STRUCTURE BEYOND PIXELS: ALIGNING VIDEO WITH STRUCTURAL PREFERENCES FOR ROBUST POLICY GENERATION

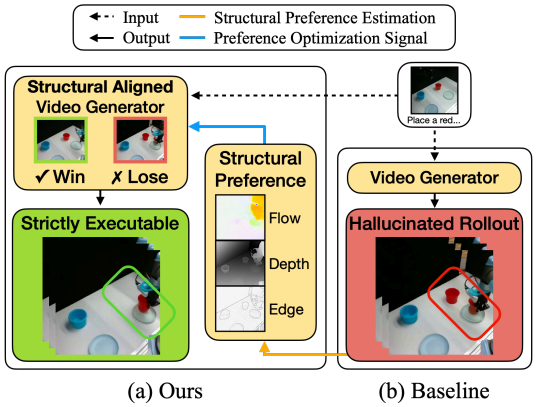
Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models can serve as world models for robotic policy generation. However, because they are trained to produce visually plausible videos without physical constraints, the generated rollouts often exhibit an *executability gap*, leading to downstream task failures. To address this, we propose Structure-Aligned Direct Preference Optimization (Struct-DPO), which aligns a pre-trained video generator with structural preferences derived from environment interaction. Specifically, we treat physically grounded simulator observations as winning samples and model rollouts as losing samples, automatically identifying the most severe violations via structural metrics for viewpoint and geometric consistency. Notably, this approach requires neither architectural changes nor explicit 3D supervision; in our experiments, it trains only 3.95% of the base model parameters at 0.27% of the additional step of standard supervised fine-tuning, yielding high parameter and computational efficiency. Experiments on the RoboCasa benchmark across 23 manipulation tasks show consistent gains in task success, and alignment learned from a subset of tasks transfers to unseen tasks, indicating that the distilled physical priors generalize across tasks and environments.

1 INTRODUCTION

Recent video generative models have transformed how robots perceive and reason about the physical world. Pre-trained on internet-scale data, large video generation models implicitly encode rich visual priors about object interactions, contact, and motion Brooks et al. (2024); He et al. (2026); Garrido et al. (2025). Leveraging these priors, contemporary work has shown that such models can be repurposed as visual world models for robotic policy generation: by predicting future frames conditioned on a goal description, they serve as visual subgoals that downstream policies decode into executable actions Liang et al. (2025); Hu et al. (2024); Du et al. (2023); Shen et al. (2026). This paradigm leverages internet-scale video pre-training to overcome a key bottleneck of action-based policies—generalization beyond their training distribution—enabling zero-shot transfer to novel objects and environments with only a handful of robot demonstrations.



(a) Ours (b) Baseline

Figure 1: Our structure-aligned generator uses physics grounded signals to produce strictly executable videos, while baselines often generate visually plausible but hallucinated rollout.

However, video models are optimized for perceptual plausibility rather than physical consistency Gu et al. (2025); Yuan et al. (2025). Trained on action-less videos that record only the visual consequences of motions without the underlying control signals or physical forces that caused them, these models learn what scenes look like as they change, not how they physically evolve, as shown in Figure 1(b). When deployed as world models for embodied agents, this manifests as an exe-

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.

cutability gap: Rollouts that appear visually coherent but contain physically impossible transitions, e.g., objects defying gravity, end-effectors moving in sync with uncontacted objects, or shapes that collapse mid-trajectory Kang et al. (2024); Motamed et al. (2026); Lin et al. (2026). Downstream policies that faithfully decode these subgoals into actions inherit and amplify these violations, so even visually coherent rollouts translate directly into failed manipulations on the robot.

To address this limitation, we propose Structure-Aligned Direct Preference Optimization (Struct-DPO). This framework preserves the generalization advantage of video-based policies while closing the executability gap, by aligning a pre-trained video policy model with environment-derived structural preferences. As shown in Figure 1(a), our core insight is that physical environments, long underutilized as data sources for video model training, provide a direct signal for what physically grounded dynamics look like. Specifically, we construct preference pairs where the environment observation serves as the physically grounded *winning* pair and the model's hallucinated rollout serves as the *losing* pair.

Importantly, our approach bypasses the architectural modifications and training burdens required by explicit 3D supervision or joint generation pipelines Chefer et al. (2025); Liu et al. (2025); Bai et al. (2025) with structural preferences. The model generates ordinary RGB frames, altering only the alignment signal. These structural preferences are computed directly by comparing the environment observation against the model's own rollout using only off-the-shelf vision tools (depth estimation, edge detection, and optical flow). Therefore, the only learning machinery our framework adds is the alignment step itself, which we make parameter- and compute-efficient with LoRA: aligning a Stable Video Diffusion-based policy with only 0.27% of the original Supervised Fine Tuning (SFT) budget while introducing just 3.95% additional parameters. Despite this minimal overhead, Structure-Aligned DPO consistently improves task success across 23 manipulation tasks on the RoboCasa benchmark. Notably, alignment on a subset of tasks transfers to improved success rates on held-out tasks not seen during alignment, suggesting that the physical priors distilled through structural preferences generalize beyond the specific task distribution used for training.

2 Related Work

2.1 Video Generation Models as Policy Learners

Recent advances in large-scale video generation Blattmann et al. (2023); Brooks et al. (2024) have enabled a new paradigm for robotic policy learning: video models serve as visual world models that synthesize future observations, which downstream controllers decode into action sequences. UniPi Du et al. (2023) formalized text-conditioned video generation as universal policies, while Video Policy Liang et al. (2025) and Video Prediction Policy (VPP) Hu et al. (2024) showed that decoupling visual prediction from action decoding yields strong zero-shot generalization to novel objects and instructions, and Unified Video-Action models Li et al. (2025) further explored joint optimization of perception and control. These approaches inherit the broad visual priors of web-scale pre-training, but the underlying internet videos record only the visual consequences of motion, not the physical forces that produced them. Lacking explicit grounding in physical dynamics, the generated rollouts frequently violate contact and object permanence constraints Gu et al. (2025); Yuan et al. (2025); Lin et al. (2026); Kang et al. (2024), an issue that downstream policies cannot recover from. This mismatch constitutes the executability gap that motivates our work. A complementary line of research has shown that video generators often default to surface-level pattern completion rather than genuine physical reasoning (Kang et al., 2024; Garrido et al., 2025), motivating mechanisms that actively close this gap rather than merely diagnose it. Our work directly targets it, aligning the generative process itself with physical constraints rather than relying on the controller to compensate.

