WEAVE-0.5: WEAVING MULTI-STREAM OBSERVATIONS INTO EMBODIED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual intelligence provides the perceptual grounding for embodied agents to learn how the physical world evolves, calling for world models that aggregate diverse and complementary visual evidence. However, interfaces built around fixed multi-view layouts or modality-specific branches constrain joint modeling across variable camera configurations, resolutions, modality sets, and observation conditions. We argue that embodied world models should represent heterogeneous observations as visual streams and flexibly predict requested streams from the available visual context. We present **WEAVE-0.5**, an exploration of multi-stream embodied visual world modeling. Each stream corresponds to one visual modality from one viewpoint and is independently encoded at its native resolution. Streams are organized on a shared virtual canvas and jointly processed with a video diffusion transformer. Canvas coordinates support multi-view and multi-resolution composition while retaining the pretrained spatiotemporal rotary positional encoding (RoPE) structure, whereas diffusion timesteps specify observed and target content. This formulation unifies future prediction, cross-view completion, modality conversion, and joint multi-view generation within a shared architecture without task-specific prediction heads. Built on Wan2.2-TI2V-5B, WEAVE-0.5 is jointly trained on real-robot, simulated, egocentric, and general videos. Component interventions support a contribution of cross-stream attention to RGB prediction of views without their own initial observations. Benchmark-specific WEAVE-0.5 checkpoints rank first in the reported RBench leaderboard snapshot with an average score of 0.642 and achieve 82.9 overall on PAI-Bench-G. These results support multi-stream visual modeling as an effective approach to integrating heterogeneous visual observations for embodied world modeling.

1 INTRODUCTION

"It's not the Thread, it's the Weave."
David Farland

Different sensors draw different threads of evidence from the same physical world. RGB and depth give these threads different colors, while camera views reveal different parts of the same interaction. Robot recordings (Bu et al., 2025; Hou et al., 2025), egocentric video (Hoque et al., 2026), and simulation (Agarwal et al., 2025; Bjorck et al., 2025) already provide a rich collection of such evidence. Yet these recordings differ in camera count, resolution, and available signals. These differences make it difficult to use recordings from different platforms in a single training process. At inference, the available visual context may likewise differ, with observations from only a subset of views or a short temporal prefix. Learning from this diversity calls for a shared visual world model that can bring these streams together and predict how the scene evolves over time (Zhu et al., 2025; Bi et al., 2026; Lu et al., 2026; Wen et al., 2026).

Two related interface questions concern how streams can be combined and how they can be used as conditions or prediction targets. For stream composition, observations can be organized through predefined spatial layouts for camera views (Liu et al., 2026) or dedicated channels and branches for additional signals (Zhang et al., 2023; Alhaija et al., 2025; Jiang et al., 2025b). Recent approaches broaden this flexibility through variable-view generation (Wang et al., 2026b) and flexible input-output configurations (Fan et al., 2026; Agarwal et al., 2026). For prediction roles, modality-specific

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: **Overview of WEAVE-0.5.** A single model processes heterogeneous visual streams conditioned on language and visual observations. The task graph assigns known (K), generated (G), or prefix-conditioned (P) roles to support future prediction, cross-view completion, bidirectional cross-modal prediction, and joint generation. Side panels show data sources and multi-view, multi-modal outputs; the lower panel presents selected public-benchmark comparisons.

noise levels vary which signals serve as conditions and which are prediction targets (Bao et al., 2023; Li et al., 2025; Huang et al., 2025). Joint training across heterogeneous corpora requires this flexibility at the stream level, where spatial resolutions and available observations can differ. A common interface needs to allow both the composition of the streams and their condition and target roles to vary across samples.

We introduce **WEAVE-0.5**, a video world model built around a common *stream* representation (Figure 1). Each stream contains a sequence of one visual modality from one viewpoint and is independently encoded at its native resolution by a shared visual encoder. Samples may contain different numbers and combinations of streams, so recordings with different camera and modality configurations can share the same model. Dense labels derived from RGB can also be included as additional streams where available. The resulting token sequences are concatenated and processed with joint attention in a pretrained video diffusion backbone (Team Wan et al., 2025), allowing streams within a sample to exchange information.

Within this joint sequence, the model must distinguish what each stream represents, how its tokens are positioned, and which tokens are observed or to be predicted. View and modality embeddings identify each stream. A shared *virtual canvas* organizes token positions across views and resolutions. Offsets place different viewpoints in separate canvas regions, while spatially aligned modalities use corresponding coordinates with spacing adjusted to their resolutions. This retains the backbone's pretrained spatiotemporal rotary positional encoding (RoPE) structure. Diffusion timesteps specify which tokens remain clean as conditions and which are diffused as prediction targets; an observed temporal prefix can likewise condition the prediction of subsequent frames (Song et al., 2025; Kong et al., 2024; Yin et al., 2025). Exchanging these assignments changes the prediction direction without changing the model: RGB can condition depth generation, depth can condition RGB generation, and one camera view can condition the completion of another. Future prediction, cross-view com-

pletion, modality conversion, and joint generation are thus expressed through different stream-role configurations (Table 4).

Built on Wan2.2-TI2V-5B (Team Wan et al., 2025), WEAVE-0.5 is jointly trained on robot, simulated, egocentric, and general video. Component interventions support a contribution of cross-stream attention to RGB prediction of views without their own initial observations. Benchmark-specific WEAVE-0.5 checkpoints rank first in the reported RBench (Deng et al., 2026) leaderboard snapshot with an average score of 0.642 and achieve the highest overall score among the compared models on PAI-Bench-G (Zhou et al., 2026) (82.9).

Main contributions.

- WEAVE-0.5 brings embodied corpora with different camera configurations, native resolutions, and modality sets into a unified training framework that also supports flexible prediction directions, allowing heterogeneous observations and visual tasks to share one pretrained video backbone.
- We combine independent native-resolution encoding and sequence packing with a shared virtual canvas that retains the pretrained spatiotemporal RoPE structure. Offsets and resolution-dependent strides organize positions, while diffusion timesteps assign prediction roles.
- We evaluate WEAVE-0.5 across embodied visual generation tasks and public benchmarks; component interventions examine the contribution of cross-stream attention to RGB prediction under missing per-view observations.

WEAVE-0.5 offers an initial exploration of extending video foundation models from single-stream generation to joint multi-stream visual modeling, providing a basis for incorporating a broader range of renderable observations within a common predictive framework.

2 Related Work

2.1 Embodied World Models

World-model research spans predictive latent or feature dynamics for planning and policy learning (Ha & Schmidhuber, 2018; Hafner et al., 2025; Hansen et al., 2024; Zhou et al., 2024), spatial world generation and simulation (World Labs, 2025; Team, 2026b), and future visual prediction for interactive simulation (Yang et al., 2024; Bruce et al., 2024; Wu et al., 2024; Alonso et al., 2024; Chi et al., 2025). Within video-based world modeling, latent diffusion and diffusion transformers provide generative architectures (Rombach et al., 2022; Peebles & Xie, 2023), while large-scale video pretraining supplies transferable appearance and motion priors (Blattmann et al., 2023; Team Wan et al., 2025). Recent models extend these research directions through embodied-oriented video training (Ali et al., 2025; Ma et al., 2026). Cross-platform learning motivates shared interfaces for heterogeneous observations. Vidar combines multi-view images with robot, camera, and task descriptions in textual instructions (Feng et al., 2025). Qwen-RobotWorld uses language as a shared action interface (Zhang et al., 2026b), whereas Cosmos 3 jointly models language, visual, audio, and action sequences (Agarwal et al., 2026). These efforts motivate video models that can accommodate differences in camera configuration, spatial resolution, and available visual signals.

2.2 Unified Multi-Stream Generation

Unified generation methods broaden which visual signals can serve as conditions and prediction targets. VACE organizes reference images, control videos, and editing masks through a unified Video Condition Unit (Jiang et al., 2025b). UniGP unifies geometry-conditioned image generation, dense prediction, and joint generation (Guo et al., 2026a), while MODUS combines autoregressive prediction and flow matching within a decoder-only framework for cross-modal and chained generation (Ye et al., 2026). For multi-view generation, MVISTA-4D combines RGB-D generation with camera-guided cross-view attention to support variable view counts and masked completion (Wang et al., 2026b). OmniView places clean observations and noisy targets from different views and times

in a joint sequence for camera-conditioned view synthesis (Fan et al., 2026). These approaches extend visual prediction across viewpoints and observation configurations.

Unified diffusion provides a complementary mechanism for assigning prediction roles. UniDiffuser and OmniFlow use modality-specific timesteps to support conditional and joint generation (Bao et al., 2023; Li et al., 2025). Diffusion Forcing and DFoT independently sample noise levels for sequence elements or video frames, enabling flexible conditioning on observed history (Chen et al., 2024a; Song et al., 2025). For heterogeneous embodied recordings, WEAVE-0.5 focuses on jointly handling variable stream composition, unequal spatial grids, and observation prefixes. Independently encoded streams are organized on a shared virtual canvas, while diffusion timesteps specify observed and target content. This formulation allows heterogeneous observations and prediction tasks to share one pretrained video backbone and training objective.

3 METHOD

WEAVE-0.5 jointly models visual streams that vary in viewpoint, modality, resolution, and prediction role. Each stream is independently encoded at its native resolution, and the resulting tokens are concatenated into a shared sequence for conditional generation (Figure 2). Stream configurations and prediction roles are specified through metadata, allowing diverse tasks to share the same backbone without task-specific prediction heads. Appendix 6.2 compares the reported visual capabilities of selected related models.

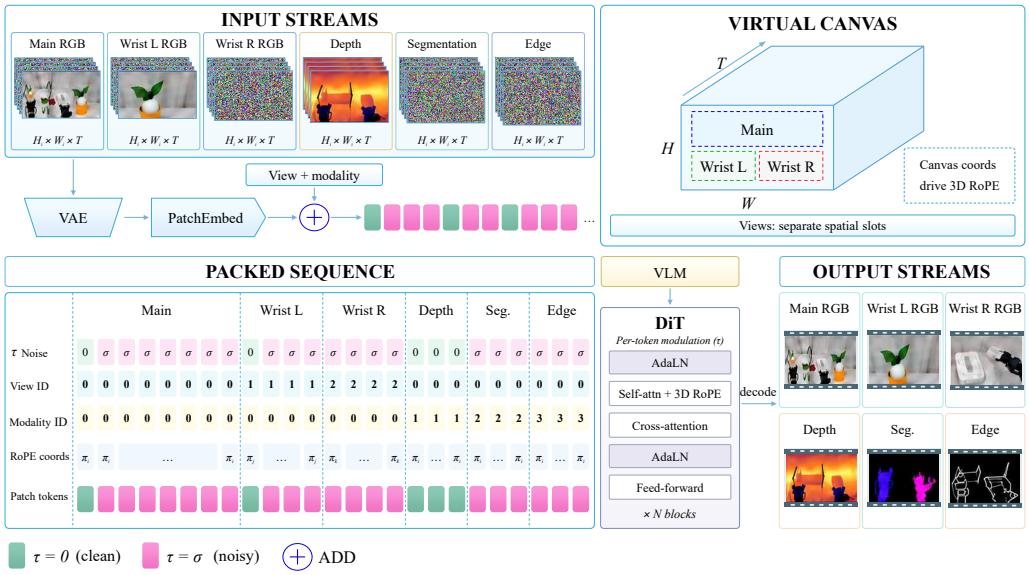

Figure 2: **Framework of WEAVE-0.5.** Temporally aligned streams are independently encoded at their respective resolutions and packed with view and modality embeddings for joint processing by a video diffusion transformer. The virtual canvas separates viewpoints, while aligned modalities share origins with resolution-dependent strides. Framewise timesteps specify observed content and prediction targets; the VLM provides semantic context from the instruction and an observed main-view RGB frame. Target latents are denoised and decoded while observations remain fixed. Thumbnail T counts input frames, whereas canvas time indexes latent frames. Noisy thumbnails illustrate latent-space noising.

