifies video, depth, embodiment segmentation, and first-frame-relative 2D flow within a common spatiotemporal latent representation, all processed by the same frozen video autoencoder. A shared backbone predicts each modality through lightweight task modulation, while the Transformer blocks and decoder remain fully shared, without modality-specific output heads. Beyond unifying representations and architecture, all tasks are jointly learned in a unified velocity prediction space, aligning dense perception with the video generator's native prediction formulation. The model supports two complementary inference regimes: multi-step future video generation with aligned dense readouts, and single-step dense perception from an observed video.

This three-level unification yields three important advantages. First, modeling all modalities within a shared representation promotes temporal and cross-modal consistency, reducing the mismatch inherent in post-hoc perception pipelines. Second, joint learning enables reciprocal transfer across tasks: the rich spatiotemporal priors of video generation support dense perception, while supervision from geometry, semantics, and motion encourages the shared representation to capture physically meaningful structures. Third, unifying the model and prediction space avoids the need for modality-specific pipelines while preserving operational flexibility: the same model can imagine future appearances or perceive an observed interaction under different inference regimes. Together, these properties transform a pretrained video generator into a unified embodied world model that jointly models multiple facets of embodied dynamics.

Our contributions are summarized as follows:

- We study the task-level unification of future video generation and dense perception for embodied world modeling, enabling a single model to both imagine future observations and infer structured states from observed videos.
- We introduce GenSight, which achieves three-level unification across heterogeneous modalities through a common spatiotemporal latent representation, a fully shared pretrained Video DiT without modality-specific output heads, and a unified velocity prediction space.
- We validate GenSight across diverse robot embodiments, human-hand interactions, and unseen tasks, demonstrating strong performance in both generation and perception and providing a systematic empirical analysis of our design choices.

2 RELATED WORK

Video generation and embodied world models. Video diffusion models, including Stable Video Diffusion (Blattmann et al., 2023) and Wan 2.2 (Wan Team, 2025), provide spatiotemporal priors

for conditional generation. Cosmos3-Nano extends video generation within an omnimodal world model (NVIDIA, 2026). Recent work adapts video generation to embodied world modeling by predicting future observations conditioned on embodied interactions, thereby providing a generative model of environment dynamics for robot planning and control (Du et al., 2023; Yang et al., 2024b). RynnWorld-4D (Zhao et al., 2026) jointly generates RGB, depth, and optical flow using modality-specific Transformer branches. In contrast, GenSight shares the Transformer backbone across modalities and supports dense perception of observed videos.

World action models. World action models couple future prediction with robot control. X-WAM and FlowWAM incorporate depth and optical flow, respectively (Guo et al., 2026a; Chen et al., 2026), while Flex-π jointly models RGB, geometry, semantic features, and actions (Yan et al., 2026). EgoWAM studies alternative visual prediction targets for human-to-robot transfer (Li et al., 2026a), and WAM4D and DreamWAM use auxiliary visual supervision with the corresponding branches removed at deployment (Li et al., 2026b; Yuan et al., 2026). Rather than directly predicting robot actions, GenSight unifies future video generation and dense perception of observed interactions within a single model.

Generative priors for dense perception. Marigold and Lotus demonstrate the value of generative priors for image perception (Ke et al., 2024; He et al., 2025), while Geo4D, DVD, and GenCeption extend this direction to video perception (Jiang et al., 2025; Zhang et al., 2026; Wang et al., 2026b). Diff-2-in-1 and UniGP further connect generation and perception at the image level (Zheng et al., 2025; Guo et al., 2026b). Building on these advances, GenSight jointly learns future video generation and dense perception for embodied interactions, using a shared Video DiT and a common velocity prediction space.

Dense visual perception. Depth Anything 3 (Lin et al., 2025), Video Depth Anything (Chen et al., 2025b), and VGGT-Ω (Wang et al., 2026a) estimate depth or scene geometry, while LingBot-Depth (Tan et al., 2026a) completes missing or unreliable sensor depth. Grounding DINO (Liu et al., 2024) enables text-conditioned object localization, and SAM 2 (Ravi et al., 2024) and SAM 3 (Meta AI, 2025) support promptable segmentation. For motion perception, RAFT (Teed & Deng, 2020) estimates optical flow and AllTracker (Harley et al., 2025) tracks dense correspondences across video frames. RAFT-3D (Teed & Deng, 2021), Any4D (Karhade et al., 2025), V-DPM (Sucar et al., 2026), and Track4World (Lu et al., 2026) further estimate 3D motion or dense spatiotemporal correspondence. These models provide complementary estimates of scene structure and can supply pseudo-labels for multimodal learning.

3 METHOD

GenSight unifies future video generation and dense perception within a single model. Given an initial image I_0 and a text prompt c, it synthesizes a future RGB video together with aligned embodiment segmentation, depth, and first-frame-relative 2D flow. Given an observed video, the same model predicts each dense modality in a single step without iterative denoising. We denote an RGB video by $V = \{V_t\}_{t=0}^{T}$ and its aligned embodiment segmentation, depth, and first-frame-relative 2D flow by $S = \{S_t\}_{t=0}^{T}$, $D = \{D_t\}_{t=0}^{T}$, and $F = \{F_{0\to t}^{\mathrm{2D}}\}_{t=0}^{T}$, respectively. These modalities characterize complementary aspects of embodied dynamics: appearance (V), semantics (S), geometry (D), and motion (F).

GenSight maps all modalities into a common image-like representation processed by the same frozen video autoencoder (Wan Team, 2025). A shared Video DiT predicts each modality through lightweight task modulation, while generation and perception are jointly learned in a unified velocity prediction space. Figure 2 illustrates the overall pipeline.

3.1 UNIFIED MODALITY REPRESENTATION

Everything-as-RGB Representation. A central challenge is to accommodate heterogeneous perceptual outputs within a model pretrained for RGB video generation. Rather than introducing modality-specific encoders or decoders, we confine modality-specific processing to the representation boundary and map each modality to a three-channel image-like representation. At training time, for each dense modality $m \in \{S, D, F\}$, an encoding mapping ψ_m converts the modality data M_m into an RGB-like sequence, which is then encoded by the same frozen video autoencoder used

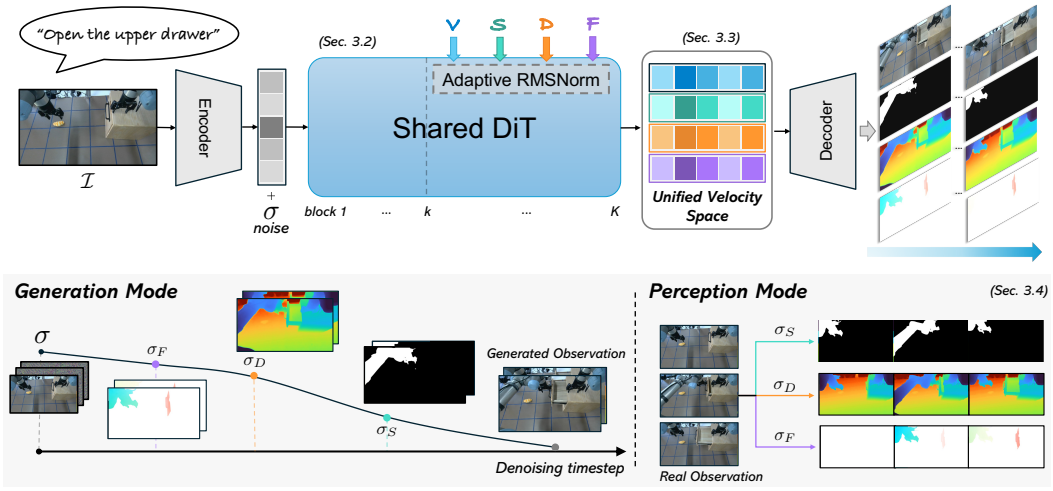


Figure 2: **Overview of GenSight. Top:** A shared Video DiT processes noisy video latents and predicts velocities for RGB video (V), embodiment segmentation (S), depth (D), and first-frame-relative 2D flow (F) in a unified prediction space. The first k blocks compute task-agnostic features, while the remaining blocks apply task-adaptive RMSNorm with shared Transformer weights.The frozen video autoencoder is shared across modalities. **Bottom left:** Given an initial image and text prompt, video generation follows a multi-step denoising trajectory, with dense predictions read at modality-specific noise levels. **Bottom right:** Given an observed video, each dense modality is predicted in one step from a noisy state at its selected noise level. Both regimes share the dense readout and decoding procedure, without independent denoising trajectories for dense modalities.

for RGB to obtain the latent target $z_m = \mathcal{E}(\psi_m(M_m))$. At inference time, the shared decoder $\mathcal{D}$ reconstructs an image-like prediction, which is converted back to the corresponding modality space through a deterministic decoding mapping ρ_m. Together, ψ_m and ρ_m define the representation interface and introduce no learnable modality-specific representation modules. Detailed definitions of the modality-specific transformations ψ_m and ρ_m are provided in Appendix A.