2.2 Geometric and Physical Grounding in Video Generation

Injecting structure into video generation has been explored along two axes: *explicit* representation by augmenting outputs with 3D-aware modalities and *implicit* regularization by adding geometry-driven preferences. **Explicit approaches.** These methods augment the generation target with 3D-aware outputs, requiring modification of either the model's output modality or its architecture. In the image domain, Intrinsic Latent Diffusion Models (Lee et al., 2026) co-generate depth, surface

normals, and segmentation maps alongside RGB images. Extending this paradigm to video, VideoJAM (Chefer et al., 2025) jointly generates RGB frames and optical flow to improve motion coherence, GeoVideo (Bai et al., 2025) predicts per-frame depth maps jointly with RGB frames and enforces multi-view reprojection consistency across them, and WorldWeaver (Liu et al., 2025) performs joint video-intrinsic generation for long-horizon scenarios. While effective, these methods either require retraining the generator with additional output channels or depend on closed-source joint generation architectures. We instead treat structure as an alignment signal applied to a frozen base model, leaving the output modality unchanged. **Implicit approaches.** A complementary line of work injects structural constraints through preference-based alignment without modifying what the model generates. Kupyn et al. (2025) construct DPO preferences from epipolar geometry errors, while Du et al. (2026) extends this with dense reprojection-based consistency signals derived from a geometry foundation model. These approaches share our high-level motivation but differ in scope: they target intrinsic video quality on standard generation benchmarks, whereas we apply structural alignment to a video model as a world model for embodied policy generation, where the executability gap directly translates to task failure.

2.3 Preference Optimization for Visual Generation

Direct Preference Optimization (DPO) (Rafailov et al., 2023) was originally introduced to align language models with human preferences by reformulating reinforcement learning from human feedback as a closed-form classification problem. Wallace et al. (2024) extended DPO to diffusion models via denoising score matching, enabling alignment of visual generation with aesthetic and prompt-following preferences. Subsequent work has applied diffusion DPO to image quality (Liang et al., 2024), text-image alignment (Karthik et al., 2025), and video (Liu et al., 2024), with preferences typically drawn from human annotations or learned reward models.

A growing body of work has shifted toward more objective preference signals that target physical plausibility rather than perceptual quality. NewtonRewards (Le et al., 2025) optimizes physical plausibility through complementary verifiable physics rewards, PhysHPO (Chen et al., 2026) aligns video generation across hierarchical granularities including motion and state-level consistency, and PhysCorr (Wang et al., 2025) introduces a dual-reward DPO pipeline guided by VLM-based physical plausibility scores. While these approaches share our motivation of moving beyond pixel-level quality, they typically rely on a single reward—whether a learned plausibility score, a multi-component reward, or a VLM judgment—evaluated on the generated video alone. Instead, we construct preference signals from multiple structured representations of the scene—3D viewpoint consistency, 2D boundary alignment, and pixel-level kinematic deviation—each derived from off-the-shelf vision models without additional training, so that every dimension surfaces a distinct class of physical violation and together they capture the executability gap more comprehensively than any single metric.

3 Physically Grounded Video Generation via Structure-Aligned DPO

In this section, we propose a method to close the executability gap. To this end, we align the video generation model with the physical constraints of the environment through DPO, guided by environment-derived structural preference data.

3.1 Problem Formulation

Let $\mathcal{M}_\theta$ denote a video generation model (specifically, Stable Video Diffusion (SVD) Blattmann et al. (2023)) pre-trained on web video and fine-tuned via SFT on robot manipulation demonstrations. Given an initial frame $\mathbf{x}_0$ and a task description c, the model generates a rollout $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_T) \sim p_\theta(\mathbf{X} \mid \mathbf{x}_0, c)$, which a downstream policy π_ϕ (Video Policy Liang et al. (2025)) decodes into an action sequence $\hat{\mathbf{a}} = \pi_\phi(\hat{\mathbf{X}})$. Executing $\hat{\mathbf{a}}$ in the environment yields a physically grounded rollout $\mathbf{X}^*$ conditioned on $(\mathbf{x}_0, c)$, which we use as the alignment target for $\hat{\mathbf{X}}$.

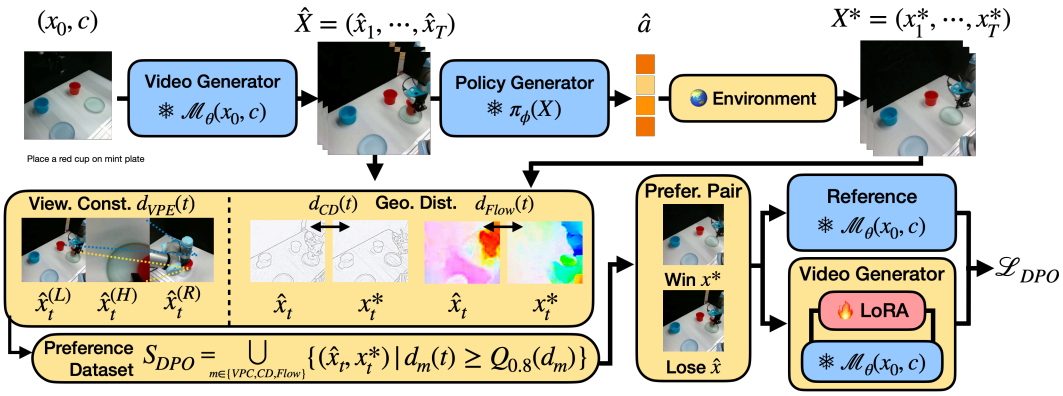


Figure 2: Given a task instruction, the video generator produces a rollout $\hat{\mathbf{X}}$, which is executed to obtain the observation $\mathbf{X}^*$. We measure viewpoint consistency and geometric distance with $(\hat{\mathbf{X}}, \mathbf{X}^*)$, and build S_{DPO} from the top 20% highest-error samples across all metrics. Using S_{DPO}, we inject LoRA adapters into the video generator and perform DPO-based structural alignment.

3.2 Environment-Based Structural Preference Data Construction

To ensure the model learns robust physical constraints, we construct the preference dataset by emphasizing pairs with the most severe physical inconsistencies. By executing the action sequence associated with each generated video in the environment, we estimate the severity of the executability gap between the generated video and the ground-truth observation. Since the executability gap must be assessed from 2D videos without relying on extensive 3D annotations, we use lightweight, off-the-shelf 2D vision tools. We deliberately select complementary metrics that capture the most critical dimensions of physical interaction based on two primary criteria: Viewpoint Consistency and Geometric Distance. Viewpoint Consistency is captured by a single metric, while Geometric Distance is characterized by two metrics: Edge Chamfer Distance and Optical Flow Displacement.

3.2.1 Viewpoint Consistency.

To evaluate whether the generated scene maintains structural integrity across viewpoints, we measure scene consistency through depth-based reprojection. Depth reprojection has been used in prior work to encourage geometric consistency in generated videos Bai et al. (2025); we adapt this tool to explicitly measure cross-view consistency at the same timestep, and introduce a per-view scale calibration that aligns the monocular depth scales across viewpoints.

Let i, j be source and target views with intrinsics K_i, K_j and relative pose $\{R_{i\to j}, t_{i\to j}\}$, and let D_i be the unscaled monocular depth from Video Depth Anything Chen et al. (2025). Warping pixel p from view i to view j with learned scale s_i gives $\mathbf{P}_{i\to j} = R_{i\to j}(s_i D_i(p) K_i^{-1} \tilde{p}) + t_{i\to j}$, and the directional reprojection error is

$$E_{i\to j}(s_i, s_j) = \frac{1}{|V|} \sum_{p \in V} \left\| [\mathbf{P}_{i\to j}]_z - s_j D_j(\pi_j(\mathbf{P}_{i\to j})) \right\|_1, \tag{1}$$

where V is the set of valid reprojected pixels.