3.1 PROBLEM FORMULATION

Definitions. An episode records a single interaction. Each training sample is a temporal segment of an episode and contains a variable number of temporally aligned *streams*, each corresponding to one visual modality from one viewpoint. Examples include main-view RGB, wrist-view RGB, and depth or other dense visual labels. To organize token positions across streams, we introduce

a shared coordinate system, termed the *virtual canvas*, for rotary positional encoding (RoPE). Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ denote a sample containing N streams, where i indexes the streams. We represent each stream as

$$s_i = (\mathbf{z}_i, v_i, m_i, \mathbf{o}_i, \delta_i, r_i), \tag{1}$$

where $\mathbf{z}_i$ denotes the latent representation of the visual clip with T_i latent frames, v_i and m_i identify its viewpoint and modality, and $\mathbf{o}_i$ and δ_i specify its offset and spatial stride on the virtual canvas, respectively. The role r_i partitions the latent-frame indices $\mathcal{I}_i = \{0, \dots, T_i - 1\}$ into observed and predicted subsets:

$$\mathcal{C}_i = \begin{cases} \mathcal{I}_i, & r_i = \texttt{clean}, \\ \{0, \dots, k-1\}, & r_i = \texttt{prefix}_k, \quad 1 \le k < T_i, \\ \varnothing, & r_i = \texttt{diffuse}, \end{cases} \qquad \mathcal{P}_i = \mathcal{I}_i \setminus \mathcal{C}_i. \tag{2}$$

Here, k is an integer prefix length measured in latent frames. A prefix stream thus provides observed history, while its remaining latent frames serve as prediction targets. Each generative query must include at least one target latent frame, i.e., $\sum_{i=1}^{N} |\mathcal{P}_i| > 0$.

Let $\mathbf{Z}_{\mathcal{C}}$ and $\mathbf{Z}_{\mathcal{P}}$ collect the observed and target latent frames selected by $\mathcal{C}_i$ and $\mathcal{P}_i$, respectively (Appendix 6.3). With metadata $\boldsymbol{\eta}(\mathcal{S}) = \{(v_i, m_i, \mathbf{o}_i, \delta_i, r_i, T_i, H_i', W_i')\}_{i=1}^{N}$ and semantic context c_{sem}, the model uses a shared parameter set θ to represent the joint conditional distribution

$$p_\theta(\mathbf{Z}_{\mathcal{P}} \mid \mathbf{Z}_{\mathcal{C}}, c_{\text{sem}}, \boldsymbol{\eta}(\mathcal{S}))\,. \tag{3}$$

Here, H_i' and W_i' denote the height and width of the latent grid. At inference, observed latents and semantic context condition generation, while stream metadata specifies the target configuration; ground-truth target latents are not provided.

3.2 Unified Stream Representation

Independent visual encoding. The shared, frozen causal video encoder $\mathcal{E}$ from Wan2.2-TI2V-5B (Team Wan et al., 2025) encodes each stream independently. For input $\mathbf{x}_i \in \mathbb{R}^{3 \times \widetilde{T}_i \times H_i \times W_i}$, the latent representation is

$$\begin{aligned} &\mathbf{z}_i = \mathcal{E}(\mathbf{x}_i) \in \mathbb{R}^{C \times T_i \times H_i' \times W_i'}, \qquad C = 48, \\ &T_i = \frac{\widetilde{T}_i - 1}{4} + 1, \qquad H_i' = \frac{H_i}{16}, \qquad W_i' = \frac{W_i}{16}. \end{aligned} \tag{4}$$

Here, $\widetilde{T}_i$ denotes the number of input frames, and H_i and W_i denote the input height and width in pixels, respectively. The latent channel dimension is $C = 48$. A shared patch projection maps non-overlapping $1 \times 2 \times 2$ latent patches to tokens. Let $\bar{H}_i$ and $\bar{W}_i$ denote the spatial token-grid dimensions and L_i the stream's token count; Appendix 6.3 gives their expressions and the divisibility assumptions. For example, a 480×640 main-view clip and a 224×320 wrist-view clip can coexist in one sample. Here, resolution refers to the clip supplied to the encoder.

Virtual canvas. We assign virtual-canvas coordinates to independently encoded tokens while retaining the pretrained temporal, height, and width RoPE axes. Different viewpoints occupy disjoint spatial coordinate regions, while spatially aligned modalities reuse the coordinate origin of their corresponding RGB stream.

Each stream i has origin $\mathbf{o}_i = (o_i^h, o_i^w)$ and scalar coordinate stride $\delta_i = 1$ by default. For a modality spatially aligned with an RGB stream, let $a(i)$ denote that RGB anchor. A lower-resolution modality reuses the anchor's offset, $\mathbf{o}_i = \mathbf{o}_{a(i)}$, with stride

$$\delta_i = \text{round}\big(\bar{H}_{a(i)} / \bar{H}_i\big)\,. \tag{5}$$

The same scalar stride is used along both spatial axes; integer-grid rounding can therefore yield only approximate coordinate alignment, without upsampling the modality stream. For $f \in \mathcal{I}_i$, $0 \le p < \bar{H}_i$, and $0 \le q < \bar{W}_i$, the rotary coordinate is

$$\boldsymbol{\pi}_i(f, p, q) = \big(f, o_i^h + \delta_i p, o_i^w + \delta_i q\big). \tag{6}$$

Using f as a shared temporal coordinate requires a common temporal origin and sampling schedule across streams. The canvas coordinates organize positional encoding rather than specify calibrated camera poses.

A stream's common offset cancels from within-stream relative coordinates, while cross-stream relative coordinates retain the differences between stream origins. For a single stream, setting $\mathbf{o}_i = (0, 0)$ and $\delta_i = 1$ recovers the pretrained positional coordinates. Within the preallocated identity tables, additional viewpoints can be accommodated by modifying the input sequence and canvas layout without changing backbone parameter shapes. Appendix 6.3 compares alternative layouts in Figure 5 and discusses the offset-invariance argument.

Stream identity. Canvas coordinates specify token positions, while separate view and modality embeddings identify the visual variables they represent. We use two learnable embedding tables $\mathbf{e}_{\mathrm{V}}, \mathbf{e}_{\mathrm{M}} \in \mathbb{R}^{32\times d}$, where d is the transformer hidden dimension. For a latent input $\mathbf{y}_i$ with the same shape as $\mathbf{z}_i$, the embedded sequence is

$$\mathbf{h}_i(\mathbf{y}_i) = \mathrm{PatchEmbed}(\mathbf{y}_i) + \mathbf{e}_{\mathrm{V}}^{v_i} + \mathbf{e}_{\mathrm{M}}^{m_i} \in \mathbb{R}^{L_i\times d}. \tag{7}$$

The sum of the view and modality embeddings is broadcast across tokens and remains independent of the stream's conditioning or prediction role. Both tables are zero-initialized; thus, tokenization initially reduces to the pretrained patch projection.

Semantic context. We replace the backbone's UMT5 text encoder with a frozen Qwen3.5-9B vision–language model (Qwen Team, 2026) and a trainable adapter. The vision–language model processes the instruction together with one observed main-view RGB frame. The adapter maps its features to the backbone's cross-attention conditioning interface, producing the shared context c_{sem}. When no main-view RGB observation is available, the vision–language model receives a neutral placeholder rather than a target RGB frame, which prevents target-frame leakage through the semantic pathway.

3.3 Role-Conditioned Multi-Stream Diffusion

We initialize the shared transformer from Wan2.2-TI2V-5B and train it with flow matching (Lipman et al., 2022; Liu et al., 2022). For each training sample, we sample a shared noise level $\sigma \in [0, 1]$ and independent standard Gaussian noise tensors $\boldsymbol{\epsilon}_i$, each with the same shape as $\mathbf{z}_i$. The interpolated latent and corresponding target velocity are

$$\mathbf{z}_i^{\sigma} = (1-\sigma)\mathbf{z}_i + \sigma\boldsymbol{\epsilon}_i, \qquad \mathbf{u}_i = \boldsymbol{\epsilon}_i - \mathbf{z}_i. \tag{8}$$

The role partition specifies the input latent and normalized timestep for each latent frame $f \in \mathcal{I}_i$:

$$\widetilde{\mathbf{z}}_i[f] = \begin{cases} \mathbf{z}_i[f], & f \in \mathcal{C}_i, \\ \mathbf{z}_i^{\sigma}[f], & f \in \mathcal{P}_i, \end{cases} \qquad \boldsymbol{\tau}_i[f] = \begin{cases} 0, & f \in \mathcal{C}_i, \\ \sigma, & f \in \mathcal{P}_i. \end{cases} \tag{9}$$

Each frame's timestep is broadcast to all spatial tokens in that latent frame. Observed frames remain clean, while all target frames within a sample share the same noise level σ, with Gaussian noise sampled independently across streams.

Packed joint attention. Tokenwise timestep conditioning distinguishes observations from prediction targets within the shared transformer. A shared MLP maps each token's timestep embedding to adaptive-layer-normalization parameters. Appendix 6.3 gives the tokenwise modulation and packed sequence length. We concatenate the stream sequences $\mathbf{h}_i(\widetilde{\mathbf{z}}_i)$ along the token axis, retaining their canvas coordinates from Eq. (6). Variable-length self-attention allows all streams within a sample to interact while keeping different samples in separate attention segments. Cross-attention conditions these streams on the sample's shared semantic context. The outputs are mapped back to individual latent grids to predict the velocity $\hat{\mathbf{u}}_i$ for each stream. Changing stream roles changes the prediction task while keeping model weights fixed. For example, exchanging the clean and diffuse roles of paired RGB and depth streams reverses the prediction direction.

3.4 Training Objective and Inference

Let $\mathcal{S}^+ = \{i \in \{1, \dots, N\} : \mathcal{P}_i \neq \varnothing\}$ denote the indices of streams containing prediction targets. We average the squared velocity error over the channels and spatial elements of target frames within each stream, then over streams containing targets. Taking the expectation over training samples and noise gives

$$\mathcal{L} = \mathbb{E}_{\mathcal{S},\sigma,\boldsymbol{\epsilon}} \left[\frac{1}{|\mathcal{S}^+|} \sum_{i \in \mathcal{S}^+} \frac{1}{C H_i' W_i' |\mathcal{P}_i|} \sum_{f \in \mathcal{P}_i} \|\hat{\mathbf{u}}_{i,f} - (\boldsymbol{\epsilon}_{i,f} - \mathbf{z}_{i,f})\|_2^2 \right]. \tag{10}$$

The squared norm sums over the channels and spatial elements of each latent frame; the denominator normalizes by the number of supervised latent elements in that stream. Clean streams contribute no supervised targets, and prefix streams contribute only their target suffixes. The objective follows the same role partition used to construct the input and applies across task configurations.