Recovering 3D Scene Flow. The predicted depth and first-frame-relative 2D flow $F^{\mathrm{2D}}_{0\to t}$ can be combined to recover the 3D scene flow $F^{\mathrm{3D}}_{0\to t}$. For a pixel p in the initial frame, its corresponding location at time t is $p_t = p + F^{\mathrm{2D}}_{0\to t}(p)$. Assuming a static camera with constant intrinsic matrix K, the world coordinate system is aligned with the camera coordinate frame, simplifying inter-frame camera motion to the identity transformation. By back-projecting the corresponding pixels into the unified 3D camera coordinate frame, the 3D scene flow is obtained as:

$$P_0 = D_0(p) \cdot K^{-1}\tilde{p}, \qquad P_t = D_t(p_t) \cdot K^{-1}\tilde{p}_t, \qquad F^{\mathrm{3D}}_{0\to t}(p) = P_t - P_0, \tag{1}$$

where $\tilde{p} = [p^\top, 1]^\top$ and $\tilde{p}_t = [p_t^\top, 1]^\top$ denote homogeneous coordinates of p and p_t.

3.2 Unified Model Architecture

Shared Video DiT Backbone. All four tasks are predicted by a single, fully shared Video DiT. The model retains the standard video-generation input formulation, while embodiment segmentation, depth, and 2D flow are predicted as additional outputs rather than provided as input modalities. Let the backbone contain L Transformer blocks. The input is processed once through the first k blocks to obtain task-agnostic spatiotemporal features,

$$h_k = \mathcal{B}_{1:k}(x_\sigma, \sigma, c), \tag{2}$$

where x_σ denotes the noisy video latent at noise level σ, c denotes the text condition, and $\mathcal{B}_{1:k}$ denotes the first k Transformer blocks. The resulting activations h_k are reused across all tasks. The remaining $L - k$ blocks and output projection process the task-conditioned streams in parallel with shared parameters, while task dependence is introduced solely through the adaptive modulation described below. The choice of k controls the balance between shared computation and task-conditioned processing. We use $k = 16$ by default and study the effect of k in Sec. 4.3.

Task Modulation via Adaptive RMSNorm. The pretrained Video DiT used in our model employs RMSNorm (Zhang & Sennrich, 2019), which we extend with lightweight task-adaptive modulation.

For a hidden representation $h \in \mathbb{R}^d$, the pretrained RMSNorm is written as

$$\mathrm{RMSNorm}(h) = \gamma \odot \frac{h}{\sqrt{d^{-1}\sum_i h_i^2 + \varepsilon}}, \tag{3}$$

where γ is the learned scaling parameter inherited from the pretrained model. For each task-conditioned block ℓ and task $\tau \in \{V, S, D, F\}$, a learned modulation table directly provides the corresponding scale, shift, and residual-gate deltas,

$$(\Delta s_{\ell,\tau}, \Delta b_{\ell,\tau}, \Delta g_{\ell,\tau}) = \mathcal{T}(\ell, \tau). \tag{4}$$

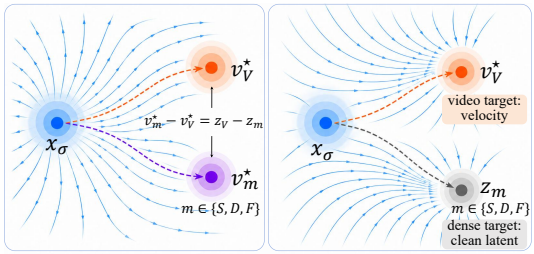
These deltas modulate the normalized features and residual update as

$$\widetilde{h}_{\ell,\tau} = (1 + \Delta s_{\ell,\tau}) \odot \mathrm{RMSNorm}(h) + \Delta b_{\ell,\tau}, \qquad h' = h + (1 + \Delta g_{\ell,\tau}) \odot \mathcal{F}(\widetilde{h}_{\ell,\tau}), \tag{5}$$

where $\mathcal{F}$ denotes the corresponding attention or feed-forward transformation. The attention and feed-forward sub-blocks use separate modulation tables, while all Transformer parameters remain shared across tasks. All modulation parameters are initialized to zero, such that the scale and residual-gate factors start at one and the shifts at zero, exactly preserving the pretrained computation at initialization. Task-specific behavior is then learned through these lightweight modulations during joint training. The RMSNorm preceding the final output projection is modulated in the same manner using task-specific scale and shift, without a residual gate. Consequently, the video autoencoder, Transformer parameters, and output projection are fully shared across modalities, with task specialization introduced only through lightweight task-adaptive modulation.

3.3 Unified Prediction Space

Unified Velocity Prediction. Beyond sharing the modality representation and model architecture, GenSight unifies the prediction parameterization of generation and dense perception. A natural alternative is to directly regress the clean dense latents while retaining velocity prediction for video (Figure 3(b)). This would require the shared model to serve two different target parameterizations: video velocities and clean perceptual latents. We instead retain the pretrained video generator's velocity parameterization across all tasks (Figure 3(a)).



(a) unified target space (b) separate target spaces

Figure 3: Comparison of prediction target spaces.

Let z_V denote the clean video latent and z_m the clean latent of dense modality $m \in \{S, D, F\}$. For a noise level $\sigma \in [0, 1]$ and a Gaussian noise sample $\epsilon \sim \mathcal{N}(0, I)$, we follow the linear interpolation path used in rectified flow and flow matching (Liu et al., 2023; Lipman et al., 2023) and define

$$x_\sigma = (1 - \sigma)z_V + \sigma\epsilon, \qquad v_V^\star = \epsilon - z_V, \qquad v_m^\star = \epsilon - z_m, \quad m \in \{S, D, F\}. \tag{6}$$

All task streams use the same noisy video input x_σ at the same noise level σ, and all velocity targets are constructed using the same noise sample ϵ. The dense latents serve as supervision targets rather than additional input modalities. The interpolation above applies only to the frames to be predicted; the initial frame remains clean as image conditioning and is excluded from all prediction losses.

Cross-Modal Residual Interpretation. The shared noise establishes a direct relation between the video and dense velocity targets:

$$v_m^\star - v_V^\star = (\epsilon - z_m) - (\epsilon - z_V) = z_V - z_m. \tag{7}$$

The noise cancels exactly in their difference, leaving the cross-modal residual between the clean video and dense latents. Thus, relative to the video target, each dense target differs by a noise-free cross-modal residual while retaining the same velocity parameterization. This complements the zero-initialized task modulation in Sec. 3.2, under which all task streams initially inherit the pretrained video generator. The formulation therefore provides a common starting point for task-specific adaptation without switching the dense streams to clean-latent prediction. The dense modalities do not require independent denoising processes. Their reconstruction from the jointly predicted velocities is described in the next subsection.

Joint Training Objective. Let $\hat{v}_\tau$ denote the model's velocity prediction for task $\tau \in \{V, S, D, F\}$. Each task is supervised by the velocity regression loss $\mathcal{L}_\tau^{\text{vel}} = \|\hat{v}_\tau - v_\tau^\star\|_2^2$. For the dense modalities, we additionally apply a spatial-gradient matching loss (Eigen & Fergus, 2015) to supervise local spatial variation. The joint training objective is

$$\mathcal{L} = \sum_{\tau \in \{V,S,D,F\}} w_\tau \lambda_\tau(n) \mathcal{L}_\tau, \qquad \mathcal{L}_V = \mathcal{L}_V^{\text{vel}}, \qquad \mathcal{L}_m = \mathcal{L}_m^{\text{vel}} + \tfrac{1}{2}\|\nabla \hat{v}_m - \nabla v_m^\star\|_2^2. \tag{8}$$

Here, w_τ denotes the task weight, and n denotes the training step. The operator ∇ computes discrete spatial gradients along the height and width dimensions of the latent velocity predictions and targets. We use $w_V = 1$ and $w_S = w_D = w_F = 0.25$ by default. We set $\lambda_V(n) = 1$, while the dense-task warmup factors $\lambda_m(n)$ gradually increase from zero to one during early training to reduce disruption to the pretrained generation capability.