Since monocular depth estimation lacks absolute metric scale and may diverge from the camera parameter scale, we calibrate the scale factors based on the first conditioning frame ($\mathbf{x}_0$). In our setup, we use the Hand, Left, and Right cameras (H, L, R) and the Left and Right views share a common scale s_{base}. We first optimize s_{base} by minimizing the bidirectional error between L, R:

$$s_{\text{base}} = \arg\min_s \left(E_{L\to R}(s, s) + E_{R\to L}(s, s) \right). \tag{2}$$

Fixing s_{base}, we subsequently determine the optimal scale for the Hand camera, s_{hand}, that minimizes the reprojection error to the calibrated auxiliary views:

$$s_{\text{hand}} = \arg\min_s \left(E_{H\to L}(s, s_{\text{base}}) + E_{H\to R}(s, s_{\text{base}}) \right). \tag{3}$$

Applying these calibrated scales across all generated frames, the final bidirectional viewpoint consistency error is evaluated as the aggregate error between the hand and the auxiliary views:

$$d_{\mathrm{VPE}}(t) = \sum_{j \in \{L,R\}} \left(E_{H \to j}(s_{\text{hand}}, s_{\text{base}}) + E_{j \to H}(s_{\text{base}}, s_{\text{hand}}) \right). \tag{4}$$

This metric explicitly penalizes spatial hallucinations, ensuring the generated scene remains geometrically coherent in 3D space rather than merely plausible in 2D.

3.2.2 Geometric Distance.

To capture fine-grained physical discrepancies, we measure geometric distance from the perspective of the hand camera, where physical interactions are most prominent. We utilize two sub-metrics based on 2D grid representations:

Edge Chamfer Distance: Measures the spatial misalignment between the object boundaries in the generated video and the environment observation. Rather than using explicit 3D point clouds, we operate directly on 2D edge maps to preserve grid structure efficiency. For a given frame at time t, let $\mathcal{E}_t^*$ and $\hat{\mathcal{E}}_t$ denote the set of 2D pixel coordinates of the edges extracted via Canny edge detection from the environment frame $\mathbf{x}_t^*$ and the generated frame $\hat{\mathbf{x}}_t$, respectively. Utilizing the Euclidean distance transform, the symmetric 2D edge Chamfer distance is computed as:

$$d_{\mathrm{CD}}(t) = \frac{1}{|\mathcal{E}_t^*|} \sum_{\mathbf{p} \in \mathcal{E}_t^*} \min_{\hat{\mathbf{p}} \in \hat{\mathcal{E}}_t} \|\mathbf{p} - \hat{\mathbf{p}}\|_2 + \frac{1}{|\hat{\mathcal{E}}_t|} \sum_{\hat{\mathbf{p}} \in \hat{\mathcal{E}}_t} \min_{\mathbf{p} \in \mathcal{E}_t^*} \|\hat{\mathbf{p}} - \mathbf{p}\|_2. \tag{5}$$

Optical Flow Displacement: Quantifies the pixel-level kinematic deviation between the generated observation and the true physical state. Instead of comparing temporal motion within the respective videos, we directly estimate the cross-modal spatial displacement between the environment frame $\mathbf{x}_t^*$ and the generated frame $\hat{\mathbf{x}}_t$ at the exact same timestep t. Using a pre-trained RAFT Teed & Deng (2020) model, we compute the dense optical flow field $\mathbf{F}_t = \mathrm{RAFT}(\mathbf{x}_t^*, \hat{\mathbf{x}}_t)$. The flow error is defined as the mean magnitude of these displacement vectors across the pixel grid Ω:

$$d_{\mathrm{Flow}}(t) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\mathbf{F}_t(p)\|_2. \tag{6}$$

For training, we calculate $d_{\mathrm{VPE}}(t), d_{\mathrm{CD}}(t)$ and $d_{\mathrm{Flow}}(t)$ and construct our DPO dataset S_{DPO} by selecting the samples above the 0.8 quantile with the highest error in each respective category. This ensures the alignment focuses on correcting the most severe physical violations.

3.3 Structure-Aligned DPO Training

We describe how preference pairs are formed and the regularization tailored to this setting. All other parameters are frozen, preserving the general visual generation capability of the pre-trained model.

DPO for Successful Tasks. If the task is successfully executed in the environment, we perform standard DPO. The environment's observation $\mathbf{X}^*$ serves as the *winning sample*, representing the physically correct execution of the task described by the text condition c. The model's generated video $\hat{\mathbf{X}}$ serves as the *losing sample*, representing a visually plausible but physically flawed rollout. This encourages the video model to favor physically consistent dynamics over superficial visual plausibility, thereby ensuring the downstream policy π_ϕ receives strictly executable visual subgoals.

Hindsight Relabeling for Failed Tasks. When a task fails (e.g., wrong grasp location), the environment rollout (winning pair) still reflects correct physical constraints but no longer demonstrates task completion, while the model's hallucinated rollout (losing pair) may instead depict a physically implausible "success." Since neither sample reflects successful task execution, conditioning standard DPO on the original task description c would be misleading. We therefore apply *hindsight relabeling* Andrychowicz et al. (2017): c is replaced with a generic interaction prompt (e.g., "Move safely and interact with objects"), redirecting the DPO objective from task completion to the underlying physical and contact dynamics.

Table 1: Quantitative results. DPO demonstrates superior physical consistency and success rates.

Method	Physical Consistency & Quality				Task Success Rate (%)		
	VPE ↓	Chamfer ↓	Flow Err. ↓	FVD ↓	PnP	Non-PnP	Overall
Video Policy	0.1191	10.075	11.570	106.466	39.17	59.56	52.46
+Struct-DPO (All)	0.1072	10.091	11.172	104.436	**42.50**	60.00	53.91
+Struct-DPO (PnP Only)	**0.1047**	**9.790**	**10.994**	**102.131**	41.25	**63.33**	**55.65**

DPO Objective for Diffusion Models. Following Diffusion-DPO Wallace et al. (2024), we adapt the preference objective to the diffusion denoising setting Ho et al. (2020): the implicit reward of a generated video is the log-likelihood ratio between the model p_θ and the frozen reference p_{ref}, and the negative log-likelihood is approximated by the per-sample denoising loss $\ell_\theta(\hat{\mathbf{X}}, \sigma) = \mathbb{E}_{\boldsymbol{\epsilon}}[\omega(\sigma)\|\boldsymbol{\epsilon}_\theta(\hat{\mathbf{X}}_\sigma, \sigma) - \hat{\mathbf{X}}\|^2]$, with $\omega(\sigma)$ the noise-schedule weighting. The resulting DPO loss (with $\beta = 2.0$) is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\Big[\log\sigma\Big(\beta(\ell_{\text{ref}}(\mathbf{X}^*) - \ell_\theta(\mathbf{X}^*)) - \beta(\ell_{\text{ref}}(\hat{\mathbf{X}}) - \ell_\theta(\hat{\mathbf{X}}))\Big)\Big]. \tag{7}$$

To prevent reward hacking, where the model degrades generation quality merely to increase divergence from the losing sample, we jointly apply an SFT loss on ground-truth environment rollouts.