We train with this shared objective on real-robot, simulated, egocentric, and general video, first on single streams and then on multiple views and modalities. Data preparation is described in Appendix 6.1, and the training recipe and implementation settings in Appendix 6.5.

Inference. In inference, the target latents are initialized with independent Gaussian noise and updated by the shared velocity model. The learned velocity field is numerically integrated from $\sigma = 1$ to $\sigma = 0$, following the noise-to-data direction of Eq. (8), while observed latents remain fixed. The generated latents and any fixed observed latents form the completed streams, which are decoded with the shared visual decoder.

4 Results

In this section, we evaluate WEAVE-0.5 on public benchmarks (Section 4.1) and present qualitative results demonstrating its multi-stream generation capabilities across views, modalities, and robot platforms (Section 4.2). Component ablations and a paired observation–depth comparison are reported in Appendices 6.4.2 and 6.4.3, respectively.

4.1 Public benchmarks

To assess the generation capability of WEAVE-0.5, we evaluate it on RBench (Deng et al., 2026) and PAI-Bench-G (Zhou et al., 2026), two public automatic benchmarks for embodied and physical video generation, following their published protocols.

RBench. RBench (Deng et al., 2026) evaluates robotic video generation across five task categories and four robot embodiments, assessing task completion, visual consistency, and physical plausibility. We report GPT-based results from the official leaderboard snapshot dated 15 September 2026, with the complete comparison provided in Tables 6 and 7 in Appendix 6.4.

WEAVE-0.5 ranks first overall with an average score of 0.642, outperforming Wan2.2-5B across all nine evaluation dimensions (Table 1). Among the listed models, it leads Spatial Relationship, Multi-entity Collaboration, and Visual Reasoning, as well as the Single-arm, Dual-arm, and Humanoid categories. Its largest advantage occurs in Single-arm evaluation, exceeding the next-best listed result by 0.089. These results indicate strengths across both task and embodiment categories, with remaining gaps in Long-horizon Planning and Quadruped generation.

PAI-Bench-G. PAI-Bench-G (Zhou et al., 2026) measures physical plausibility and domain-specific correctness through its Domain Score, and visual quality and reference fidelity through its Quality Score. Overall is the mean of these two scores. Baseline results follow the published benchmark; detailed breakdowns appear in Tables 8 and 9 in the appendix.

WEAVE-0.5 achieves the highest Overall (82.9) and Domain (88.9) scores among the generative models in Table 2. Relative to its Wan2.2-TI2V-5B backbone, it improves Overall by 1.6 points and Domain by 4.2 points, accompanied by a 0.9-point decrease in Quality. The overall improvement

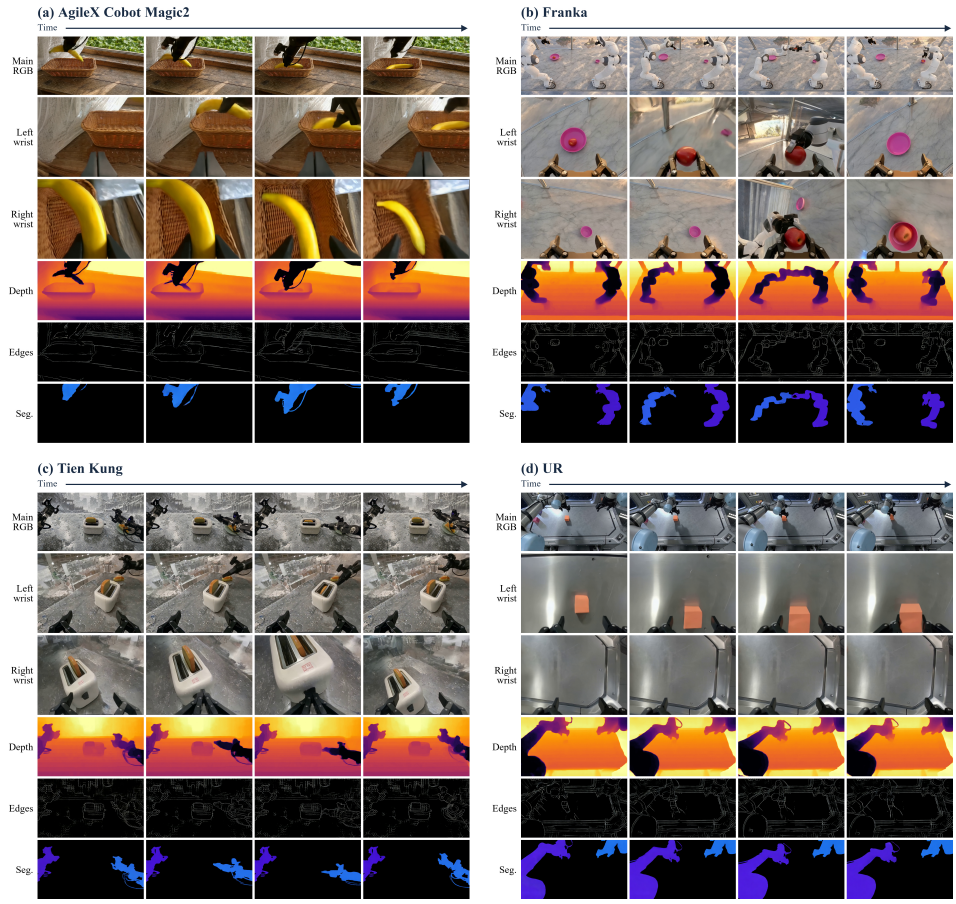

Figure 3: **Joint multi-stream generation across robot embodiments:** (a) AgileX Cobot Magic2, (b) Franka, (c) Tien Kung, and (d) UR. Given text and only the initial main-camera RGB frame, the model jointly predicts the main-view continuation and five unobserved streams. Rows show three-view RGB, followed by main-view depth, edges, and segmentation. Columns show four selected timestamps aligned across streams, with time progressing from left to right.

Table 1: **Results on RBench.** Higher scores are better. Bold and underlined values indicate the best and second-best results among the listed models, respectively. S-arm: single-arm; D-arm: dual-arm; Quad: quadruped.

		Tasks					Embodiments			
Model	Avg.	Manip	Spat	Multi	Long	Reas	S-arm	D-arm	Quad	Human
Wan2.2-A14B (Team Wan et al., 2025)	0.507	0.381	0.454	0.373	0.501	0.330	0.608	0.582	0.690	0.648
Wan2.2-5B (Team Wan et al., 2025)	0.380	0.331	0.313	0.142	0.318	0.234	0.436	0.448	0.590	0.607
LingBot-Video (Ma et al., 2026)	<u>0.620</u>	**0.578**	0.643	0.444	**0.634**	0.505	0.636	0.639	**0.758**	<u>0.689</u>
Wan 2.6 (Alibaba Cloud, 2025)	0.607	0.546	0.656	<u>0.479</u>	0.514	<u>0.531</u>	<u>0.666</u>	<u>0.681</u>	0.723	0.667
Veo 3 (Google DeepMind, 2025)	0.563	0.521	0.508	0.430	0.530	0.504	0.634	0.610	0.689	0.637
Cosmos 2.5 (Ali et al., 2025)	0.464	0.358	0.338	0.201	0.496	0.399	0.544	0.560	0.658	0.626
Cosmos 3 Nano (Agarwal et al., 2026)	0.584	0.550	<u>0.670</u>	0.466	0.576	0.394	0.591	0.611	<u>0.731</u>	0.666
Ours	**0.642**	<u>0.575</u>	**0.689**	**0.483**	<u>0.601</u>	**0.548**	**0.755**	**0.690**	0.713	**0.727**

therefore reflects stronger domain-specific performance, while preserving visual quality remains an area for improvement.

Table 2: **Results on PAI-Bench-G.** Higher scores are better. Source videos are included as a reference. Bold and underlined values indicate the best and second-best scores among the listed generative models, respectively.

Model	Overall	Domain Score	Quality Score
Source Videos	83.9	89.8	78.0
Proprietary			
Veo 3 (Google DeepMind, 2025)	82.2	86.8	77.6
Open-source			
MAGI-1-4.5B (Sand.ai et al., 2025)	77.5	78.7	76.3
HunyuanVideo-I2V (Kong et al., 2024)	77.6	77.1	**78.0**
CogVideoX1.5 (Yang et al., 2025b)	78.4	80.3	76.6
LTX-Video-13B (HaCohen et al., 2024)	78.4	79.5	77.4
MAGI-1-24B (Sand.ai et al., 2025)	79.4	82.4	76.5
Cosmos-Predict2-14B (NVIDIA, 2025)	80.0	84.3	75.8
Wan2.2-TI2V-5B (Team Wan et al., 2025)	81.3	84.7	<u>77.9</u>
Wan2.2-I2V-A14B (Team Wan et al., 2025)	<u>82.3</u>	<u>87.1</u>	77.5
Ours	**82.9**	**88.9**	77.0

4.2 Multi-Stream Generation

Figure 3 shows six-stream joint generation across four robot embodiments from a text instruction and only the initial main-camera RGB frame. The generated views depict corresponding objects and interaction stages from complementary perspectives. In (a), the banana descends toward the basket in the main view, while the wrist views reveal its position relative to the basket and gripper. In (b), the main view shows the Franka arms converging and separating, while the wrist views reveal the red object between the grippers during their interaction. In (c), the toaster interaction is visible from both the main and wrist views; in (d), the main and left-wrist views show the approach toward the orange block.

The jointly generated depth, edge, and segmentation streams show the main-view scene in different ways. In (a) and (b), the gripper movements are also visible in the depth and segmentation maps. Depth shows the overall layout of the robot and workspace, edges outline the robot and surrounding objects, and segmentation marks the robot regions as they move. Together, these examples show how the RGB views and auxiliary streams depict corresponding stages of the same interaction.

5 Conclusion and Future Directions

Conclusion. We present WEAVE-0.5, a multi-stream embodied world model that represents observations across viewpoints, resolutions, and modalities as independently encoded visual streams. A shared virtual canvas organizes token positions, while tokenwise diffusion timesteps specify observed and target content for joint prediction with one pretrained video backbone. Ablations support cross-stream attention's contribution to multi-view RGB prediction quality. The model ranks first in the reported RBench leaderboard snapshot and achieves an overall score of 82.9 on PAI-Bench-G.

Future Directions. WEAVE-0.5 represents an initial step toward multi-stream embodied world modeling, with current evaluation focused on visual observations from a main camera and two wrist cameras. Future work could extend the framework to more diverse camera configurations and sensing modalities, including point clouds and tactile signals, through suitable stream representations and aligned training data. Another direction is to systematically investigate how scaling training data and model capacity affects multi-stream prediction and generalization to unseen camera and modality configurations.

AI USE STATEMENT

The manuscript builds on material written and checked by the authors. Generative AI tools assisted with condensing and revising this material, translation for internal manuscript review, LaTeX and table formatting, reference and cross-reference checks, and the preparation and revision of schematic figures. AI assistance was also used to draft selected equation expressions and mathematical explanations, and to review the consistency between experimental descriptions, tables, and interpretations. AI tools also assisted with the design of supplementary controlled ablations, the review and debugging of evaluation scripts, statistical checks, and the interpretation of results.

Separately, learned models are components of the research pipeline: model-assisted annotation and dense-label generation prepare training streams (Appendix 6.1), a frozen vision–language encoder provides semantic conditioning (Section 3.2), and vision–language judges are used in public-benchmark evaluation (Section 4.1). These uses are components of the method and evaluation protocols, distinct from the general-purpose AI assistance described above.