3.4 Inference for Generation and Perception

Building on the unified velocity formulation, GenSight supports two complementary inference regimes with shared parameters and a common dense readout. In generation, dense predictions are read from intermediate states along the video denoising trajectory. In perception, noisy states are constructed directly from an observed video. Figure 2 illustrates the two regimes.

Multi-Step Video Generation with Dense Readouts. Given an initial image I_0 and a text prompt c, GenSight generates a future RGB video through K denoising steps, $x_{\sigma_K} \to x_{\sigma_{K-1}} \to \cdots \to x_{\sigma_0}$, with decreasing noise levels. For each dense modality $m \in \{S, D, F\}$, we empirically select a readout noise level $\sigma_m^\star$, which is used in both inference regimes. The choice of $\sigma_m^\star$ is studied in Sec. 4.3. During generation, the corresponding dense task stream is run only when the video denoising trajectory reaches this readout state.

At $x_{\sigma_m^\star}$, the video and dense task streams produce $\hat{v}_V$ and $\hat{v}_m$ from the same input. The video prediction yields the noise estimate $\hat{\epsilon} = x_{\sigma_m^\star} + (1 - \sigma_m^\star)\hat{v}_V$, and the dense latent is reconstructed as

$$\begin{aligned} \hat{z}_m = \hat{\epsilon} - \hat{v}_m &= x_{\sigma_m^\star} + (1 - \sigma_m^\star)\hat{v}_V - \hat{v}_m \\ &= \hat{z}_V - (\hat{v}_m - \hat{v}_V), \end{aligned} \tag{9}$$

where $\hat{z}_V = x_{\sigma_m^\star} - \sigma_m^\star \hat{v}_V$ is the model's clean video latent estimate at the readout state. The difference $\hat{v}_m - \hat{v}_V$ estimates the cross-modal residual $z_V - z_m$ in Eq. (7). Thus, the dense latent is reconstructed relative to the model's video estimate. The corresponding prediction is obtained as $\hat{M}_m = \rho_m(\mathcal{D}(\hat{z}_m))$, using the shared video decoder $\mathcal{D}$ and the modality decoding mapping ρ_m. Dense predictions are therefore obtained without independent denoising trajectories.

Single-Step Dense Perception. Given an observed video, we encode it as z_V. For a requested modality m, we directly construct the noisy state at its selected noise level,

$$x_{\sigma_m^\star} = (1 - \sigma_m^\star) z_V + \sigma_m^\star \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{10}$$

A single forward pass at this state produces both the video velocity $\hat{v}_V$ and the dense velocity $\hat{v}_m$. We then apply the readout in Eq. (9) and the same decoding procedure to obtain $\hat{M}_m$. Each requested modality is therefore predicted from its selected noisy state without iterative denoising.

4 Experiments

We evaluate GenSight as both a video generator and a dense perception model. The experiments examine *video quality*, the accuracy of *depth*, *embodiment segmentation*, and *scene flow* prediction in both **generation mode** and **perception mode**, and generalization to unseen manipulation tasks and a simulated environment. Our ablation studies address two questions: (1) How many task-agnostic blocks should be used? (2) Which readout noise level is best suited to each dense task?

4.1 Datasets and Metrics

We train GenSight on 246,330 episodes across four embodiments from RoboMIND 2.0 (Hou et al., 2025) and our egocentric human dataset *EgoTable*. Additionally, we evaluate task-level generalization on held-out tasks and assess cross-domain transfer to simulation via RoboTwin 2.0 (Chen et al., 2025c). Dataset composition, split protocols, and dense labels are detailed in Appendix B and C.

Table 1: **Generation mode** on RoboMIND and RoboTwin. Depth and flow are scored with VGGT-Ω and AllTracker, respectively. AbsRel and δ_1 are in %; $\mathrm{AAE}_{\mathrm{top}\,1\%}$ is in degrees; EPE is in pixels.

	Video				Depth		2D Flow		
Method	FVD ↓	FID ↓	PSNR ↑	LPIPS ↓	AbsRel ↓	δ_1 ↑	nEPE ↓	$\mathrm{AAE}_{\mathrm{top}\,1\%}$ ↓	EPE ↓
RoboMIND									
Wan 2.2	122.38	85.32	13.68	0.4725	–	–	–	–	–
RynnWorld4D	73.45	47.90	**18.07**	0.2661	8.53	92.55	1.133	45.14	**4.936**
Cosmos3-Nano	44.97	42.09	16.90	0.3002	–	–	–	–	–
GenSight	**14.55**	**16.12**	18.01	**0.2296**	**7.88**	**93.17**	**0.427**	**23.10**	7.089
RoboTwin 2.0									
RynnWorld4D	71.18	67.65	16.53	0.3839	**4.86**	95.69	1.225	47.49	**7.11**
GenSight	**61.14**	**46.28**	**16.71**	**0.2927**	4.98	**97.25**	**0.892**	**26.78**	16.45

Table 2: **Perception mode** on RoboMIND. RAFT† and RAFT‡ share the 2D EPE, and their 3D scores use ground-truth and our predicted depth, respectively. Ranking is per metric: 1st, 2nd.

(a) **Segmentation**

Method	IoU ↑
SAM 3	73.5
GenCeption	67.5
GenSight	88.7

(b) **Depth**

Method	RMSE ↓	AbsRel ↓	δ_1 ↑
VDA	13.12	7.4	94.0
$\mathrm{DA3}_{\mathrm{mono}}$	15.65	13.2	81.7
$\mathrm{DA3}_{\mathrm{stream}}$	16.17	13.2	85.3
DVD	15.91	7.6	92.9
VGGT-Ω	13.11	5.9	94.5
GenCeption	15.89	8.5	90.7
GenSight	4.76	2.4	99.1

(c) **Flow**

Method	EPE ↓	$\mathrm{ADE}_{\mathrm{abs/disp}}$ ↓	$\mathrm{FDE}_{\mathrm{abs/disp}}$ ↓
RAFT†	11.2	7.04/6.88	7.49/7.40
RAFT‡		9.63/7.68	9.95/7.92
Any4D	18.4	15.26/6.72	16.93/7.97
V-DPM	25.2	24.17/4.88	26.24/7.14
Track4World	6.8	20.31/6.92	23.99/9.06
GenSight	5.2	7.39/7.68	9.51/9.87

For video generation, we evaluate 49-frame text-image-to-video generation using FVD (Unterthiner et al., 2018), FID (Heusel et al., 2017), PSNR, and LPIPS (Zhang et al., 2018), conditioned on the initial frame and caption. **For embodiment segmentation,** we report intersection over union (IoU, %) for masks, with robot arm/gripper or human hand as foreground. **For depth,** we report root mean squared error (RMSE, centimeters), absolute relative error (AbsRel, %), and threshold accuracy (δ_1, %) over valid pixels. **For scene flow,** we report first-frame-relative 2D endpoint error (EPE) and 3D average/final errors (ADE/FDE) in absolute and displacement forms. For generated 2D flow, we additionally report nEPE and $\mathrm{AAE}_{\mathrm{top}\,1\%}$. Definitions appear in Appendix D.

4.2 Comparisons with Baselines

4.2.1 Generation Mode

Baselines. We compare generated RGB video with Wan 2.2 (Wan Team, 2025), RynnWorld4D (Zhao et al., 2026), and Cosmos3-Nano (NVIDIA, 2026). RynnWorld4D builds on Wan 2.2, while GenSight builds on Cosmos3-Nano. RynnWorld4D jointly generates RGB video, depth, and optical flow, but not embodiment segmentation; we therefore compare it with GenSight on TI2V generation, depth, and flow.