Sigma Sampling Adjustments for DPO training. We additionally shift the EDM Karras et al. (2022) sigma sampling distribution from the SFT baseline, $\sigma_{mean}, \sigma_{std} = (1.0, 1.6)$ to lower-noise values $\sigma_{mean}, \sigma_{std} = (0.0, 1.0)$, concentrating the DPO signal on late-stage denoising where structural hallucinations primarily manifest.

4 Experiments

4.1 Experimental Setup

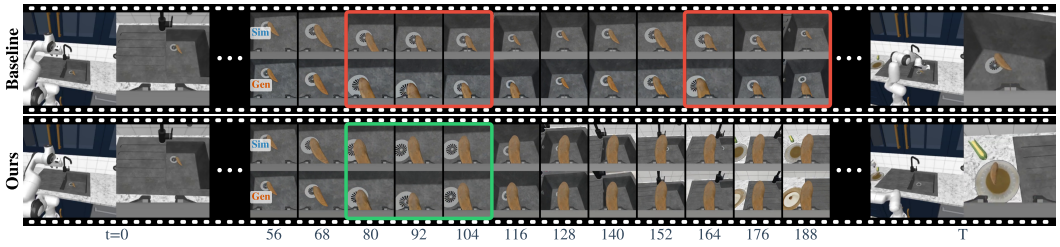
We evaluate our framework across diverse environments and tasks from the RoboCasa Nasiriany et al. (2024) benchmark, which provides procedurally generated kitchen environments with diverse object configurations and textures, as well as on a physical UR5e robot platform. We report downstream task success rates (%), alongside the executability gap metrics ($d_{\text{VPE}}, d_{\text{CD}}, d_{\text{Flow}}$) used to construct the preference data, to verify whether DPO minimizes its objective. We additionally measure Fréchet Video Distance (FVD) Unterthiner et al. (2018), which is not directly optimized by DPO, as an independent metric to assess whether the overall quality of the rollouts also improves. For training, we build upon the Video Policy (VP) Liang et al. (2025) architecture and perform Structure-Aligned DPO (Struct-DPO) by injecting LoRA modules into the denoising U-Net. For efficiency, we shorten the per-cycle inference: the original VP executes 16 of every 32 predicted action steps over 1,000 rollout steps, whereas we execute 24 steps and run 480 rollout steps in total. Detailed environment configurations, dataset construction, and hyperparameters are in Appendix B.

4.2 Benchmark Experimental Results

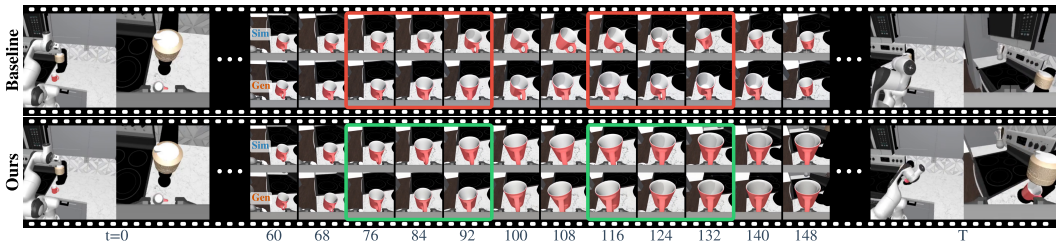
Quantitative Results on RoboCasa. This experiment addresses the following question: "Does Struct-DPO improve performance across diverse tasks?" To this end, we compare three models: the VP baseline, a model DPO-aligned on the entire dataset (*All Tasks*), and a model DPO-aligned exclusively on the Pick-and-Place subset (*PnP Only*). We evaluate these three models across 23 RoboCasa tasks and report the average of each metric over all tasks. As shown in Table 1, both DPO variants improve over the baseline in task success and in physical consistency, indicating stronger physical grounding. More notably, *PnP Only* improves performance on both seen PnP tasks and *unseen* Non-PnP tasks (overall 52.46% → 55.65%), showing that the constraints learned via Structure-Aligned DPO are not task-specific but reflect universal structural priors. We further observe that *PnP Only* often surpasses *All Tasks*, which we attribute to reduced task interference: PnP isolates fundamental interactions without the constrained dynamics of articulated objects, yielding a cleaner universal prior that better closes the executability gap across the full benchmark.

Table 2: Granular phase-wise success rates for Pick-and-Place tasks. Success rates are reported over 30 executions for each of the eight tasks.

Method	Grasp Success (%)	Transport Success (%)	Final Task Success (%)
Baseline	60.42	45.42	38.33
Struct-DPO	**62.50 (+2.08)**	**47.92 (+2.50)**	**41.25 (+2.92)**



(a) *Pick-and-Place* task.



(b) *Non-PnP* task (unseen during training).

Figure 3: Qualitative comparison. (a) Ours preserves contact and spatial coherence (green box), whereas the baseline violates physics (red box). (b) Ours transfers learned physics across tasks, preserving consistency in unseen (green box), whereas the baseline fails (red box).

Granular Analysis of Execution Trajectory in Pick-and-Place. While the overall task success rate provides a holistic measure of performance, it can obscure the model's capabilities in specific physical interactions. Specifically, we observed improvements not only in the final task outcome but also in intermediate physical interactions. To explicitly isolate these improvements, we decompose the execution trajectory of Pick-and-Place tasks into three sequential milestones: (1) *Grasp Success* (successfully making contact and lifting the target object), (2) *Transport Success* (maintaining the grasp and carrying the object to the target region without dropping it), and (3) *Final Task Success* (accurately placing the object to satisfy the completion condition). This granular breakdown allows us to characterize how the benefits of structural alignment manifest across different stages of the manipulation process. As shown in Table 2, our method consistently outperforms the baseline across all three stages, indicating that these benefits extend across the entire manipulation process.

Qualitative Analysis. Figures 3a and 3b show representative rollouts from the SFT baseline and our DPO-aligned model on a trained Pick-and-Place task and an unseen Non-PnP task, respectively. The SFT model frequently generates rollouts where object shapes collapse mid-sequence, contact dynamics are physically implausible (e.g., objects floating or interpenetrating surfaces), and auxiliary viewpoints are geometrically inconsistent with the primary view. In contrast, the DPO-aligned model maintains consistent object geometry across frames and viewpoints, produces physically plausible contact trajectories, and generates rollouts that are more faithfully executable by the robot policy. Notably, this holds even for the Non-PnP task in Figure 3b, which is absent from the alignment data: the physical priors acquired from Pick-and-Place interactions transfer to the unseen category, mirroring the quantitative trend in Table 1.