The authors retain responsibility for verifying AI-assisted text, equations, references, figures, and any code against the underlying method and experimental records. AI assistance does not establish correctness; the authors take responsibility for the final content, claims, and artifacts of this work.

ETHICS STATEMENT

This work studies visual generation and prediction with WEAVE-0.5; it does not validate a robot-control policy. Plausible videos and high benchmark scores do not guarantee geometric or physical correctness or safe behavior. Any downstream use in robot planning or control requires separate safety evaluation, appropriate safeguards, and human oversight; the reported results should not be treated as evidence of safe deployment.

Training draws on robot, egocentric, simulated, and general video. Such data can contain identifiable people or environments and can underrepresent particular embodiments, tasks, or settings. Model-derived annotations and visual labels, as well as vision–language evaluation judges, can introduce additional errors and biases. Reuse of data and pretrained models, and any subsequent release of derived artifacts, must respect source-specific terms, access restrictions, and applicable privacy requirements. Some training sources cannot be redistributed; their inclusion does not grant permission to release the underlying recordings. These limitations warrant caution when interpreting or transferring the model's predictions.

REPRODUCIBILITY STATEMENT

Section 3 defines the stream representation, virtual-canvas coordinates, role-conditioned diffusion, training objective, and inference procedure. Appendix 6.5 reports the staged training recipe and sampling configuration, while Appendices 6.1 and 6.5 describe the data sources, preparation pipeline, and implementation details, including a late-phase optimization configuration. The offset-invariance argument is given in Appendix 6.3. Section 4.1 describes the public benchmarks and presents the comparison tables. For the ablations, Appendix 6.4.2 reports the component comparisons, conditioning settings, and per-view results; Appendix 6.4.3 provides the checkpoint, sampling, and evaluation details for the paired comparison. These descriptions distinguish inference-time interventions, continued-training variants, and a paired control of RGB observations and joint depth generation. The latter specifies the common future-frame mask, paired sample size, metric aggregation, and bootstrap intervals separately from the original component-scoring protocol. Exact retraining remains constrained by training-data access: some sources cannot be redistributed. Omitting those sources changes the effective training distribution, so a reproduction using only accessible sources should not be assumed to match the reported results exactly.

REFERENCES

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

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

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

Alibaba Cloud. Alibaba Unveils Wan2.6 Series Enabling Everyone to Star in Videos. Alibaba Cloud Community. https://www.alibabacloud.com/blog/alibaba-unveils-wan2-6-series-enabling-everyone-to-star-in-videos_602742, 2025. Published 16 December 2025.

Alibaba Cloud. Video generation and editing. Alibaba Cloud Model Studio documentation. https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model, 2026. Section: Wan 2.5. Updated 2 September 2026.

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 (NeurIPS)*, volume 37, pp. 58757–58791, 2024.

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

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

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning (ICML)*, pp. 1692–1717. PMLR, 2023.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35101–35113, 2026.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video Generation Models as World Simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

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

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Boyuan Chen, Diego Martí Monsó, 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 (NeurIPS)*, volume 37, pp. 24081–24125, 2024a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length Film Generative Model. *arXiv preprint arXiv:2504.13074*, 2025a. URL https://arxiv.org/abs/2504.13074.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025b.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024b.

Jen-Hao Cheng, Yipeng Wang, Hao Zhang, Gengshan Yang, and Jenq-Neng Hwang. Flex4dhuman: Flexible multi-view video diffusion for 4d human reconstruction. *arXiv preprint arXiv:2606.13655*, 2026.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking Video Generation Model for the Embodied World. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2601.15282. Official leaderboard: https://huggingface.co/spaces/DAGroup-PKU/RBench-Leaderboard. Snapshot: 15 September 2026.

DreamX Team, Rui Chen, Xiangxiang Chu, Geng Li, Jifan Li, Qingfeng Shi, Datao Tang, Jing Tang, Jun Wang, and Pengfei Zhang. Dreamx-phi 1.0: Action-conditioned video world model for robotic manipulation. *arXiv preprint arXiv:2608.13489*, 2026.

Xiang Fan, Sharath Girish, Vivek Ramanujan, Chaoyang Wang, Ashkan Mirzaei, Petr Sushko, Aliaksandr Siarohin, Sergey Tulyakov, and Ranjay Krishna. OmniView: An All-Seeing Diffusion Model for 3D and 4D View Synthesis. *arXiv preprint arXiv:2512.10940*, 2026. URL https://arxiv.org/abs/2512.10940.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the Boundaries of Video Generation Models. *arXiv preprint arXiv:2506.09113*, 2025. URL https://arxiv.org/abs/2506.09113.

Google DeepMind. Veo: A Text-to-Video Generation System. Technical report, Google DeepMind, 2025. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022.

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. In *European Conference on Computer Vision (ECCV)*, 2026a.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2026b.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. DOI: https://doi.org/10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *International Conference on Learning Representations (ICLR)*, 2024.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.

Jiehui Huang, Yuechen Zhang, Xu He, Yuan Gao, Zhi Cen, Bin Xia, Yan Zhou, Xin Tao, Pengfei Wan, and Jiaya Jia. 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.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. In *Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194, 2025. URL https://proceedings.mlr.press/v305/jang25a.html.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 10 2025b.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Kuaishou Technology. Kling AI Launches Video 2.6 Model with "Simultaneous Audio-Visual Generation" Capability, Redefining AI Video Creation Workflow. Official announcement. https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-video-26-model-simultaneous-audio-visual, 2025. Published 5 December 2025.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13178–13188. IEEE, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiagang Zhu, Jiaxiong Qiu, Guan Huang, and Zhizhong Su. RoboTransfer: Controllable geometry-consistent video diffusion for manipulation policy transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1410–1420, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, et al. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, et al. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. LongCat-Video Technical Report. *arXiv preprint arXiv:2510.22200*, 2025. URL https://arxiv.org/abs/2510.22200.

MiniMax. MiniMax Hailuo 02, World-Class Quality, Record-Breaking Cost Efficiency. MiniMax News. https://www.minimax.io/news/minimax-hailuo-02, 2025. Published 18 June 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

NVIDIA. Cosmos-Predict2: World Simulation Model for Physical AI. GitHub repository. https://github.com/nvidia-cosmos/cosmos-predict2, 2025.

OpenAI. Sora 2 System Card, 2025. URL https://openai.com/index/sora-2-system-card/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

RealSource. RealSource World: A Large-Scale Real-World Dual-Arm Manipulation Dataset. Hugging Face dataset. https://huggingface.co/datasets/RealSourceData/RealSource-World, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, et al. MAGI-1: Autoregressive Video Generation at Scale. *arXiv preprint arXiv:2505.13211*, 2025. URL https://arxiv.org/abs/2505.13211.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

AgiBot World Team. Agibot world 2026. https://huggingface.co/datasets/agibot-world/AgiBotWorld2026, 2026a.

World Labs Team. Atlas: A world model for spatial intelligence. *World Labs Blog*, 2026b. https://www.worldlabs.ai/blog/atlas.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A Native Audio-Visual Joint Generation Foundation Model. *arXiv preprint arXiv:2512.13507*, 2025. URL https://arxiv.org/abs/2512.13507.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 976–985, 2026.

Unitree. UnifoLM-WMA-0: A World-Model-Action (WMA) Framework under UnifoLM Family, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42592–42603, 2026a.

Jiaxu Wang, Yicheng Jiang, Tianlun He, Jingkai Sun, Qiang Zhang, Junhao He, Jiahang Cao, Zesen Gan, Mingyuan Sun, Qiming Shao, et al. Mvista-4d: View-consistent 4d world model with test-time action inference for robotic manipulation. *arXiv preprint arXiv:2602.09878*, 2026b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/tip.2003.819861. DOI: https://doi.org/10.1109/tip.2003.819861.

Liangjian Wen, Linjie Li, Jiang Duan, Yong Dai, Jianzhuang Liu, and Zhao Kang. Dependency, compression, and synergy: A unified information-theoretic view of multimodal learning. *arXiv preprint arXiv:2609.14421*, 2026. URL https://arxiv.org/pdf/2609.14421v1.

World Labs. Marble: A Multimodal World Model. World Labs Blog. https://www.worldlabs.ai/blog/marble-world-model, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 Technical Report. *arXiv preprint arXiv:2511.18870*, 2025a. URL https://arxiv.org/abs/2511.18870.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 68082–68119, 2024.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18682–18692. IEEE, 2025b.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025c.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. *arXiv preprint arXiv:2310.12190*, 2023.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025a.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Mingqiao Ye, Zhaochong An, Zhitong Gao, Xian Liu, François Fleuret, Chuan Li, Amir Zadeh, Serge Belongie, Afshin Dehghan, Jesse Allardice, et al. Modus: Decoder-only any-to-any modeling of diverse modalities. *arXiv preprint arXiv:2607.25948*, 2026.

Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, et al. Genie sim 3.0: A high-fidelity comprehensive simulation platform for humanoid robot. *arXiv preprint arXiv:2601.02078*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 6 2025.

He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, et al. Hy-embodied-0.5-vla: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026a.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-RobotWorld Technical Report: Unifying Embodied World Modeling through Language-Conditioned Video Generation. *arXiv preprint arXiv:2606.17030*, 2026b. URL https://arxiv.org/abs/2606.17030.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 10 2023.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 30546–30566, 2025. doi: 10.52202/085713-1024. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2bde8fef08f7ebe42b584266cbcfc909-Paper-Conference.pdf.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595. IEEE, 2018.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A Comprehensive Benchmark For Physical AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21522–21536, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_PAI-Bench_A_Comprehensive_Benchmark_For_Physical_AI_CVPR_2026_paper.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/2411.04983v1.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

6 APPENDIX

6.1 DATA OVERVIEW

WEAVE-0.5 is jointly trained on real-robot, simulated, egocentric, and general video. Figure 4 summarizes the collected corpora and visual-stream preparation pipeline. Table 3 lists the corresponding sources and references. The training and sampling schedule is described in Appendix 6.5.