Results and discussion. Table 1 reports the quantitative comparison, with qualitative examples shown in Figure 4. On RoboMIND, GenSight achieves the lowest FVD, FID, and LPIPS among the four generators, with PSNR comparable to RynnWorld4D. Its generated depth also improves both AbsRel and δ_1. For 2D flow, GenSight has lower nEPE and $\mathrm{AAE}_{\mathrm{top}\,1\%}$, whereas RynnWorld4D has lower raw EPE, indicating how EPE alone can favor low-motion generations. On RoboTwin 2.0, GenSight improves all four video metrics and depth δ_1. The flow trend is consistent across datasets: GenSight has lower nEPE and top-motion angular error, while RynnWorld4D has lower raw EPE.

4.2.2 Perception Mode

With ground-truth RGB video as input, we compare GenSight against task-specific foundation models on embodiment segmentation, depth, and flow, as shown in Figure 4.

Embodiment segmentation. We compare with text-prompted SAM 3 (Meta AI, 2025) and GenCeption (Wang et al., 2026b), using the same embodiment prompt for both baselines. Table 2a

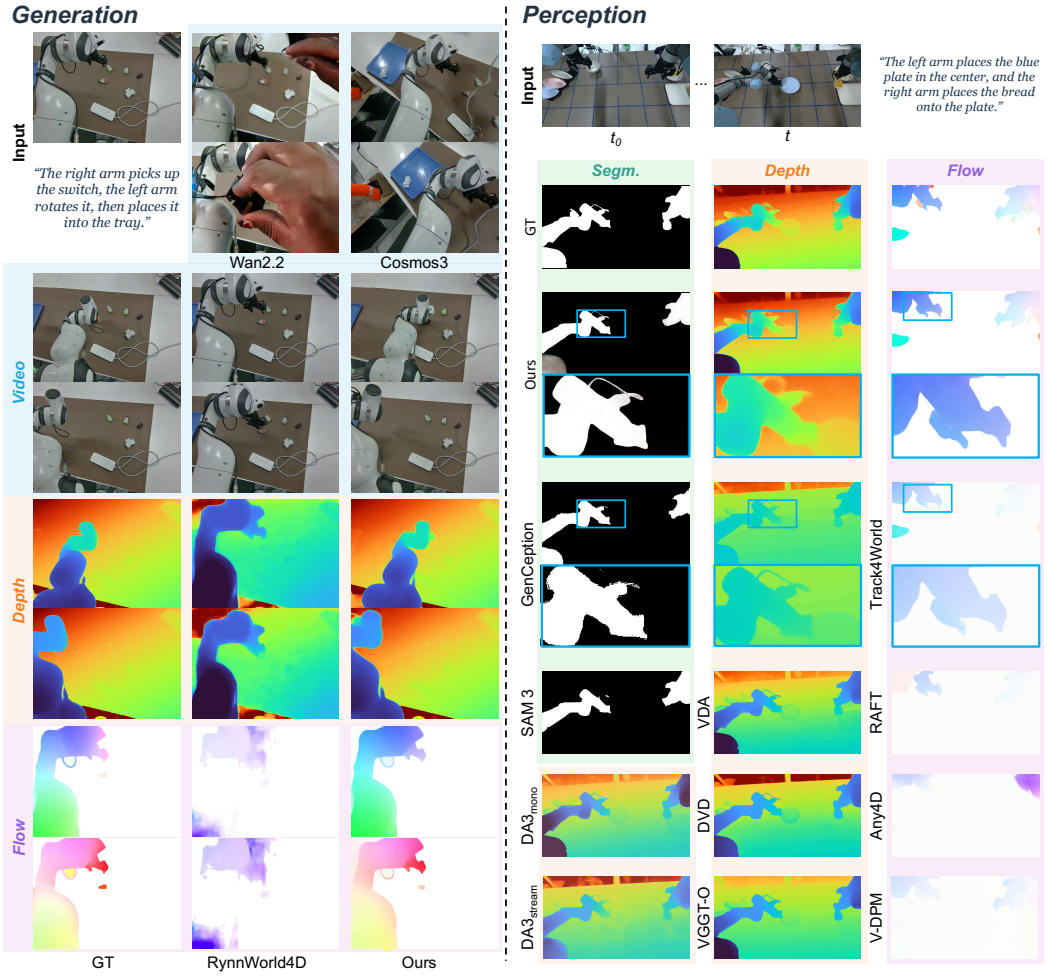


Figure 4: **Qualitative comparison in generation and perception modes. Left (generation):** Given an initial frame and action prompt, the top panels show video examples from Wan 2.2 and Cosmos3-Nano; below, two time frames of video, depth, and 2D flow compare the reference, RynnWorld4D, and GenSight. **Right (perception):** Given observed RGB frames, we compare GenSight with reference maps and task-specific baselines for embodiment segmentation, depth, and 2D flow. Blue boxes mark zoomed-in regions.

shows that GenSight achieves the highest IoU, indicating that its shared representation can also capture embodiment semantics effectively.

Depth estimation. We compare with DA3 (Lin et al., 2025) in monocular and streaming configurations, VGGT-Ω (Wang et al., 2026a), DVD (Zhang et al., 2026), VDA (Chen et al., 2025b), and GenCeption (Wang et al., 2026b). As shown in Table 2b, GenSight performs best across RMSE, AbsRel, and δ_1, while VGGT-Ω is the strongest baseline across these metrics. This suggests that the shared video backbone also supports accurate geometric perception.

Flow estimation. For 2D flow, we compare with RAFT (Teed & Deng, 2020), Any4D (Karhade et al., 2025), V-DPM (Sucar et al., 2026), and Track4World (Lu et al., 2026). For 3D scene flow, we compare with RAFT-3D (Teed & Deng, 2021), Any4D, V-DPM, and Track4World. Table 2c shows GenSight achieves the lowest 2D EPE. For 3D flow, it ranks behind only RAFT-3D with ground-truth depth and outperforms all methods without ground-truth depth. Note that we use absolute trajectory errors as the primary 3D metric, since displacement errors can partially cancel depth discontinuities and alter the ranking. Further analysis is provided in Appendix Figure I.

Generalization across embodiments and to simulation. Figure 5 qualitatively demonstrates GenSight's embodiment controllability on RoboTwin 2.0. Given the same initial frame, it

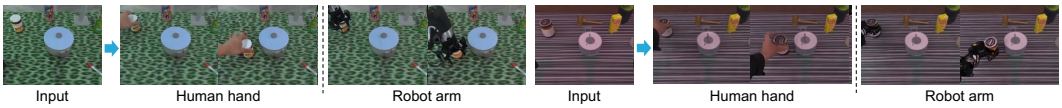


Figure 5: **Embodiment-conditioned generation on RoboTwin 2.0.** Given the same initial frame (left), GenSight generates a human-hand continuation (middle) or a robot-arm continuation (right) under the corresponding embodiment condition.

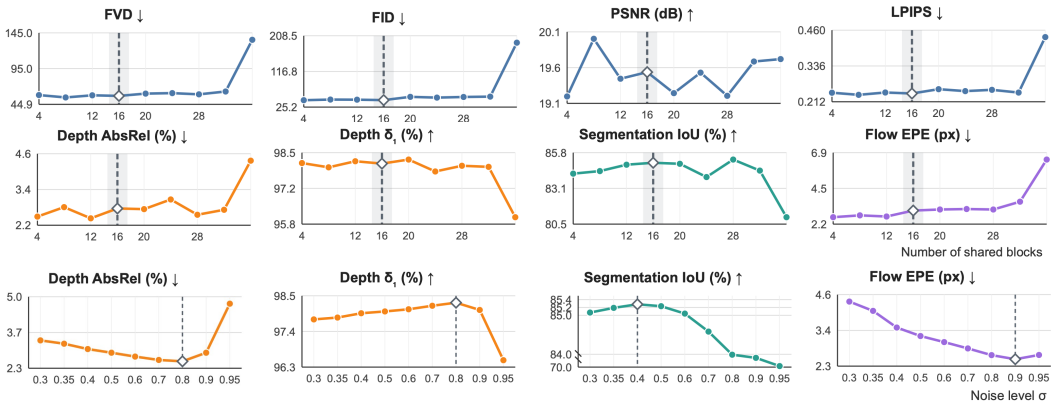


Figure 6: **Ablation studies on UR5. Top two rows:** Video generation and dense perception metrics across task-agnostic prefix lengths k; the dashed line marks our selected configuration. **Bottom row:** Evaluation of a fixed checkpoint across readout noise levels σ, where markers indicate optima.

successfully synthesizes continuation videos for either a human hand or a robot arm in simulated scenes. Additional zero-shot generalization results are shown in Appendix Figure VII.