Figure 4 illustrates that our structural preference exposes physical inconsistencies that are not apparent from raw RGB alone. Although the baseline generation looks perceptually similar to the real observation, the three structural representations—edge maps, optical flow against the real frame, and depth with cross-view reprojection—reveal clear deviations that perceptual comparison would

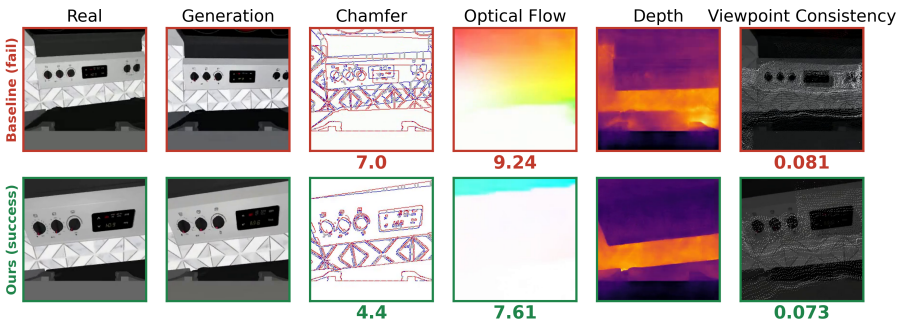


Figure 4: Per-metric visualization of Baseline (top) vs. Ours (bottom) on a representative frame. From left: real and generated RGB, Chamfer edge maps (d_{CD}), optical flow between real and generated (d_{Flow}), monocular depth and cross-view point cloud (d_{VPE}). Our method aligns each representation more tightly with the real observation.

Table 3: Ablation study on the components of the Structure-Aligned DPO framework.

Ablation Variant	**Success Rate (%)**
Full Method, Struct-DPO (PnP Only)	**55.65**
w/o Viewpoint Consistency Metric	52.17 (-3.48)
w/o Geometric Distance (Chamfer) Metric	53.91 (-1.74)
w/o Geometric Distance (Flow) Metric	53.91 (-1.74)
w/o Hindsight Relabeling (Failed Tasks Excluded)	50.43 (-5.22)
w/o Hindsight Relabeling (Task-Specific Description Retained)	53.91 (-1.74)
w/o Sigma Sampling Modification ($\sigma_{mean} = 1.0,\ \sigma_{std} = 1.6$)	52.17 (-3.48)

miss. We guide our model to minimize these deviations, and accordingly its generations score higher under our structural preference, mirroring the physical consistency improvement reported in Table 1.

Ablation Studies. As shown in Table 3, we conduct ablation studies to validate the individual contributions of our methodological choices. Among the inconsistency metrics, removing d_{VPE} causes the largest drop underscoring that enforcing 3D structural integrity is the most critical signal; omitting either d_{CD} or d_{Flow} also reduces success to 53.91%, showing that 2D boundary preservation and kinematic alignment remain essential. The two hindsight-relabeling ablations confirm that both components are important for learning task-agnostic dynamics. Consistent with our claim that severe physical hallucinations cause failures, 71% and 59% of Ours (PnP) and Ours (All) preference data, respectively, are failed rollouts. In contrast, successful trajectories may exhibit metric-level discrepancies without sufficiently severe violations to cause failure, potentially introducing noise during training. Thus, we argue that these results stem from the more abundant and informative training signal provided by failed rollouts. Finally, reverting the sigma-sampling distribution to the SFT default degrades performance, supporting our hypothesis that DPO updates are most effective when concentrated on late-stage denoising, where structural hallucinations primarily occur.

Training Efficiency and Parameter Overhead. The Structure-Aligned DPO introduces minimal computational and parameter overhead. Since the pre-trained VP already captures broad visual understanding and basic manipulation skills from 369K SFT steps, DPO only refines physical dynamics, converging in 1K steps (**0.27%** of SFT). Moreover, LoRA uses only 60.1M parameters (**3.95%** of the 1.52B VP model). These results highlight that strong physical priors can be efficiently injected into large generative models without full fine-tuning or expensive data collection.

4.3 Deployment on a Physical Robot

To verify that Structure-Aligned DPO transfers beyond simulation, we deploy it on a physical UR5e arm with three Intel RealSense D405 cameras (left, right, and wrist). We design three tasks: Task 1 moves the red cup to the mint-colored plate, Task 2 moves the blue cup to the mint-colored plate, and Task 3 moves the red cup to the blue plate. Figure 5 illustrates a representative setup and successful execution: the robot approaches the target cup, establishes a stable grasp, transports it, and completes

Table 4: Real-world success rate. DPO substantially improves over the baseline.

Single-task training (Task 1 only)

Method	Succ. (%)
Video Policy (Baseline)	52.0
+ DPO (All pairs)	78.0
+ DPO (Geometric subset)	**82.0**

Multi-task training (Task 1–3)

Task	Method	Succ. (%)
Task 1	Video Policy (Baseline)	48.0
	+ DPO (Geometric subset)	**65.0**
Task 2	Video Policy (Baseline)	62.0
	+ DPO (Geometric subset)	**72.0**
Task 3	Video Policy (Baseline)	24.0
	+ DPO (Geometric subset)	**32.0**



(a) Overview (b) Physical setup (c) t_1 (d) t_2 (e) t_3

Figure 5: Deployment setup (a–b) and temporal progression(c–e) of the robot manipulation task. (a) Task 1: red cup to mint plate; Task 2: blue cup to mint plate; Task 3: red cup to blue plate.

placement. We use the VP baseline, while the video and policy generator are further fine-tuned on 50 teleoperated demonstrations Wu et al. (2024). For DPO, preference pairs are constructed by treating real observations as the winning samples and the outputs of the further fine-tuned VP model as the losing samples. We report success rates over 50 runs. Full hardware setup, calibration of d_{VPE} on the moving wrist camera, and preference-data construction are in Appendix D. We evaluate our method under two experimental settings. First, in the *Single-task* setting, we train and compare models exclusively on Task 1 to isolate the effect of Structure-Aligned DPO, without potential performance degradation from jointly learning multiple tasks. Second, in the *Multi-task* setting, we jointly train on Tasks 1–3 to examine whether the benefits of our method extend across Pick-and-Place tasks.

Results As summarized in Table 4, Structure-Aligned DPO substantially improves task completion on hardware, and the gain over the SFT baseline is markedly larger than in simulation (Table 1). We attribute this to the greater visual variability in real-world (e.g., lighting, shadows, and sensor artifacts), which makes physically consistent video generation more challenging and thereby amplifies the benefits of structural grounding. Moreover, the geometric subset outperforms training on all pairs despite using only 45% of the data and far fewer optimization steps (102 vs. 951), supporting the effectiveness of our proposed S_{DPO} selection strategy over full training.

5 Conclusion

In this paper, we identify the executability gap in SFT-trained video world models: visually coherent rollouts that fail to respect physical constraints. To bridge this gap, our Structure-Aligned DPO framework constructs structural preference pairs from high-error samples under multi-view depth consistency and geometric distances, directly penalizing physical violations. Furthermore, we introduce hindsight relabeling with generic text conditioning for failed environment tasks. This separates the learning of fundamental contact dynamics from task-specific objectives, allowing the model to learn physics even from unsuccessful interactions and promoting robust structural alignment. Ultimately, this environment-grounded optimization shifts the paradigm from computationally expensive explicit structural generation to implicit alignment, providing a scalable mechanism to inject physical priors into video generation models.