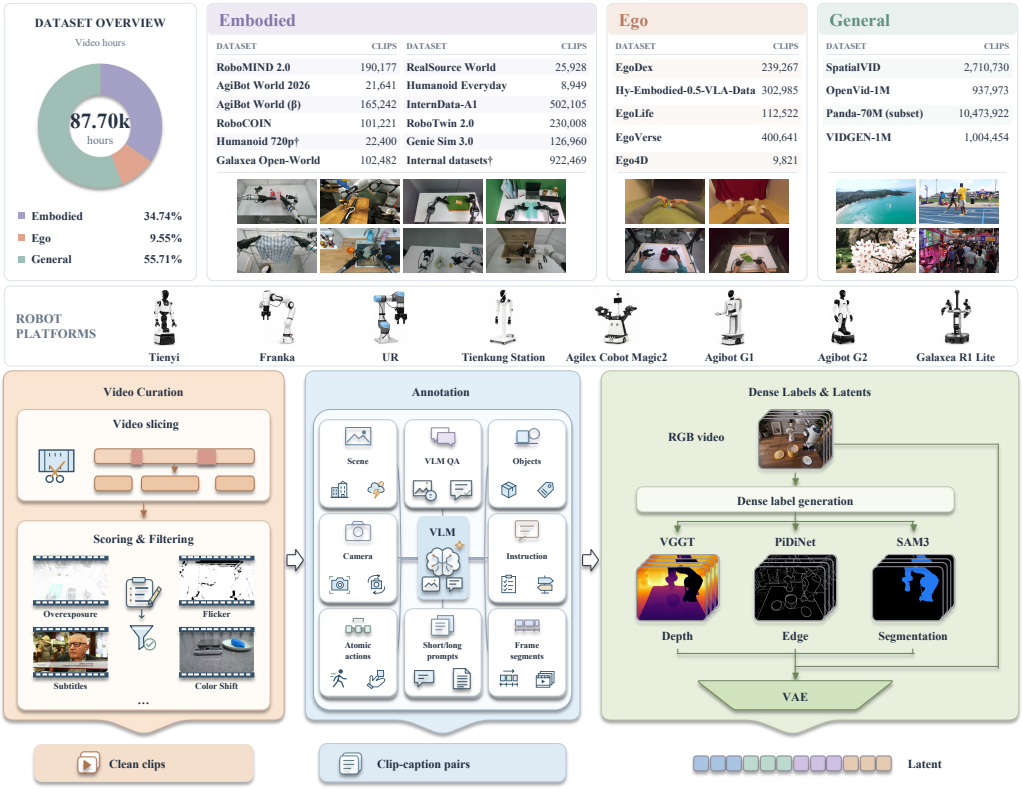

Figure 4: Overview of the collected corpora and visual-data preparation. Top: embodied, egocentric, and general video sources. Video-hour proportions describe corpus composition, not the training sampling weights in Table 14. † marks internally collected datasets, which cannot be redistributed. Bottom: video slicing, scoring, and filtering produce clean clips. A shared vision–language tool generates category-dependent structured annotations. Depth, edge, and segmentation labels are estimated from RGB where available; RGB and dense-label streams are encoded by a frozen VAE into cached latents.

Corpora and inventory. We distinguish what was collected from what the sampler draws. Clip counts in Figure 4 are recorded before per-source filtering and can exceed distinct-episode counts because a long recording may be sliced into several clips. Some public sources are used as selected subsets rather than full releases, including Panda-70M; their collected durations can therefore differ from the totals reported by the providers. The figure reports the collected inventory, while stage-dependent sampling weights are given in Table 14.

Clip construction. Processing depends on whether a source contains untrimmed web videos or pre-segmented robot and egocentric episodes. For web video, shot detection proposes boundaries, and candidate clips are assessed for visual quality, temporal continuity, and semantic coherence. Robot and head-worn episodes provide initial clip boundaries and undergo quality filtering and input normalization. Only temporally aligned camera recordings are assembled into multi-view samples; otherwise, the source contributes single-view samples.

Table 3: **Data sources.** Sources and types corresponding to Figure 4. † marks internally collected datasets, which cannot be redistributed.

Source	Type
Embodied: real-robot	
RoboMIND 2.0 (Hou et al., 2025)	multi-embodiment
AgiBot World 2026 (Team, 2026a)	humanoid
AgiBot World (β) (Bu et al., 2025)	bimanual mobile
RoboCOIN (Wu et al., 2025c)	bimanual, multi-embodiment
Humanoid 720p†	humanoid
Galaxea Open-World (Jiang et al., 2025a)	mobile manipulation
RealSource World (RealSource, 2025)	dual-arm humanoid
Humanoid Everyday (Zhao et al., 2025)	humanoid
Embodied: simulation	
InternData-A1 (Tian et al., 2026)	simulation
RoboTwin 2.0 (Chen et al., 2025b)	simulation
Genie Sim 3.0 (Yin et al., 2026)	simulation
Egocentric: head-worn	
EgoDex (Hoque et al., 2026)	head-worn, dexterous hands
Hy-Embodied-0.5-VLA-Data (Zhang et al., 2026a)	head-worn, UMI-style bimanual
EgoLife (Yang et al., 2025a)	head-worn, long-form
EgoVerse (Punamiya et al., 2026)	head-worn, multi-device
Ego4D (Grauman et al., 2022)	head-worn, daily activity
Internal	
Internal datasets†	mixed
General web video	
SpatialVID (Wang et al., 2026a)	web video, camera pose
OpenVid-1M (Nan et al., 2025)	web video, captioned
Panda-70M (subset) (Chen et al., 2024b)	web video, selected subset
VIDGEN-1M (Tan et al., 2024)	web video, captioned

Annotation and visual streams. The shared annotation tool produces structured captions containing instructions, short and long prompt variants, scene and object descriptions, and camera metadata, with fields varying by video category. Semantic conditioning uses the instruction and observation interface in Section 3.2; prompt construction is described in Appendix 6.5.2. Where available, depth, edge, and segmentation labels are estimated from RGB rather than supplied by a second sensor. RGB and dense-label videos are independently encoded into cached latent streams before training by the shared, frozen causal autoencoder, at the resolutions specified in Appendix 6.5.3.

6.2 Capability coverage

Table 4 summarizes reported visual capabilities of selected related models. The entries describe capability coverage rather than benchmark performance; an unreported capability is not evidence that the architecture cannot support it. The comparison is not exhaustive. MVISTA-4D (Wang et al., 2026b) also supports variable view counts and masked view completion, while OmniView (Fan et al., 2026) jointly represents clean observations and noisy targets across views and time (Section 2.2).

Table 4: Reported visual capability coverage of selected embodied world models. ✓ indicates reported support, ○ indicates restricted support, and × indicates that support has not been described or demonstrated in the cited work. Model references are Cosmos 3 (Agarwal et al., 2026), LingBot-Video (Ma et al., 2026), Ctrl-World (Guo et al., 2026b), DreamX-Phi (DreamX Team et al., 2026), and RoboTransfer (Liu et al., 2026). In the task labels, V denotes RGB video and C denotes depth, edge, or segmentation streams; support for one does not imply support for all three. DreamX-Phi's restricted support uses an auxiliary depth-prediction branch rather than interchangeable RGB/depth conditioning roles.

Task Type	Cosmos 3	LingBot	Ctrl-World	DreamX-Phi	RoboTransfer	**WEAVE-0.5**
Single-view generation	✓	✓	✓	✓	✓	✓
Multi-view generation	✓	×	✓	○	✓	✓
Variable view count	×	×	×	×	×	✓
Cross-view completion	×	×	×	×	×	✓
Modality pred. (V→C)	×	×	×	○	×	✓
Modality-cond. gen. (C→V)	✓	×	×	×	✓	✓
Joint RGB + modality	×	×	×	○	×	✓

6.3 METHOD DETAILS

Observed and target latent collections. For any index set $\mathcal{A} \subseteq \mathcal{I}_i$, let $\mathbf{z}_i[\mathcal{A}]$ denote the corresponding subsequence of latent frames in temporal order. The observed and target latent collections are defined as

$$\mathbf{Z}_{\mathcal{C}} = \{\mathbf{z}_i[\mathcal{C}_i]\}_{i=1}^{N}, \qquad \mathbf{Z}_{\mathcal{P}} = \{\mathbf{z}_i[\mathcal{P}_i]\}_{i=1}^{N}. \tag{11}$$

These collections enter the conditional distribution in Eq. (3).

Token grid and stream length. After the patch projection in Section 3.2, the resulting spatial token grid has dimensions $\bar{H}_i = H_i'/2 = H_i/32$ and $\bar{W}_i = W_i'/2 = W_i/32$, yielding a sequence of length

$$L_i = T_i \bar{H}_i \bar{W}_i. \tag{12}$$

These expressions assume that $\widetilde{T}_i - 1$ is divisible by 4 and that H_i, W_i are divisible by 32.

Tokenwise conditioning and sequence packing. For token ℓ, a shared MLP maps its timestep embedding to the shift, scale, and gate parameters used by adaptive layer normalization:

$$\mathbf{m}_\ell = \mathrm{MLP}\big(\mathrm{emb}(\tau_\ell)\big) \in \mathbb{R}^{6\times d}. \tag{13}$$

Here, emb denotes the timestep embedding function. For sample n with N_n streams, let $L_i^{(n)}$ denote the token count of stream i. The packed sequence length is

$$L^{(n)} = \sum_{i=1}^{N_n} L_i^{(n)}. \tag{14}$$

Analytic checks rather than trained ablations. The role partition and canvas stride follow directly from their definitions. The `clean` role holds an entire stream at timestep zero for modality prediction or modality-conditioned generation. Eq. (5) assigns stride 1 to an aligned modality at its RGB anchor's resolution and stride 2 when its token grid is half the anchor's size along both spatial axes. These are interface properties, not measured prediction-quality gains.

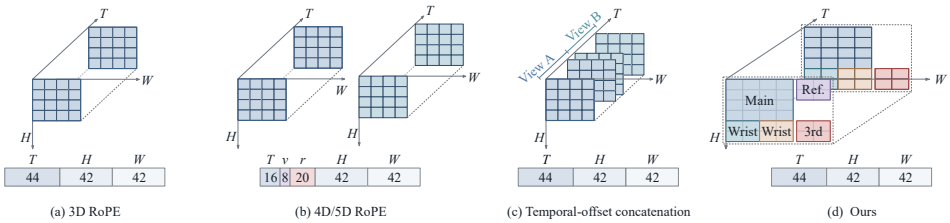

Figure 5: **Multi-stream positional layouts.** The bars denote allocations of a 128-channel head, not sequence lengths. (a) Wan's $44/42/42$ split over (t, h, w). (b) Flex4DHuman's five-band design assigns $16/8/20/42/42$ channels to time, view, SE(3) geometry, height, and width (Cheng et al., 2026). (c) Frame-wise concatenation assigns consecutive temporal coordinates to different views while retaining the $44/42/42$ split, as in the public Wan-based implementation of ReCamMaster (Bai et al., 2025a). (d) Our canvas shares time and separates views through spatial offsets, preserving the pretrained partition.

Four layouts for multi-stream RoPE. Figure 5 and Table 5 distinguish channel allocation from token coordinates. Flex4DHuman's five-band design partitions a 128-channel head into 16 time, 8 view, 20 SE(3) geometry, and 42 channels for each spatial axis (Cheng et al., 2026). The time, view, and geometry bands occupy the original 44 temporal channels; this is a specific design, not a requirement of every multi-view encoding. ReCamMaster instead concatenates source and target videos along the frame dimension (Bai et al., 2025a). Its public Wan-based implementation assigns consecutive temporal coordinates across the concatenated sequence without changing the $44/42/42$ channel allocation. Temporally aligned frames from different views therefore receive different positional timestamps; this does not imply that the generated videos cannot be synchronized.

Table 5 also includes a conceptual embedding-only layout that reuses positional coordinates across views. It is not a characterization of SynCamMaster's complete architecture: that method combines camera embeddings with cross-view synchronization attention (Bai et al., 2025b). In an image-domain study, removing the offset reduces DINO subject similarity from 0.542 to 0.386, a relative drop of approximately 29% (Wu et al., 2025b). This is contextual evidence rather than an ablation of WEAVE-0.5. Our formulation combines spatial offsets with identity embeddings (Section 3.2) while preserving the rotary axes and their partition. Proposition 1 establishes invariance to a common offset within a single stream, not equivalence of the full trained model to the pretrained backbone.