4.3 Ablations

We examine the task-agnostic prefix length and dense-readout noise levels on held-out UR5 tasks. Experimental settings and detailed per-metric results are provided in Appendix F.

Number of task-agnostic blocks. As shown in Figure 6 (top two rows), the optimal value of k varies across metrics. A long task-agnostic prefix degrades both video generation and dense perception, whereas an intermediate prefix length provides a better overall balance. These results support preserving task-adaptive modulation in the later shared blocks.

Noise level for dense prediction. Figure 6 (bottom row) illustrates that different dense modalities favor distinct readout noise levels. Depth performs best around $\sigma = 0.8$, while segmentation prefers a lower noise level of $\sigma = 0.4$, and flow continues to improve up to $\sigma = 0.9$. These results motivate using separate readout noise levels for different dense tasks.

5 Conclusion

We presented GenSight, a unified embodied world model that brings future video generation and dense perception into a single pretrained Video DiT. GenSight unifies heterogeneous modalities through a common spatiotemporal representation, a shared model architecture with lightweight task adaptation, and a unified velocity prediction space. This formulation enables multi-step future video generation with aligned dense readouts and single-step dense perception from observed videos. Experiments across diverse embodiments, human-hand interactions, unseen tasks, and simulated environments demonstrate strong performance in both regimes and validate the key design choices. Overall, our results show that seeing and imagining can be jointly modeled within a single embodied world model, providing a unified foundation for understanding embodied dynamics.

Limitations and Future Work. GenSight relies partly on sensor completion and off-the-shelf estimators for dense supervision, which may introduce annotation noise and bias. Moreover, the current model focuses on world modeling and does not yet close the loop with action prediction or control. Future work will explore more reliable dense supervision and tighter integration of generation, perception, and action for closed-loop embodied interaction.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing and for research ideation or execution. We have not used generative AI tools for retrieval and discovery, to draft sections of the paper, for generating synthetic datasets, and proving mathematical claims is not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Caiyi Zhang, Peihao Li, Kiwhan Song, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025a.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025c.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, Jianlou Si, Jun Huang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. Flowwam: Optical flow as a unified action representation for world action models, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2650–2658, 2015.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising, 2026a.

Qin Guo, Hao Luo, Dongxu Yue, Weixuan Jin, Xiao Fu, Fan Wang, and Dan Xu. UniGP: Taming diffusion transformer for prior-preserved unified generation and perception, 2026b.

Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas Guibas. Alltracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5253–5262, October 2025.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, 2025.

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

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, Chenyang Gu, Zhuoyang Liu, Nuowei Han, Xiangju Mi, Yaoxu Lv, Yankai Fu, Gaole Dai, Langzhe Gu, Tao Li, Yuheng Zhang, Yixue Zhang, Xinhua Wang, Shichao Fan, Meng Li, Zhen Zhao, Ning Liu, Zhiyuan Xu, Pei Ren, Junjie Ji, Haonan Liu, Kuan Cheng, Shanghang Zhang, and Jian Tang. RoboMIND 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20658–20671, 2025.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4D: Unified feed-forward metric 4D reconstruction. *arXiv preprint arXiv:2512.10935*, 2025.

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*, pp. 9492–9502, 2024.

Michael King, Aravindh Mahendran, Matthew Koichi Grimes, Fedor Kitashov, Adham Elarabawy, Pedro Velez, Maks Ovsjanikov, and Viorica Pătrăucean. Gen4u: Unifying video generation and understanding via diffusion. *arXiv preprint arXiv:2607.06856*, 2026.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. Egowam: World action models beyond pixels with in-the-wild egocentric human data, 2026a.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, Jian Tang, Sirui Han, and Shanghang Zhang. Wam4d: Fast 4d world action model via spatial register tokens, 2026b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4World: Feedforward world-centric dense 3D tracking of all pixels. *arXiv preprint arXiv:2603.02573*, 2026.

Meta AI. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D video reconstruction with dynamic point maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14502–14511, 2026.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception, 2026a.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026b.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Zachary Teed and Jia Deng. RAFT-3D: Scene flow using rigid-motion embeddings. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Video generation models are general-purpose vision learners, 2026b.

Ge Yan, Jinghao Liu, Yuzhi Fan, Lei Cai, Minwen Liao, Jesse Zhang, and Dieter Fox. Flex-π: A multi-stream world-action model with compute flexibility. *arXiv preprint arXiv:2608.10860*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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 (NeurIPS)*, 2024a.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *International Conference on Learning Representations (ICLR)*, 2024b.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. Dreamwam: Beyond rgb future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Jing He, Zixin Zhang, Haodong Li, Yihao Liang, Kanghao Chen, Bin Ren, Xu Zheng, Shuai Yang, Kun Zhou, Yinchuan Li, Nicu Sebe, and Ying-Cong Chen. DVD: Deterministic video depth estimation with generative priors, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Haoyu Zhao, Xingyue Zhao, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. Rynnworld-4d: 4d embodied world models for robotic manipulation, 2026.

Shuhong Zheng, Zhipeng Bao, Ruoyu Zhao, Martial Hebert, and Yu-Xiong Wang. Diff-2-in-1: Bridging generation and dense perception with diffusion models. In *International Conference on Learning Representations*, 2025.

Supplementary Material for "Seeing and Imagining as One: Unified Video Generation and 4D Perception for Embodied World Modeling"

A Dense modality encoding and decoding

1. Depth Representation

To effectively process metric depth, we operate in the logarithmic domain, apply per-clip percentile min–max normalization, and convert the normalized values into a three-channel thermometer representation. This design choice brings three main benefits: (1) logarithmic scaling compresses distant ranges while preserving close-range details; (2) percentile normalization prevents extreme depth outliers from destabilizing scale and gradients; (3) a 768-level RGB thermometer code significantly expands dynamic range beyond standard 8-bit grayscale to reduce quantization errors. Specifically, these processes are formulated as follows:

Depth Normalization Given a metric depth map D, we first transform it to the logarithmic domain $y = \ln D$. We then apply min–max normalization within the $[1\text{st}, 99\text{th}]$ non-sky percentile interval $[y_{\min}, y_{\max}]$:

$$\hat{D} = \frac{\ln D - y_{\min}}{y_{\max} - y_{\min}} \tag{11}$$

Depth Encoding The normalized depth $\hat{D} \in [0, 1]$ is mapped onto a 768-level discrete scale $S = \lfloor 767 \cdot \hat{D} \rceil \in \{0, \dots, 767\}$, and then converted into a three-channel RGB thermometer code $\mathbf{C} = [R, G, B]^\top$:

$$\mathbf{C}(\hat{D}) = \min\left(\max\left(\begin{bmatrix} S \\ S - 256 \\ S - 512 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right), \begin{bmatrix} 255 \\ 255 \\ 255 \end{bmatrix}\right) \tag{12}$$

Depth Decoding The metric depth D is reconstructed directly from the RGB vector $\mathbf{C}$ via:

$$S = R + G + B, \qquad \hat{D} = \frac{S}{767} \tag{13}$$

$$D = \exp\Big(y_{\min} + \hat{D} \cdot (y_{\max} - y_{\min})\Big) \tag{14}$$

2. Segmentation Representation

Robot, arms, grippers, and human hands are represented as three-channel binary masks and recover clean binary targets via thresholding during decoding to ensure sharp binary outputs. Specifically:

Segmentation Encoding Given a binary mask $M \in \{0, 1\}^{H \times W}$, we replicate it across three RGB channels to form $\mathbf{M} \in \{0, 255\}^{3 \times H \times W}$:

$$\mathbf{M} = 255 \cdot \begin{bmatrix} M \\ M \\ M \end{bmatrix} \tag{15}$$

Segmentation Decoding Given the reconstructed continuous values $\hat{\mathbf{M}} \in [0, 255]^{3 \times H \times W}$, the binary mask $\hat{M}$ is recovered by channel-averaging and applying a threshold $\tau = 127$:

$$\hat{M} = \mathbb{I}\left(\frac{1}{3} \sum_{c \in \{R,G,B\}} \hat{\mathbf{M}}_c > 127\right) \tag{16}$$

3. FLOW REPRESENTATION

We encode the two-channel displacement field on an HSV color wheel, so that flow becomes an ordinary RGB video consumable by the unchanged visual tokenizer.