Limitations. We highlight three limitations: (i) the use of a single LoRA-aligned backbone with late-stage-focused DPO updates, (ii) the inference-time cost of generating video rollouts as visual subgoals, and (iii) the use of purely visual-structural preference signals. See Appendix E for details.

AI USE STATEMENT

We used generative AI tools including literature review, experimental feedback, method implementation, translation, editing, and readability improvement. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final submission.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide complete details of our method and experiments at multiple levels. The Structure-Aligned DPO training objective and the three structural inconsistency metrics (d_{VPE}, d_{CD}, d_{Flow}) are formally defined in Sections 3.2 and 3.3, with a full diffusion-DPO objective in Appendix A. All hyperparameters including LoRA rank and scaling factor, learning rate, β, λ_{sft}, sigma distribution, and batch configuration are reported in Appendix B. Also, preference dataset construction, evaluation protocols, and per-task results for all 23 RoboCasa tasks are provided in Appendix B. For the real-world experiments, hardware specifications, camera calibration procedure, calibration for the wrist camera, and preference-pair construction are detailed in Appendix D.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 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, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. *Advances in Neural Information Processing Systems*, 38:133919–133951, 2026.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, et al. " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. Pre-trained video generative models as world simulators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4645–4653, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, Sergey Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18399–18410, 2025.

Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training Newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Hyundo Lee, Suhyung Choi, Inwoo Hwang, and Byoung-Tak Zhang. Towards spatially consistent image generation: On incorporating intrinsic scene properties into diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5863–5871, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. *arXiv preprint arXiv:2406.04314*, 2024.

Shubo Lin, Xuanyang Zhang, Wei Cheng, Weiming Hu, Gang Yu, and Jin Gao. Mmphysvideo: Scaling physical plausibility in video generation via joint multimodal modeling. *arXiv preprint arXiv:2604.02817*, 2026.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. VideoDPO: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. *arXiv preprint arXiv:2508.15720*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, 38:95597–95621, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Peiyao Wang, Weining Wang, and Qi Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. *arXiv preprint arXiv:2511.03997*, 2025.

Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2024. URL `https://arxiv.org/abs/2309.13037`.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. *arXiv preprint arXiv:2510.11512*, 2025.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

A Derivation of the Diffusion-DPO Objective

For completeness, we summarize how the DPO objective is adapted to diffusion models, following Wallace et al. (2024). The original DPO objective Rafailov et al. (2023) aligns a policy π_θ against a reference π_{ref} from preference triplets (c, y_w, y_l):

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid c)}{\pi_{\text{ref}}(y_w \mid c)} - \beta\log\frac{\pi_\theta(y_l \mid c)}{\pi_{\text{ref}}(y_l \mid c)}\right)\right], \tag{8}$$

where σ is the logistic function and β controls the divergence from the reference policy. This formulation imposes a negative margin on dispreferred (here, unphysical) outputs, restricting the generative process to a physically plausible manifold and bridging the executability gap without relying on extensive real-world robot data.

To transfer this to the diffusion denoising setting Ho et al. (2020), we define the implicit reward for a generated video $\hat{\mathbf{X}}$ as the log-likelihood ratio between p_θ and p_{ref}:

$$r(\hat{\mathbf{X}} \mid \mathbf{x}_0, c) = \beta\log\frac{p_\theta(\hat{\mathbf{X}} \mid \mathbf{x}_0, c)}{p_{\text{ref}}(\hat{\mathbf{X}} \mid \mathbf{x}_0, c)}. \tag{9}$$

Since the exact likelihood is intractable, it is approximated via the denoising score matching objective. For a noised sample $\hat{\mathbf{X}}_\sigma = \hat{\mathbf{X}} + \sigma\boldsymbol{\epsilon}$:

$$-\log p_\theta(\hat{\mathbf{X}} \mid \mathbf{x}_0, c) \approx \ell_\theta(\hat{\mathbf{X}}, \sigma) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\omega(\sigma)\left\|\boldsymbol{\epsilon}_\theta(\hat{\mathbf{X}}_\sigma, \sigma) - \hat{\mathbf{X}}\right\|^2\right], \tag{10}$$

where $\omega(\sigma)$ is the loss weighting from the noise schedule. Substituting these approximations into Equation (8) yields the diffusion DPO loss used in the main text (Equation (7)).

B Detailed Experimental Setup

Environment Details. The RoboCasa Nasiriany et al. (2024) benchmark, built upon robosuite Zhu et al. (2020), provides procedurally generated kitchen environments with diverse object configurations and textures. The 24 candidate tasks consist of *Pick-and-Place* (8 tasks) and *Non-PnP* tasks (Door/Drawer Manipulation, Appliance Control, Faucet/Sink, 16 tasks). One *Non-PnP* task (*OpenDoubleDoor*) is excluded from evaluation as its completion horizon exceeds our 480-step cap, leaving 15 evaluated Non-PnP tasks. Each task is evaluated across multiple unseen layout-and-style configurations. The video model takes three RGB cameras (*robot0_agentview_left*, *robot0_agentview_right*, and *robot0_eye_in_hand*) at 256×256 resolution and predicts $T{=}25$-frame latent rollouts: one conditioning frame plus 8 keyframes per camera.

Preference Dataset Construction. The base Video Policy (VP) Liang et al. (2025) model was previously trained via SFT for 369,000 steps on RoboCasa using 50 human demonstrations per task. To construct the structural preference dataset S_{DPO}, we deploy this VP model to generate four task rollouts per environment configuration used during VP training, each initialized with a different random model seed. Each VP cycle outputs a 25-frame latent (1 conditioning frame plus 8 keyframes for each of the three viewpoints) and a corresponding action chunk; the simulator rollout under that action chunk, VAE-encoded to match the generator's latent layout, serves as the *winning* latent $\mathbf{X}^*$, while the generator's own latent serves as the *losing* sample. A pair is retained in S_{DPO} whenever any of $d_{\mathrm{VPE}}, d_{\mathrm{CD}}, d_{\mathrm{Flow}}$ lies above the 0.8 quantile of its empirical distribution.

DPO Hyperparameters. For Structure-Aligned DPO, we insert LoRA adapters into the cross-attention projections $\{W_Q, W_K, W_V, W_O\}$ and the feed-forward layers (`ff.net.0.proj` and `ff.net.2`) of the denoising U-Net, with rank $r = 64$ and scaling factor $\alpha = 64$ (60.1M trainable parameters, 3.95% of the 1.52B base model). Training uses Adam with learning rate 5×10^{-6}, batch size 1 with gradient accumulation 4 across 4 GPUs (effective batch 16), bf16-mixed precision, and gradient clipping at 1.0 for 1,000 steps. Reference-model log-likelihoods are obtained by toggling LoRA off in the same forward pass. We set $\beta = 2.0$, $\lambda_{\mathrm{sft}} = 2.0$, and shift the EDM sigma distribution from $(\sigma_\mu, \sigma_\sigma) = (1.0, 1.6)$ to $(0.0, 1.0)$ to concentrate updates on late-stage denoising.