Table 5: **Positional-layout properties relative to the pretrained layout (a).** (b) Flex4DHuman's five-band design; (c) concatenation in time; (e) a conceptual embedding-only layout, not drawn in Figure 5; (d) virtual canvas. This is a comparison of encoding properties, not a prediction-quality ablation.

vs. pretrained (a)	(b) five bands	(c) time concat.	(e) embed. only	(d) **canvas**
Rotary partition retained	×	✓	✓	✓
Temporal band preserved	× (44 → 16)	✓	✓	✓
Single-stream RoPE retained	×	✓	✓	✓
Shared time across views	✓	×	✓	✓
Distinct view coordinates	✓	✓	×	✓
View-position encoding	view/geometry	time offsets	embeddings	spatial offsets

Offset invariance. A common spatial offset cancels from within-stream rotary attention.

Proposition 1 (Single-stream offset invariance)**.** *Let a sample consist of a single stream with offset* $\mathbf{o}$ *and stride* δ*. Holding the token features, stride, rotary frequencies, and all other inputs fixed, every rotary self-attention logit is invariant to the common offset* $\mathbf{o}$*.*

Proof. Rotary embeddings act on queries and keys by a block-diagonal rotation $R(\cdot)$ satisfying the homomorphism property $R(\mathbf{p})^\top R(\mathbf{q}) = R(\mathbf{q} - \mathbf{p})$, so the logit between a query at canvas position $\mathbf{p}$ and a key at $\mathbf{q}$ is $\mathbf{u}^\top R(\mathbf{q} - \mathbf{p})\mathbf{v}$ and depends on $\mathbf{q} - \mathbf{p}$ alone. By Eq. (6), two tokens at local

indices (t, a, b) and (t', a', b') of the same stream have canvas coordinates $\big(t,\ o^h + \delta a,\ o^w + \delta b\big)$ and $\big(t',\ o^h + \delta a',\ o^w + \delta b'\big)$, hence coordinate difference $\big(t' - t,\ \delta(a' - a),\ \delta(b' - b)\big)$, in which $\mathbf{o}$ cancels. Although the offset changes the absolute coordinates used to construct the rotations, their product depends only on the relative coordinate. All self-attention pairs of a single-stream sample lie within that stream, so the common offset does not change these logits.

Single-view episodes may therefore be placed at the origin without changing their within-stream rotary logits. This property does not establish that mixing single-view and multi-view samples introduces no distribution shift, nor that changing other model inputs leaves the output unchanged. For $i \neq j$, the coordinate difference retains the difference between $\mathbf{o}_i$ and $\mathbf{o}_j$; this residual term encodes the assigned cross-view layout.

Visual-stream admissibility. The interface accepts a signal when it can be represented as a visual stream in the training pipeline. Such a stream uses the same encoder, canvas coordinates, role assignment, and loss; this is an interface property, not a claim that every possible signal was evaluated.

6.4 Additional Experimental Results

6.4.1 Public benchmark comparisons

Tables 6 and 7 record the task and embodiment dimensions of the same full RBench leaderboard snapshot used for the comparison in Section 4.1. Higher scores are better; bold and underlined values indicate the best and second-best scores within each table, respectively. The full PAI-Bench-G comparison appears in Tables 8 and 9 below.

Table 6: **RBench leaderboard snapshot: task dimensions.** All 30 entries from the 15 Sep. 2026 official snapshot (Deng et al., 2026), using the GPT evaluator. Values are the five task-category scores; Avg. is the overall score reported in that snapshot. Manip.: manipulation; Spat.: spatial; Multi.: multi-entity; Long.: long-horizon; Reas.: visual reasoning. The complementary embodiment scores are in Table 7. "#" marks the reported API limitation under which approximately 50 of 650 videos could not be generated.

			Task dimensions				
Model	Rank	Avg.	Manip.	Spat.	Multi.	Long.	Reas.
Open-source							
LingBot-Video (Ma et al., 2026)	2	<u>0.620</u>	**0.578**	0.643	0.444	**0.634**	0.505
Cosmos 3 Nano (Agarwal et al., 2026)	4	0.584	0.550	<u>0.670</u>	0.466	0.576	0.394
Cosmos 3 Super (Agarwal et al., 2026)	6	0.581	0.487	0.642	0.444	0.591	0.395
Cosmos 3 edge (Agarwal et al., 2026)	10	0.551	0.503	0.540	0.313	0.523	0.393
Wan2.2-A14B (Team Wan et al., 2025)	13	0.507	0.381	0.454	0.373	0.501	0.330
HunyuanVideo 1.5 (Wu et al., 2025a)	15	0.460	0.442	0.316	0.312	0.438	0.364
LongCat-Video (Meituan LongCat Team et al., 2025)	16	0.437	0.372	0.310	0.220	0.384	0.186
Wan2.1-14B (Team Wan et al., 2025)	19	0.399	0.344	0.268	0.282	0.335	0.205
LTX-2 (HaCohen et al., 2026)	20	0.381	0.284	0.304	0.233	0.386	0.164
Wan2.2-5B (Team Wan et al., 2025)	21	0.380	0.331	0.313	0.142	0.318	0.234
SkyReels (Chen et al., 2025a)	23	0.361	0.203	0.276	0.203	0.254	0.234
LTX-Video (HaCohen et al., 2024)	24	0.344	0.302	0.176	0.210	0.280	0.241
FramePack (Zhang et al., 2025)	25	0.339	0.206	0.258	0.173	0.169	0.170
HunyuanVideo (Kong et al., 2024)	26	0.303	0.177	0.180	0.108	0.147	0.035
CogVideoX-5B (Yang et al., 2025b)	28	0.256	0.116	0.112	0.098	0.212	0.079
Commercial							
Wan 2.6 (Alibaba Cloud, 2025)	3	0.607	0.546	0.656	0.479	0.514	<u>0.531</u>
Seedance 1.5 pro (Team Seedance et al., 2025)	4	0.584	<u>0.577</u>	0.495	**0.484**	0.570	0.470
Wan 2.5 (Alibaba Cloud, 2026)	7	0.570	0.527	0.576	0.402	0.496	0.437
Hailuo v2 (MiniMax, 2025)	8	0.565	0.560	0.637	0.386	0.545	0.474
Veo 3 (Google DeepMind, 2025)	9	0.563	0.521	0.508	0.430	0.530	0.504
Seedance 1.0 (Gao et al., 2025)	10	0.551	0.542	0.425	0.448	0.454	0.442
Kling 2.6 pro (Kuaishou Technology, 2025)	12	0.534	0.529	0.598	0.364	0.530	0.358
Sora v2 Pro# (OpenAI, 2025)	22	0.362	0.208	0.268	0.186	0.255	0.115
Sora v1 (Brooks et al., 2024)	27	0.266	0.151	0.223	0.111	0.166	0.139
Robotics-specific							
Cosmos 2.5 (Ali et al., 2025)	14	0.464	0.358	0.338	0.201	0.496	0.399
DreamGen (gr1) (Jang et al., 2025)	17	0.420	0.312	0.372	0.297	0.334	0.215
DreamGen (droid) (Jang et al., 2025)	18	0.405	0.358	0.348	0.214	0.316	0.339
Vidar (Feng et al., 2025)	29	0.206	0.073	0.106	0.050	0.054	0.050
UnifoLM-WMA-0 (Unitree, 2025)	30	0.123	0.036	0.040	0.018	0.062	0.000
Ours	1	**0.642**	0.575	**0.689**	<u>0.483</u>	<u>0.601</u>	**0.548**

Table 7: **RBench leaderboard snapshot: embodiment dimensions.** The same 30-entry snapshot and ordering as Table 6; these columns report the four embodiment scores. S-arm: single-arm; D-arm: dual-arm; Quad.: quadruped robot; Human.: humanoid robot. "#" marks the reported API limitation under which approximately 50 of 650 videos could not be generated.

		Embodiment dimensions			
Model	Rank	S-arm	D-arm	Quad.	Human.
Open-source					
LingBot-Video (Ma et al., 2026)	2	0.636	0.639	**0.758**	0.689
Cosmos 3 Nano (Agarwal et al., 2026)	4	0.591	0.611	0.731	0.666
Cosmos 3 Super (Agarwal et al., 2026)	6	0.615	0.623	0.739	0.691
Cosmos 3 edge (Agarwal et al., 2026)	10	0.639	0.620	<u>0.744</u>	0.683
Wan2.2-A14B (Team Wan et al., 2025)	13	0.608	0.582	0.690	0.648
HunyuanVideo 1.5 (Wu et al., 2025a)	15	0.513	0.526	0.634	0.595
LongCat-Video (Meituan LongCat Team et al., 2025)	16	0.586	0.576	0.681	0.621
Wan2.1-14B (Team Wan et al., 2025)	19	0.464	0.497	0.595	0.599
LTX-2 (HaCohen et al., 2026)	20	0.453	0.424	0.622	0.555
Wan2.2-5B (Team Wan et al., 2025)	21	0.436	0.448	0.590	0.607
SkyReels (Chen et al., 2025a)	23	0.507	0.477	0.586	0.509
LTX-Video (HaCohen et al., 2024)	24	0.440	0.456	0.526	0.464
FramePack (Zhang et al., 2025)	25	0.440	0.464	0.626	0.548
HunyuanVideo (Kong et al., 2024)	26	0.454	0.480	0.625	0.524
CogVideoX-5B (Yang et al., 2025b)	28	0.338	0.385	0.465	0.496
Commercial					
Wan 2.6 (Alibaba Cloud, 2025)	3	0.666	<u>0.681</u>	0.723	0.667
Seedance 1.5 pro (Team Seedance et al., 2025)	4	0.648	0.641	0.680	<u>0.692</u>
Wan 2.5 (Alibaba Cloud, 2026)	7	<u>0.680</u>	0.634	0.726	0.654
Hailuo v2 (MiniMax, 2025)	8	0.594	0.611	0.640	0.635
Veo 3 (Google DeepMind, 2025)	9	0.634	0.610	0.689	0.637
Seedance 1.0 (Gao et al., 2025)	10	0.622	0.641	0.698	0.686
Kling 2.6 pro (Kuaishou Technology, 2025)	12	0.570	0.605	0.637	0.613
Sora v2 Pro# (OpenAI, 2025)	22	0.476	0.513	0.664	0.561
Sora v1 (Brooks et al., 2024)	27	0.314	0.324	0.544	0.419
Robotics-specific					
Cosmos 2.5 (Ali et al., 2025)	14	0.544	0.560	0.658	0.626
DreamGen (gr1) (Jang et al., 2025)	17	0.564	0.532	0.579	0.575
DreamGen (droid) (Jang et al., 2025)	18	0.499	0.476	0.542	0.556
Vidar (Feng et al., 2025)	29	0.382	0.410	0.374	0.357
UnifoLM-WMA-0 (Unitree, 2025)	30	0.268	0.194	0.293	0.200
Ours	1	**0.755**	**0.690**	0.713	**0.727**

Table 8: **Detailed PAI-Bench-G results: overall and domain scores.** The compact subset is in Table 2; complementary quality scores are in Table 9. CS: Common Sense; AV: Autonomous Vehicle; RO: Robot; IN: Industry; HU: Human; PH: Physics. Overall is the mean of Domain and Quality. Baseline values are reproduced from the published benchmark results (Zhou et al., 2026); source videos are a reference and are not ranked as a model.