Flow Normalization Let $F^{\text{grid}} \in \mathbb{R}^{T\times g_h\times g_w\times 2}$ be the raw first-frame-anchored flow, expressed in pixels of a flow grid of width g_w. We first rescale it onto a common canvas of width $W = 1280$ using a single scalar, which leaves the displacement direction invariant:

$$F = \frac{W}{g_w} F^{\text{grid}} \tag{17}$$

We then remove global spurious motion (such as camera shake or collisions) via a per-pixel temporal gate: a pixel p is retained only if its trajectory ever exceeds $\tau = 100$, and is otherwise zeroed across the entire clip:

$$K(p) = \mathbb{I}\Big(\max_t \|F_t(p)\|_2 \;\geq\; 100\Big), \qquad F_t(p) \leftarrow K(p)\, F_t(p) \tag{18}$$

Flow Encoding Write $\tilde{F} = (u, v)$ for the normalized flow at a pixel, expressed in polar coordinates:

$$\theta = \operatorname{atan2}(v, u) \in [-\pi, \pi), \qquad m = \sqrt{u^2 + v^2} \tag{19}$$

Direction is mapped to hue, magnitude to saturation, and value is held at its maximum. Using the 8-bit convention ($H \in \{0, \ldots, 179\}$):

$$H = \left\lfloor 179 \cdot \frac{\theta + \pi}{2\pi} \right\rfloor, \qquad S = \left\lfloor 255 \cdot \min\Big(\frac{m}{\sigma}, 1\Big) \right\rfloor, \qquad V = 255 \tag{20}$$

where $\sigma > 0$ is a single dataset-wide magnitude scale estimated as the 99th percentile of moving pixel magnitudes. The final RGB code is obtained via the standard conversion:

$$\mathbf{C} = \mathcal{H}^{-1}(H,\, S,\, V) \in \{0, \ldots, 255\}^3 \tag{21}$$

Zero motion and invalid pixels are both assigned the neutral white code ($S = 0, V = 255$) to prevent injecting spurious hue.

Flow Decoding Given a reconstructed RGB code $\hat{\mathbf{C}}$, the optical flow direction and magnitude are recovered by inverting the color wheel:

$$\Big(\hat{H}, \hat{S}, \hat{V}\Big) = \mathcal{H}\Big(\hat{\mathbf{C}}\Big) \tag{22}$$

$$\hat{\theta} = \frac{\hat{H}}{179} \cdot 2\pi - \pi, \qquad \hat{m} = \frac{\hat{S}}{255} \cdot \sigma \tag{23}$$

$$\hat{u} = \hat{m} \cos\hat{\theta}, \qquad \hat{v} = \hat{m} \sin\hat{\theta} \tag{24}$$

Flow quality is evaluated using the End-Point Error (EPE) over the valid pixel region Ω:

$$\text{EPE} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \|(\hat{u}, \hat{v})\,(p) - (u, v)\,(p)\|_2 \tag{25}$$

B DATASETS AND EVALUATION PROTOCOL

Data sources. We combine RoboMIND 2.0 (Hou et al., 2025) episodes from UR5, Franka, Tianyi, and Tienkung with *EgoTable*, our in-house collected egocentric dataset for human tabletop manipulation, as exemplified in Figure VI. It encompasses over 5,000 distinct tasks, featuring 200+ manipulated objects, 100+ reference objects, and 6 explicit spatial relations. For both RoboMIND 2.0 and EgoTable, each episode provides RGB video, synchronized sensor depth, and an action caption. We augment captions with domain tags specifying the data type, data source, and arm type, so generation can be conditioned on the input embodiment. Dataset statistics are shown in Table I.

Table I: Dataset statistics across different sub-datasets.

Category	Sub-dataset	Count
RoboMIND 2.0	UR5	49,941
	Tianyi	7,143
	Tienkung	7,352
	Franka	135,423
EgoTable	*In-house Tasks*	46,471
Total		**246,330**

Evaluation splits. For the training datasets, we reserve all episodes from 5% of tasks for out-of-distribution evaluation and use the remaining 95% of tasks for training. The RoboMIND 2.0 test set contains 600 episodes from 24 unseen real-world tasks. For real-to-simulation evaluation, we use RoboTwin 2.0 (Chen et al., 2025c), comprising 300 episodes from 50 simulated tasks.

C Pseudo-label generation

Dense modalities are labeled offline. *Depth:* The dataset provides ground-truth depth acquired via depth cameras; however, the raw depth maps suffer from missing values (holes). We employ Lingbot-Depth (Tan et al., 2026b) to perform depth completion and recover dense depth representations. *Segmentation:* GroundingDINO (Liu et al., 2024) detects the arm/hand with embodiment-specific prompts (robot arm/gripper vs. human hand), seeding SAM2 (Ravi et al., 2025), which is propagated through the clip with periodic re-detection. *Scene flow:* AllTracker (Harley et al., 2025) produces dense t_0-relative fields, followed by a per-pixel temporal max-flow filter to remove outliers. The 2D fields can be combined with depth for 3D scene-flow evaluation.

D Flow evaluation metrics

We evaluate flow relative to the first frame. For 2D flow, endpoint error (EPE, pixels) is averaged over valid pixels. Let $e_i = \|\hat{\mathbf{f}}_i - \mathbf{f}_i\|_2$ and $g_i = \|\mathbf{f}_i\|_2$ be the error and ground-truth flow magnitude at valid pixel i. For generated flow, we also report normalized EPE, $\mathrm{nEPE} = \sum_i e_i / \sum_i g_i$, to account for the amount of reference motion. We additionally report $\mathrm{AAE}_{\mathrm{top}\,1\%}$, the mean angular error (in degrees) over the 1% of valid pixels with the largest ground-truth flow magnitude. These metrics complement raw EPE, which can favor near-static predictions when reference motion is small. For 3D scene flow, we report average and final trajectory errors (ADE and FDE, centimeters) in both absolute-position and first-frame-relative displacement forms. The displacement variants can partly cancel depth jumps, so we emphasize absolute errors in the main comparison.

E Training details

Backbone The backbone is a 36-block diffusion transformer with hidden width 4096, initialized from a Cosmos3-Nano (NVIDIA, 2026); the visual tokenizer is a frozen Wan2.2 VAE (Wan Team, 2025) with $(4, 16, 16)$ spatiotemporal compression, and text is tokenized by a Qwen3-VL-8B-Instruct tokenizer (Yang et al., 2025).

Trainable and Frozen Components The backbone carries two parallel expert pathways per block, a reasoner and a generator, both initialized from Qwen3-VL-8B. We train the generator and freeze the reasoner throughout: within the transformer blocks, this splits the parameters into 396 trainable and 396 frozen tensors.

Resolution, Clip Length, and Token Budget Rather than fixing a batch size, samples are packed sequentially into a fixed token budget of $12{,}288$ vision tokens per device per step, drawing samples

until the budget is filled. The token count of a clip is

$$N_{\text{tok}} = \left\lfloor \frac{H}{32} \right\rfloor \cdot \left\lfloor \frac{W}{32} \right\rfloor \cdot \left(1 + \frac{T-1}{4}\right) \tag{26}$$

where the spatial divisor is the product of the VAE spatial compression 16 and the patch size 2. The budget therefore trades resolution against duration: it admits 49 frames at 704×1280, 149 frames at 480×640, and 577 frames at 256×320. Because packing is pull-until-budget, shortening individual clips does not reduce memory, it only admits more clips per step; the token budget is consequently the only effective memory control. Gradients are not accumulated.

Optimizer and Learning-Rate Schedule We use fused AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-6}$, no weight decay, and gradient-norm clipping at 1.0. The peak learning rate is 8×10^{-5}. The schedule is a linear warmup over the first 500 steps from 0 to the peak, followed by a single cosine decay cycle scheduled over $8{,}000$ steps toward a floor of 1×10^{-5}.