Evaluation Setups. We modify the per-cycle inference horizon of VP for evaluation throughput. The original VP rolls out 1,000 environment steps and executes 16 of every 32 predicted action steps per cycle. We instead execute all 24 predicted action steps and cap the rollout at 480 environment steps (20 cycles). The reduced budget excludes the *OpenDoubleDoor* task, whose nominal completion horizon exceeds 480 steps and which is unsolvable under this cap by either model; we therefore evaluate the remaining 23 tasks (8 PnP, 15 Non-PnP). For the baseline we generate 5 layout/style configurations $\times$ 6 demonstrations per configuration with deterministic policy seed 42, yielding 30 trials per task and 690 trials in total. Struct-DPO is evaluated on a paired subset of 5 configurations $\times$ 6 demonstrations per configuration, yielding 30 trials per task and 690 trials in total.

For physical-grounding metrics, d_{VPE} is computed with Video Depth Anything (ViT-B, metric-depth, input size 518). Per-view depth scales are calibrated on the first conditioning frame via grid search with inlier tolerance 0.2, then applied to all 6 subsampled keyframes per cycle (subsample factor 4, so 24 raw simulator frames yield 6 depth frames). d_{CD} uses Canny edges of the hand-camera view with the symmetric distance-transform Chamfer distance of Equation (5), and d_{Flow} uses RAFT to estimate the cross-modal displacement between the simulator render and the generated frame at the same timestep. FVD is computed per camera view with an I3D backbone over 24-frame chunks and averaged across the three views.

C Detailed Experimental Results

Task-wise breakdown for Pick-and-Place. Per-task success rates across all 23 tasks are reported in Table 5. To complement the per-task aggregates of Table 5, we decompose Pick-and-Place execution into three sequential milestones (*Grasp*, *Transport*, *Final Task Success*) and report the per-task breakdown in Table 6. The improvements introduced by Structure-Aligned DPO accumulate consistently across phases ($+2.08$ on Grasp, $+2.50$ on Transport, $+2.92$ on Final), reproducing the main paper Table 2 aggregate at per-task resolution. Phase-wise values are aggregated over the 30

Table 5: Per-task success rates over the 23 evaluated RoboCasa tasks (*OpenDoubleDoor* excluded). Baseline and Struct-DPO (*PnP Only*) aggregates 30 trials per task (5 layout/style configurations $\times$ 6 demonstrations). Subtotals reproduce the aggregates in Table 1. Bold marks the larger of each pair.

Task	**Baseline (%)**	**Struct-DPO (%)**
Pick-and-Place (8 tasks)		
PnPCabToCounter	10.00	**13.33**
PnPCounterToCab	43.33	**56.67**
PnPCounterToMicrowave	**43.33**	30.00
PnPCounterToSink	**46.67**	**46.67**
PnPCounterToStove	**40.00**	33.33
PnPMicrowaveToCounter	13.33	**26.67**
PnPSinkToCounter	70.00	**73.33**
PnPStoveToCounter	46.67	**50.00**
PnP Subtotal	39.17	**41.25**
Non-Pick-and-Place (15 tasks)		
CloseDoubleDoor	73.33	**86.67**
CloseDrawer	73.33	**93.33**
CloseSingleDoor	**100.00**	93.33
OpenDrawer	30.00	**46.67**
OpenSingleDoor	46.67	**50.00**
CoffeePressButton	93.33	**96.67**
CoffeeServeMug	73.33	**76.67**
CoffeeSetupMug	13.33	**33.33**
TurnOffMicrowave	93.33	**100.00**
TurnOnMicrowave	**100.00**	**100.00**
TurnOffStove	3.33	**6.67**
TurnOnStove	**20.00**	**20.00**
TurnOffSinkFaucet	**60.00**	53.33
TurnOnSinkFaucet	**73.33**	50.00
TurnSinkSpout	40.00	**43.33**
Non-PnP Subtotal	59.56	**63.33**
Overall	52.46	**55.65**

DPO-paired trials per task (the demonstration subset used for Struct-DPO in Table 5); the baseline column therefore differs slightly in absolute level from the 30-trial baseline of Table 5 but shows the same trend.

Alternative preference by epipolar consistency To further investigate performance gains arise specifically from our proposed structural preference or from geometric consistency signals in general, we conducted an additional experiment using epipolar consistency as the sole preference-selection criterion Kupyn et al. (2025).

Specifically, instead of constructing the DPO dataset from the union of high-error samples under viewpoint consistency (d_{VPE}), Chamfer distance (d_{CD}), and optical-flow displacement (d_{Flow}), we computed the Sampson distance between corresponding observations across camera views and selected 0.8 quantile with the highest error samples according to the resulting epipolar-consistency error. The remainder of the DPO training pipeline is kept identical to the main experiments.

Table 7 reports the task success rates. We evaluated both alignment on all tasks and alignment restricted to the Pick-and-Place subset, following the same protocol as Table 1.

The results indicate that epipolar consistency provides a useful geometric signal for identifying physically inconsistent generations, improving performance over the baseline Video Policy. However, its gains transfer less effectively compared to original Struct-DPO method, resulting in a lower success rate than Struct-DPO. We attribute this difference to the fact that epipolar consistency primarily captures cross-view geometric agreement, whereas our structural preference combines complementary signals from 3D viewpoint consistency, boundary alignment, and motion displacement. Together, these signals expose a broader range of physical violations and provide a more transferable alignment signal across task categories.

Table 6: Granular phase-wise success rates for Pick-and-Place tasks. Success rates are reported over 30 executions for each of the eight tasks. Decomposing the execution trajectory reveals that Structure-Aligned DPO yields particularly strong improvements in the initial grasping phase, confirming its superior understanding of spatial layout and contact dynamics.

Task	Method	Grasp (%)	Transport (%)	Final Success (%)
PnPCabToCounter	Baseline	33.33	20.00	10.00
	DPO	33.33	**23.33**	**13.33**
PnPCounterToCab	Baseline	50.00	43.33	43.33
	DPO	**70.00**	**56.67**	**56.67**
PnPCounterToMicrowave	Baseline	**76.67**	**60.00**	**43.33**
	DPO	73.33	43.33	30.00
PnPCounterToSink	Baseline	56.67	53.33	43.33
	DPO	56.67	**56.67**	**46.67**
PnPCounterToStove	Baseline	**73.33**	**50.00**	**40.00**
	DPO	63.33	40.00	33.33
PnPMicrowaveToCounter	Baseline	33.33	13.33	13.33
	DPO	**43.33**	**26.67**	**26.67**
PnPSinkToCounter	Baseline	80.00	70.00	70.00
	DPO	**90.00**	**73.33**	**73.3**
PnPStoveToCounter	Baseline	**80.00**	53.33	43.33
	DPO	70.00	**63.33**	**50.00**
Total	Baseline	60.42	45.42	38.33
	DPO	**62.50 (+2.08)**	**47.92 (+2.50)**	**41.25 (+2.92)**

Table 7: Alternative preference construction using epipolar consistency. 20 demos per task.