Model	Overall	CS	AV	RO	IN	HU	PH	Avg.
Source Videos	83.9	96.6	78.7	91.8	89.0	86.6	93.1	89.8
Proprietary								
Veo 3 (Google DeepMind, 2025)	82.2	93.0	72.1	88.8	89.0	84.4	89.7	86.8
Open-source								
DynamiCrafter (Xing et al., 2023)	68.3	73.9	43.9	48.9	70.9	60.7	80.8	63.0
LTX-Video-2B (HaCohen et al., 2024)	77.5	88.1	60.5	69.4	82.6	77.1	86.2	78.0
MAGI-1-4.5B (Sand.ai et al., 2025)	77.5	88.3	63.4	71.3	82.5	77.3	86.3	78.7
HunyuanVideo-I2V (Kong et al., 2024)	77.6	86.7	60.3	68.9	82.4	75.8	85.8	77.1
CogVideoX1.5 (Yang et al., 2025b)	78.4	88.7	64.4	73.6	84.6	79.1	88.7	80.3
LTX-Video-13B (HaCohen et al., 2024)	78.4	88.6	60.0	71.8	84.6	79.1	88.8	79.5
CogVideoX (Yang et al., 2025b)	78.6	88.4	62.7	75.3	85.0	81.0	88.5	80.9
MAGI-1-24B (Sand.ai et al., 2025)	79.4	91.4	68.0	76.5	85.4	81.5	87.0	82.4
Cosmos-Predict2-2B (NVIDIA, 2025)	79.7	91.2	70.7	81.9	86.4	83.1	88.1	84.2
Cosmos-Predict2-14B (NVIDIA, 2025)	80.0	92.5	69.8	81.3	86.6	82.4	89.2	84.3
Wan2.1-I2V-14B (Team Wan et al., 2025)	80.8	92.5	69.4	79.6	87.2	83.4	89.5	84.3
Wan2.2-TI2V-5B (Team Wan et al., 2025)	81.3	92.2	69.1	82.9	86.4	83.3	90.3	84.7
Cosmos-Predict2.5-2B (Ali et al., 2025)	81.4	91.5	70.4	82.7	86.8	83.5	91.7	84.9
Wan2.2-I2V-A14B (Team Wan et al., 2025)	82.3	94.1	73.0	86.8	88.6	84.5	92.7	87.1
Ours	**82.9**	**95.0**	**75.5**	**93.0**	**89.4**	**85.4**	**94.9**	**88.9**

Table 9: **Detailed PAI-Bench-G results: quality scores.** The same models and row order as Table 8. SC/BC: Subject/Background Consistency; MS: Motion Smoothness; AQ: Aesthetic Quality; IQ: Imaging Quality; OC: Overall Consistency; IS/IB: image-to-video Subject/Background fidelity. Avg. averages the eight quality sub-metrics. Baseline values are reproduced from the published benchmark results (Zhou et al., 2026); source videos are a reference and are not ranked as a model.

Model	SC	BC	MS	AQ	IQ	OC	IS	IB	Avg.
Source Videos	93.3	94.2	99.1	51.7	68.4	21.5	97.8	98.2	78.0
Proprietary									
Veo 3 (Google DeepMind, 2025)	91.4	93.1	99.2	51.9	69.8	**21.7**	97.0	96.9	77.6
Open-source									
DynamiCrafter (Xing et al., 2023)	91.1	92.5	94.9	51.5	68.0	21.2	84.5	86.2	73.7
LTX-Video-2B (HaCohen et al., 2024)	89.2	92.7	98.7	53.2	**71.3**	21.1	95.0	95.9	77.1
MAGI-1-4.5B (Sand.ai et al., 2025)	92.1	93.3	99.0	50.4	61.8	21.6	94.5	**98.1**	76.3
HunyuanVideo-I2V (Kong et al., 2024)	**94.3**	**95.3**	**99.5**	52.1	65.2	21.5	**98.6**	97.6	**78.0**
CogVideoX1.5 (Yang et al., 2025b)	91.6	93.9	98.5	50.0	66.5	21.2	95.0	96.1	76.6
LTX-Video-13B (HaCohen et al., 2024)	90.6	93.5	99.0	**53.5**	69.5	21.4	95.7	96.0	77.4
CogVideoX (Yang et al., 2025b)	91.4	93.4	98.0	51.2	64.6	21.3	94.1	95.9	76.3
MAGI-1-24B (Sand.ai et al., 2025)	90.0	92.4	99.0	50.2	64.2	21.4	96.8	97.9	76.5
Cosmos-Predict2-2B (NVIDIA, 2025)	88.7	92.1	97.6	49.3	65.9	21.6	91.9	94.6	75.2
Cosmos-Predict2-14B (NVIDIA, 2025)	89.6	92.8	98.0	49.8	67.5	21.5	92.2	94.9	75.8
Wan2.1-I2V-14B (Team Wan et al., 2025)	90.4	93.0	98.0	51.7	70.0	21.4	96.5	97.2	77.3
Wan2.2-TI2V-5B (Team Wan et al., 2025)	92.0	93.9	98.8	52.3	69.6	21.4	97.4	98.0	77.9
Cosmos-Predict2.5-2B (Ali et al., 2025)	92.3	94.2	99.1	52.8	70.7	21.2	96.2	97.2	**78.0**
Wan2.2-I2V-A14B (Team Wan et al., 2025)	91.2	93.2	98.3	51.8	69.2	21.5	97.2	97.8	77.5
Ours	93.5	94.8	99.1	49.9	62.6	21.4	97.2	97.6	77.0

6.4.2 COMPONENT ABLATIONS

Our internal evaluations use two non-public test sets: ID-128 for in-domain single-view future prediction and MV-261 for the reported main-camera and two wrist-camera RGB comparisons. These evaluations use exponential moving average (EMA) weights; multi-view scores average the three views. For the component comparisons, the RGB-only reference observes the main-camera first frame on both datasets. Inference-time ablations modify a fixed checkpoint; the continued-training comparison uses matched initialization, data mixtures, and training budgets.

Cross-stream attention and identity embeddings. Table 10 compares the RGB-only reference with variants that restrict cross-stream attention or zero the view and modality identity embeddings. Restricting attention to within-stream tokens at inference lowers MV-261 PSNR from 15.74 to 13.59 dB. Restricting attention during matched-budget continued training (R2) also reduces MV-261 PSNR to 14.01 dB. On ID-128, the continued-training variant changes PSNR only from 19.77 to 19.65 dB, compared with the larger decrease in matched-budget MV-261 continued training. These results support the contribution of cross-stream attention to multi-view RGB prediction in this setting.

Zeroing both view and modality identity embeddings at inference lowers MV-261 PSNR to 14.65 dB and raises LPIPS from 0.441 to 0.499, supporting their contribution to RGB reconstruction. Both interventions leave the canvas rotary coordinates unchanged.

Table 10: **Cross-stream attention and identity embedding ablations.** R0 is the RGB-only reference. Attention masking and identity-table zeroing are inference-time interventions; R2 restricts attention during matched-budget continued training. PSNR is in dB. MV-261 component PSNR values use the original protocol, including frame zero.

	ID-128		MV-261	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
R0 (reference)	19.77	0.214	15.74	0.441
Cross-stream attention off (inference)	—	—	13.59	0.551
R2: attention restricted (continued training)	19.65	0.217	14.01	—
Both identity tables zeroed (inference)	—	—	14.65	0.499

Attention restriction during continued training. The inference-time intervention and continued-training restriction in Table 10 answer different questions. R2 inherits an earlier checkpoint; the restriction does not imply that the model has never learned cross-stream interactions.

Identity embeddings, attention, and rotary layout. Table 10 tests the joint removal of the view and modality embedding tables at inference. This is a sensitivity test of their combined input in an RGB-only configuration; it does not separate the contributions of the two tables or directly test discrimination between different modalities. Rotary coordinates are not disabled in any reported training run. Proposition 1 establishes a common-offset property within one stream, while the image-domain comparison of Wu et al. (2025b) provides external context. Neither substitutes for a matched prediction-quality ablation of canvas RoPE.

Per-view component sensitivity. Table 11 expands the RGB-only comparison in Table 10 using the same original scoring protocol, including frame zero. Restricting cross-stream self-attention mainly affects the wrist views, which have no RGB observations of their own; the main-view PSNR changes little. The shared vision–language context is retained in both configurations.

Table 11: **Per-view component comparison under the original protocol.** RGB-only prediction with the main-camera first frame observed. These PSNR values (dB) include frame zero and are not directly comparable with the future-only results in Table 12. Attention off masks self-attention between distinct streams.

Cross-stream attention	Main	Wrist 1	Wrist 2	Mean
On	18.81	14.12	14.29	15.74
Off	18.78	10.83	11.17	13.59

6.4.3 Controlling observations and joint depth generation

At the same short-budget R0 checkpoint, we independently vary wrist-camera RGB first-frame observations and joint depth generation in a matched 2×2 comparison (Table 12). A and D observe only the main-camera first frame; B and C additionally observe both wrist-camera first frames. C and D generate depth entirely from noise, without ground-truth depth input, while A and B generate only the three RGB streams. All four configurations use the same sampling settings and common future-frame scoring protocol. Cross-stream attention is enabled in all four configurations.

Checkpoint and sampling. All runs use EMA weights from the short-budget R0 reference checkpoint at $5{,}000$ training steps. A reuses the RGB-only reference outputs; B–D are additional runs of the specified configurations. Sampling uses FlowMatch Euler with 25 steps, guidance scale 5.0, and sigma shift 5. Each case uses its original evaluation-list index as the sampling seed. RGB streams precede the optional depth stream in the noise-drawing order. The main-camera first frame and the original caption provide the shared vision–language context in every configuration; no depth-specific instruction is added. The main RGB resolution is 480×640 and each wrist resolution is 224×320 ($H \times W$). Temporal buckets contain 81, 161, or 321 pixel frames.

Paired evaluation and reference videos. All four configurations use the same matched cases from MV-261 (Table 12). Predictions are compared with VAE-reconstructed reference videos, rather than raw dataset pixels. Decoded reference pixels on the scored future frames are identical across the four configurations for each case and view. Scoring retains each view's native resolution and requires matching prediction/reference shapes; it does not resize or truncate mismatched videos.

Common future-frame protocol. Every configuration is scored on the same pixel-frame indices $1, \ldots, \widetilde{T} - 1$, where $\widetilde{T}$ is the pixel-frame count, excluding frame zero from *all* RGB views, including the unobserved wrists in A and D. The causal VAE encodes the first pixel frame separately, so the observed first-frame latent does not include later pixel frames. PSNR is computed per frame with pixel range $[0, 255]$ and then averaged. SSIM (Wang et al., 2004) uses a 7×7 uniform window, valid spatial support, sample covariance, and RGB-channel averaging. LPIPS (Zhang et al., 2018) uses AlexNet with inputs normalized to $[-1, 1]$. All three metrics average over every scored future frame. We first average the three view scores within each case, then average cases with equal weight.

Paired uncertainty. For each contrast, we compute within-case differences between configuration scores. We compute percentile 95% confidence intervals using $10{,}000$ bootstrap resamples grouped by recorded episode identifiers, retaining paired configurations and all views together; each recorded identifier corresponds to one evaluated case. The intervals are computed separately for each contrast and quantify case-level uncertainty at the fixed checkpoint and sampling seeds, not variation across training runs.