Dense-task Warmup Factor The loss weights of the three dense modalities are additionally ramped by a warmup factor $\lambda_m(n)$, which rises from 0 to 1 on a cosine over the first 10% of the scheduled run, i.e. $n_\lambda = 800$ steps:

$$\lambda_m(n) = \tfrac{1}{2}\left(1 - \cos \pi \min\left(\frac{n}{n_\lambda}, 1\right)\right) \tag{27}$$

The video branch is never scaled by λ. The purpose is to protect the pretrained image-to-video capability.

Compute Budget Training ran on 64 NVIDIA A800 GPUs under hybrid sharded data parallelism. Each optimizer step takes 37.5 s wall clock, dominated by the $K = 4$ fanout through the specialized tail. The run consumed 85 hours for $8{,}000$ steps, approximately $5{,}400$ GPU-hours.

F Ablation settings and detailed results

Protocol. All ablation models are trained exclusively on the 95% training-task split of RoboMIND 2.0 UR5 and evaluated on the remaining 5%: eight unseen tasks with 25 episodes each, for 200 test episodes. We evaluate checkpoints after 3,000 training steps. Unless varied, the task-agnostic prefix contains 16 transformer blocks and the prediction noise level is $\sigma = 0.6$. The sharing and noise sweeps use equal modality weights; the modality-composition study assigns a weight of 0.25 to each active dense objective and 1 to TI2V.

Number of task-agnostic blocks. Figure 6 shows that the optimal degree of task-agnostic sharing varies across metrics, with no single prefix length dominating all tasks. As the task-agnostic prefix becomes too long, both video generation and dense perception exhibit a broad performance degradation, suggesting that excessive sharing limits modality-specific adaptation. An intermediate prefix provides the best overall trade-off across video quality, depth, segmentation, and flow. We therefore adopt a balanced intermediate configuration, supporting the design choice of preserving task-adaptive capacity in the later blocks.

Dense-readout noise. Figure 6 evaluates a fixed checkpoint across different prediction-noise levels. The video metrics remain unchanged because the sweep only alters the dense readout state. In contrast, the dense modalities exhibit distinct trends: depth improves toward relatively high noise levels before degrading at the extreme end of the sweep; segmentation performs best at lower noise levels and gradually deteriorates as the noise increases; and flow improves over most of the range before saturating and slightly reversing at the highest noise level. These modality-dependent trends motivate using separate readout noise levels for depth, segmentation, and flow.

G Additional results

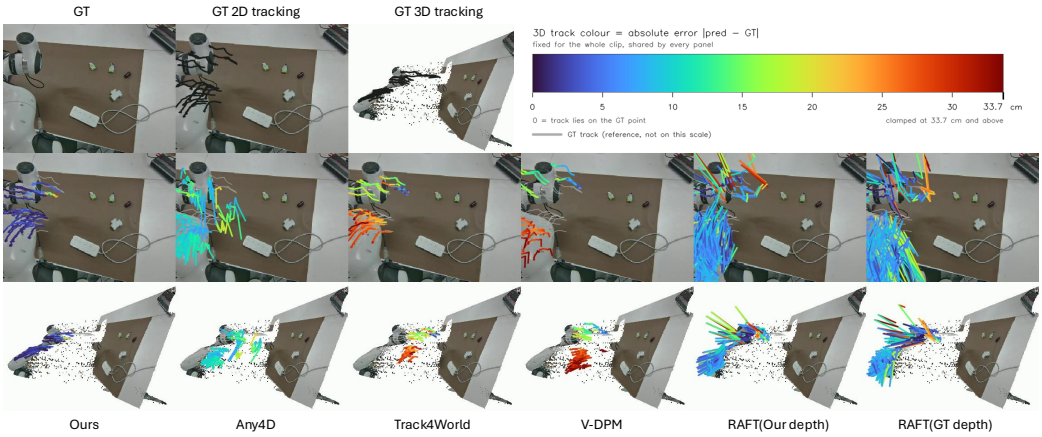


Figure I: **Qualitative 3D Tracking Results.** We estimate 3D scene flow by combining depth and 2D optical flow. Benefiting from accurate Z-axis depth estimation, our method significantly outperforms existing approaches in absolute ADE/FDE metrics. This advantage stems from the fact that alternative methods suffer from residual spatial misalignments even after depth-scale alignment with ground truth. Conversely, our slightly inferior performance in displacement metrics (e.g., ADE-disp/FDE-disp) is primarily attributed to frame-to-frame depth jitter. As shown in the figure, although Tracking4World's Lu et al. (2026) trajectory aligns well with the ground truth on the 2D image plane, a clear depth misalignment with the ground truth is visible when viewed from the right perspective. Trajectory color denotes the magnitude of absolute error.

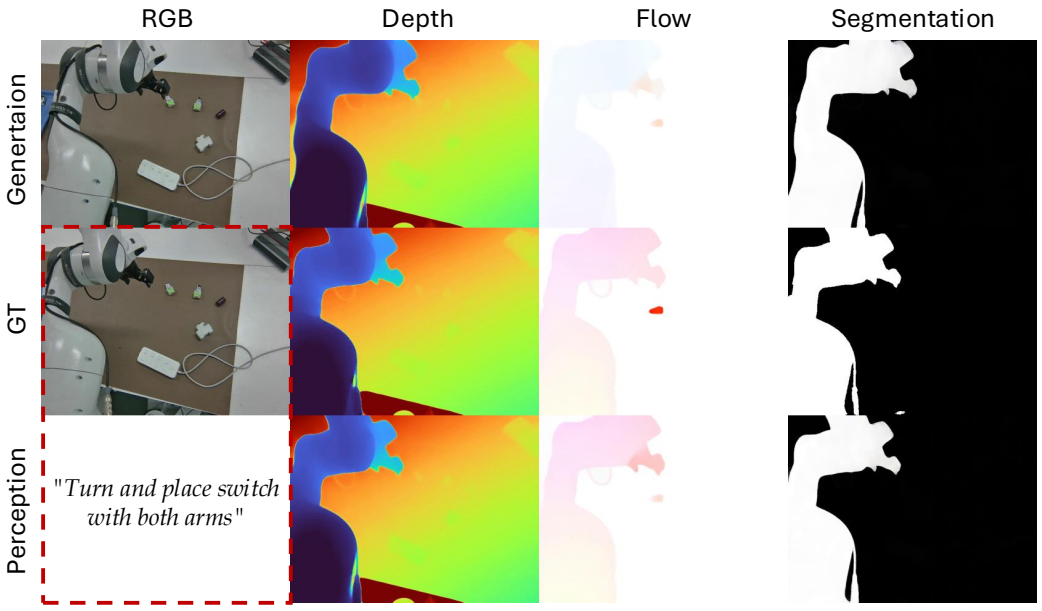


Figure II: **Qualitative Results on RoboMIND 2.0 (Franka Embodiment)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.

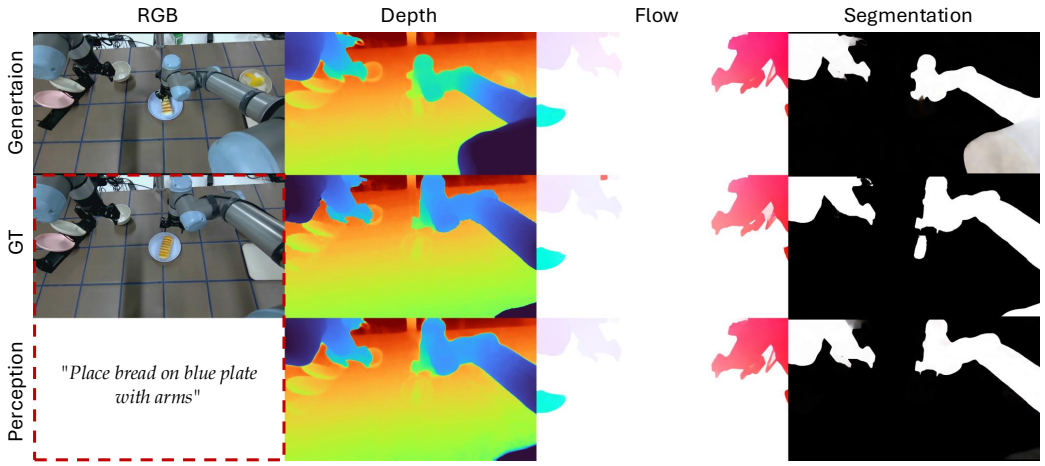


Figure III: **Qualitative Results on RoboMIND 2.0 (UR5 Embodiment)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.