Method	PnP (%)	Non-PnP (%)	Overall (%)
Video Policy	36.50	58.53	50.87
+Epipolar-DPO (All)	**41.88**	58.44	52.92
+Epipolar-DPO (PnP Only)	41.25	57.19	51.88
+Struct-DPO (PnP Only, Ours)	41.25	**63.33**	**55.65**

D Real-World Experiment Details

Setup. We deploy Structure-Aligned DPO on a UR5e arm fitted with a Robotiq 2F-85 gripper, observed by three Intel RealSense D405 RGB cameras placed at the left, right, and wrist positions. This mirrors the three-view conditioning used in simulation, with the wrist camera taking the hand-centric role. On the table, we place a red cup and a mint-colored plate alongside a blue cup and a blue plate; successful execution requires physically consistent grasping, transport, and placement.

Using Gello Wu et al. (2024) teleoperation, we collect 50 demonstrations, each recorded at 30 Hz as a 7-DoF trajectory (six arm joints and a continuous gripper). All real-domain training is performed on two NVIDIA RTX A6000 GPUs.

Training and metric adaptation. We first adapt the video policy to the real domain through the standard two-stage SFT (6 epochs, roughly 90000 steps), then attach the Structure-Aligned DPO LoRA to the Stage-1 video U-Net (rank 64, $\alpha = 64$, $\beta = 2.0$, $\lambda_{\text{sft}} = 2.0$, learning rate 5×10^{-6}, effective batch size 8). The same stopping rule governs every model (the SFT baseline and both DPO variants alike): we train until the training loss stops decreasing rather than fixing a step budget, and given the limited real-world data we do not hold out a separate validation set. Because all three models share this loss-based stopping rule, the subset model's higher success rate—achieved with fewer pairs and far fewer optimizer steps—reflects the sufficiency of geometrically selected pairs rather than an artifact of unequal training.

We construct preference pairs from SFT baseline deployment data: for every captured cycle the real observation is the *winning* sample and the model's own generation is the *losing* sample, with hindsight relabeling applied to unsuccessful trajectories as in Section 3.3.

Evaluating d_{VPE} on hardware additionally requires the inter-camera geometry that the simulator otherwise supplies. We recover it once through calibration (per-view intrinsics from a ChArUco target, the fixed left/right extrinsics from a stereo solve, and the wrist-camera-to-tool transform from hand–eye calibration) and reconstruct the moving wrist-camera pose per frame from the robot's forward kinematics, yielding the relative rotations and translations that the depth-reprojection error requires.

E Limitations and Future Work

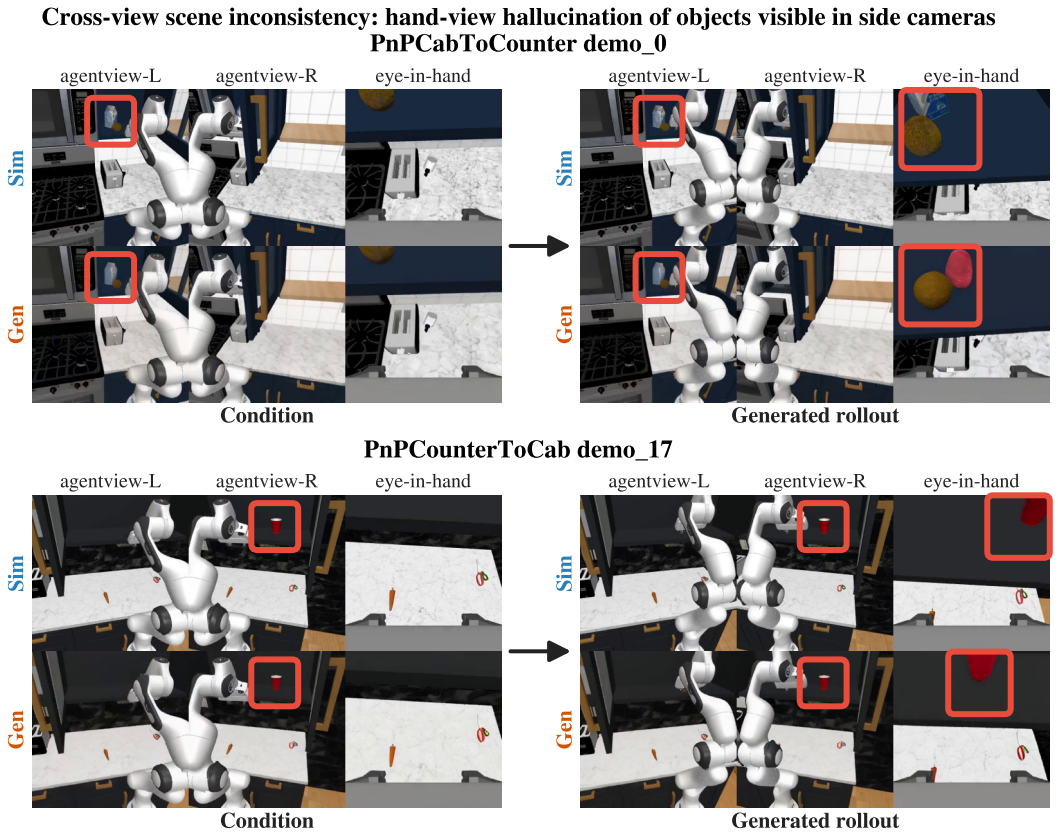


Figure 6: Failure cases not corrected by Structure-Aligned DPO. In each row, the left panel shows the conditioning frame and the right panel shows a later generated frame from the same rollout. The hand camera (rightmost sub-panel of each cell) hallucinates an object with incorrect shape and position as it enters the field of view, even though the side cameras already depict the same object consistently. Because our DPO LoRA concentrates updates on late-stage denoising (Section 3.3), it corrects fine-grained contact errors but does not recover such coarse, scene-level structure.

Backbone and denoising-stage coverage. Our experiments are conducted on a single video-policy backbone (SVD-based VP), and the proposed alignment is realized through LoRA on top of an SFT-initialized model rather than full-parameter training. Combined with the sigma-sampling adjustment that concentrates DPO updates on late-stage denoising (Section 3.3), this means our framework is most effective at correcting fine-grained, contact-level errors such as positional offsets during grasping, while coarser, low-frequency failures—e.g., an object being hallucinated at an entirely wrong location or appearing inconsistently across viewpoints at the scene level—remain largely uncorrected (Figure 6). Extending Structure-Aligned DPO to additional backbones and to earlier denoising stages where global scene layout is determined is an important direction for future work.

Inference-time cost of video generation. Our pipeline still requires generating full video rollouts to serve as visual subgoals, which limits its applicability to high-frequency embodied control. Bypassing or amortizing this iterative video synthesis at inference time—e.g., through distillation or direct action prediction conditioned on the aligned representation—would be needed to support real-time deployment.

Visual-only preference signals. Our structural preference signals are derived purely from visual modalities (depth, edges, optical flow) and therefore capture only what is observable from RGB cameras. Incorporating non-visual physical feedback such as contact forces, tactile signals, or object mass would provide a more complete view of physical interactions and could further tighten the alignment between generated videos and the underlying dynamics.