Interpretation. Providing wrist-camera observations improves future-frame RGB PSNR by 1.97 dB without depth generation (B–A) and 1.98 dB with depth generation (C–D), with consistent SSIM and LPIPS improvements (Table 13). The corresponding intervals exclude zero for all three metrics. For B–A, the per-view PSNR differences are 0.0685, 3.2493, and 2.5927 dB for the main camera and the two wrists, respectively; the average improvement is concentrated in the newly observed wrists. With RGB observations fixed, adding a generated depth stream does not show a reliable RGB-quality improvement: all three 95% intervals include zero for both C–B and D–A. The interaction intervals also include zero. These results support additional RGB observations as

Table 12: **Matched observation–depth configurations on MV-261.** All rows use the same R0 checkpoint and the common future-frame protocol on the same $n = 181$ paired cases. *All* denotes the main camera and both wrists; *Main* denotes the main camera only. Depth, when present, is generated without ground-truth depth observations. Scores are means of case-level three-view averages; PSNR is in dB.

	RGB first frames	Generate depth	PSNR↑	SSIM↑	LPIPS↓
A	Main	No	15.6635	0.5634	0.44172
B	All	No	17.6337	0.6361	0.36122
C	All	Yes	17.6548	0.6371	0.36116
D	Main	Yes	15.6764	0.5634	0.44270

Table 13: **Paired contrasts under the common future-frame protocol.** Each cell reports the mean difference and its 95% bootstrap interval. B–A and C–D add wrist-camera observations; C–B and D–A add a generated depth stream with RGB observations fixed. $I = (C - B) - (D - A)$ is their interaction. Positive PSNR/SSIM differences and negative LPIPS differences indicate improved RGB prediction.

Contrast	ΔPSNR (dB)	ΔSSIM	ΔLPIPS
B–A	$+1.9702$ $[1.7241, 2.2320]$	$+0.07276$ $[0.06357, 0.08241]$	-0.08050 $[-0.09002, -0.07121]$
C–B	$+0.0212$ $[-0.0196, 0.0663]$	$+0.00099$ $[-0.00057, 0.00294]$	-0.00005 $[-0.00235, 0.00183]$
D–A	$+0.0129$ $[-0.0188, 0.0440]$	$+0.00002$ $[-0.00090, 0.00094]$	$+0.00098$ $[-0.00048, 0.00250]$
C–D	$+1.9784$ $[1.7252, 2.2484]$	$+0.07373$ $[0.06423, 0.08348]$	-0.08153 $[-0.09117, -0.07219]$
I	$+0.0082$ $[-0.0355, 0.0540]$	$+0.00097$ $[-0.00078, 0.00307]$	-0.00103 $[-0.00350, 0.00109]$

the source of the measured improvement in this fixed-checkpoint inference comparison. The depth contrasts do not establish an RGB-quality benefit or equivalence. These contrasts concern fixed-checkpoint inference, not the effect of auxiliary-modality training or the quality of generated depth. Single-view auxiliary comparisons under the original protocol are reported below.

Single-view auxiliary configurations. Under the original scoring protocol on ID-128, adding depth, edges, or segmentation as generation targets, with RGB first-frame conditioning unchanged, changes mean RGB PSNR by -0.05, -0.03, and -0.06 dB, respectively. These descriptive point differences do not establish equivalence or a general auxiliary-modality benefit.

6.5 Training and Implementation Details

Data sources and visual-stream preparation are described in Appendix 6.1. This subsection specifies the model configuration, sampling, training curriculum, and execution settings.

6.5.1 Model and trainable parameters

The backbone is a 5B-parameter video diffusion transformer ($d = 3072$, 30 layers, 24 heads). The causal autoencoder and Qwen3.5-9B vision–language encoder remain frozen. We train the DiT and semantic adapter using the shared objective in Eq. (10), with a reduced learning-rate scale for the adapter. During the multi-stream phase, the full DiT, including the two zero-initialized identity embedding tables, is trained.

6.5.2 Sampling and conditioning

Source and clip sampling. Table 14 gives the stage-dependent weights of general, egocentric, and embodied video. Sampling weights are renormalized over available sources. Within the data hierarchy, we weight each node by its subtree size at temperature $\alpha = 0.5$ and cap a clip's probability

at $10\times$ the uniform probability. On a collection of 1.02M clips, this raises the effective sample size, $\mathrm{ESS} = 1/\sum_e p_e^2$ for clip probabilities p_e, from $5{,}985$ under uniform tree descent to 289K, while reducing the effective number of distinct tasks by roughly 10%.

Variable stream composition. Training samples contain one to eight streams, subject to a 50K-token budget. During the packed-stream phase, temporally aligned RGB views and dense visual modalities are assembled jointly; their numbers and combinations vary with each source's available observations. Samples span three temporal buckets and about eight spatial resolutions, ranging from roughly 240p to 720p. The host denotes the designated main RGB view; stage-dependent resolution selection is specified in Appendix 6.5.3.

Conditioning and prediction roles. From S3 onward, RGB, depth, edge, and segmentation streams participate according to their availability. Each stream is assigned a role from Section 3.1, specifying a clean observation, an observed prefix with a predicted suffix, or a fully predicted stream. These assignments support visual prediction, cross-view completion, modality-conditioned generation, modality prediction, and joint generation within the same interface. Stream composition and role assignment vary together rather than restricting training to a fixed RGB–modality tuple.

Prompt construction. Prompts provide the semantic context described in Section 3.2. For modality-conditioned generation, we retain scene and object descriptions with probability 0.95 and increase the prompt-length budget to preserve appearance information. In samples with a main-view RGB observation, these descriptions are usually omitted because that observation provides the corresponding visual information.

6.5.3 Training stages and resolutions

Training proceeds through a single-stream phase (S1–S2) and a packed-stream phase (S3–S5). Table 14 brings together the sampling weights, resolution mix, and approximate compute for each stage. Each stage warm-starts from the previous EMA weights; S1 starts from a single-view video generator (Section 3.3). The late-phase optimization configuration is specified in Appendix 6.5.4.

Table 14: **Training recipe and approximate compute.** Sampling weights by category (%) and main-view training resolution. The three categories are renormalized to 100 after setting aside a small internal residual. Resolution shares are approximate; the 720p share includes high-resolution source clips. In Phase II, resolution refers to the host stream; wrist and auxiliary streams remain at 240p. GPU-hours are H200-class equivalent estimates rounded to the nearest thousand, not wall-clock measurements, and exclude backbone pretraining.

Stage	Gen.	Ego.	Emb.	Main-view resolution	GPU-h
Phase I: single stream					
S1	65	10	25	240p	∼10k
S2	45	12	43	240p / 480p mix	∼4k
Phase II: packed streams					
S3	25	14	61	240p (20% 480p)	∼8k
S4	15	16	69	240p (35% 480p)	∼1k
S5	10	17	73	240p; 45% 480p; ∼15% 720p	∼1k

Phase I: alignment and domain adaptation. S1 aligns the semantic adapter with the pretrained backbone using a mixture dominated by general video. S2 increases the proportion of egocentric and embodied footage, establishing the single-view model used in subsequent stages. Both stages use one stream per sample, selected from an available viewpoint in the clip. At the end of S2, the two identity tables are introduced with the initialization defined in Section 3.2.

Phase II: multi-stream expansion. S3 introduces packed RGB views and dense modalities together, using the variable stream composition and role assignment above while retaining single-view samples. S4 and S5 increase the embodied-data share without adding stream types. Task roles are

sampled within each stage over its available visual stream types, so stages expand the supported variables rather than specialize the model to individual prediction directions. After the identity tables are introduced, these transitions change the sampler without changing parameter shapes. A residual general-video share is retained throughout to preserve broad motion and appearance coverage.

Resolution schedule. The schedule selects the input resolutions (Section 3.2). S1 uses 240p clips; S2 mixes 240p and 480p. S3 begins with predominantly 240p host views to accommodate the additional streams: the host is drawn at 240p with initial probability 0.80, which decays linearly across the packed-stream phase. High-resolution source clips retain their encoding resolutions. In S5, the approximate host-resolution training mix includes 45% at 480p and 15% at 720p, with the remainder at 240p; the 720p share includes the high-resolution source clips. Wrist views and auxiliary streams remain at 240p throughout the packed-stream phase, and aligned modalities use the coordinate stride in Eq. (5).

6.5.4 Optimization and execution

Late-phase optimization. For the late multi-stream configuration, we use AdamW with $\beta = (0.9, 0.999)$ and weight decay 0.01. The peak learning rate is 2×10^{-6} for the DiT and 2×10^{-7} for the adapter. A 250-update warmup is followed by cosine decay to 5% of the peak. The run uses 88 data-parallel workers, one sample per worker and one gradient-accumulation step, giving a global batch size of 88. Optimization uses bf16 and ZeRO stage-2. EMA is updated after every optimizer step, with target decay 0.9998 and EMA warmup enabled. These settings describe this late-phase configuration rather than a uniform recipe for all stages.

Execution. A sample is one packed sequence and fits on a single device, so we do not use tensor or sequence parallelism. Packing keeps self-attention block-diagonal across samples (Section 3.3).

Cached latents. Latent resolutions are fixed at encoding, limiting pixel-space augmentation. Each stage selects among the available cached resolutions without modifying the model.

Online context encoding. The frozen vision–language encoder processes the instruction and one observed main-view RGB frame *online* at every step, using the neutral placeholder described in Section 3.2 when no such observation is available. It does not encode the full observation prefix. Instructions range from short commands to long scene descriptions. A small trained adapter maps the resulting variable-length context onto the backbone's fixed cross-attention interface, so prompt length and style do not change the DiT's weight shapes or the packing of the visual tokens.

Geometry buckets. A sample's *geometry* comprises task mode, horizon bucket, stream count, and per-stream resolution, forming a discrete product space. At each step, all ranks use the same geometry sampled with a shared seed, while each rank selects its own episode and clip.

Geometry varies across training steps, preserving the mixture's marginal distribution over shapes while constraining the joint distribution across ranks.

The reported run uses one sample per rank, so geometry buckets synchronize shapes across ranks without packing multiple samples into a sequence.

6.5.5 TOKEN COST

The token count of a sample is the sum of the token counts of its streams, each counted at its own resolution: a stream of $\widetilde{T}$ pixel frames at $H \times W$ contributes $\left(\frac{\widetilde{T}-1}{4}+1\right) \cdot \frac{H}{32} \cdot \frac{W}{32}$ tokens after the autoencoder and patching. For a 161-frame sample with a 480×640 main view and two 224×320 wrist views, the three streams require $12{,}300 + 2{,}870 + 2{,}870 = 18{,}040$ tokens. For an equal-resolution pixel-stitching baseline that upsamples every view to 480×640, a layout with no empty cells costs $3 \times 12{,}300 = 36{,}900$ tokens, a factor of 2.05. A seven-view configuration with one 480×640 main view and six 224×320 views costs $29{,}520$ tokens as streams, versus $86{,}100$ for that stitching baseline, a factor of 2.9; the latter exceeds the 50K per-rank budget. A lower-resolution alternative is to resize all seven views to 224×320 before stitching, yielding $20{,}090$ tokens but reducing the main view's spatial pixel count by roughly three quarters. These counts compare the specified layouts, not a lower bound for every possible pixel-stitching arrangement.

Table 15: **Tokens per sample** after the autoencoder and patching, 161 pixel frames. The stitching baseline resizes every view to 480×640 with no empty cells. Native streams retain one 480×640 view and two or six 224×320 views. The seven-view stitching baseline exceeds the 50K per-rank budget.

Layout	3-view	7-view	vs. 50K budget
Streams (native res.)	18,040	29,520	fits
Pixel-stitched (480×640/view)	36,900 (2.05×)	86,100 (2.9×)	7-view does not fit