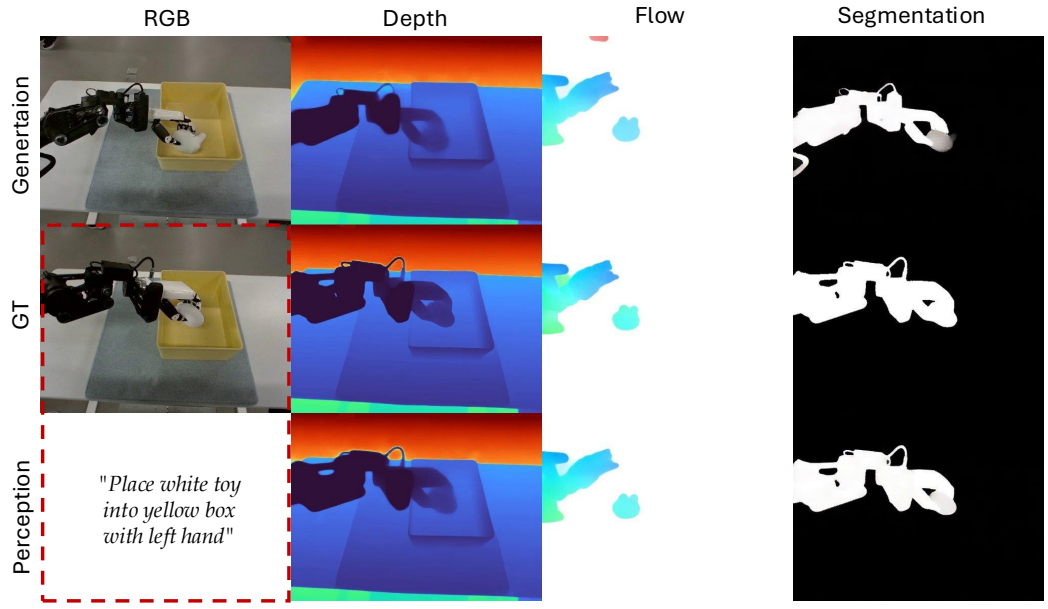


Figure IV: **Qualitative Results on RoboMIND 2.0 (Tianyi Embodiment)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.

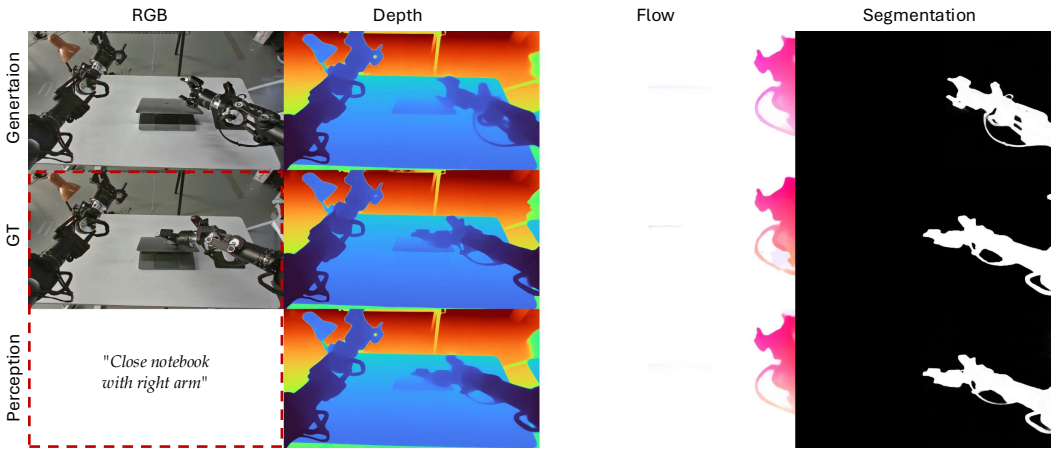


Figure V: **Qualitative Results on RoboMIND 2.0 (Tienkung Embodiment)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.

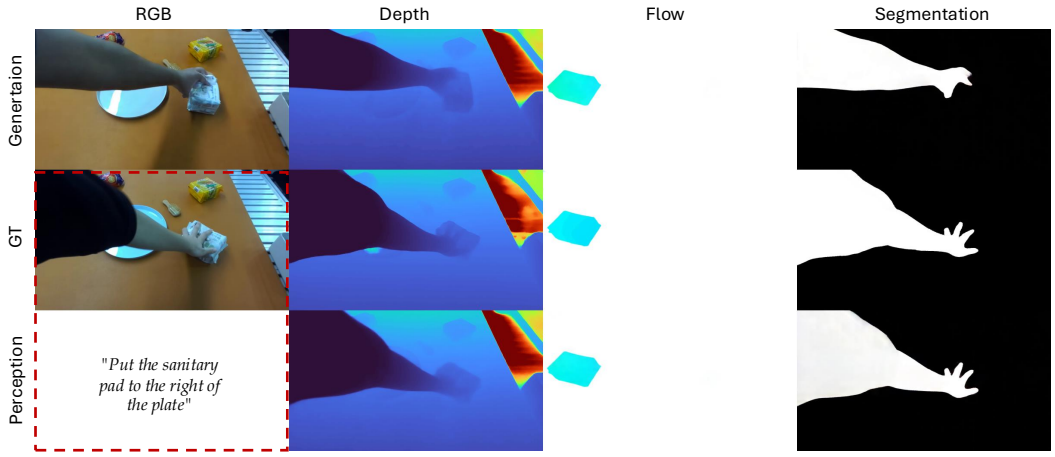


Figure VI: **Qualitative Results on EgoTable (Huaman Hand)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.

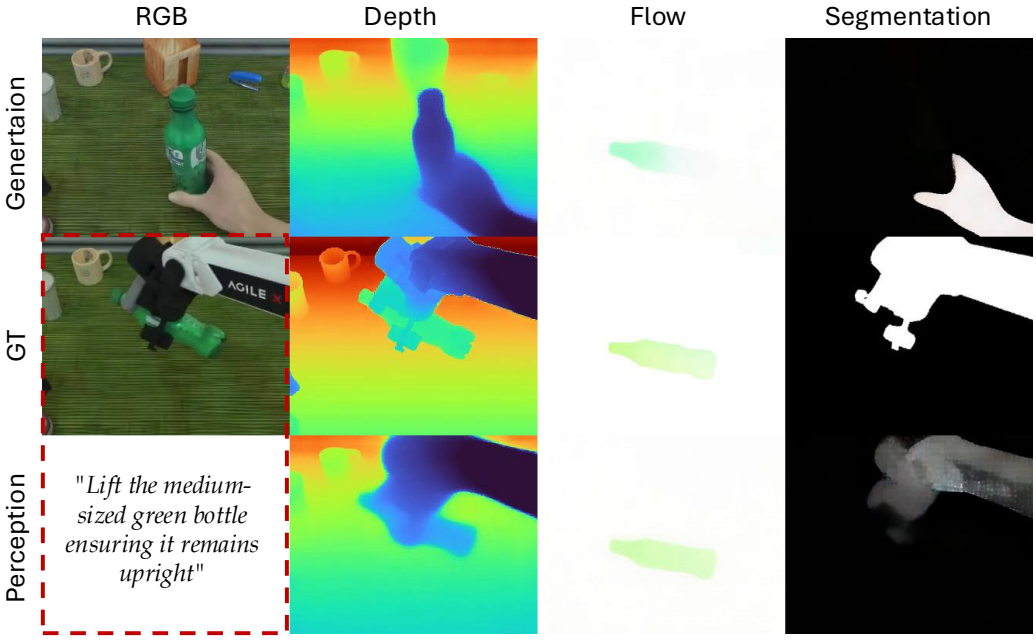


Figure VII: **Qualitative Results on Robotwin2.0 (Simulated AgileX Embodiment)** Middle row: Ground truth (GT). Top row: Generative mode result. Bottom row: Perception mode prediction.