WORLDALIGN: DECOUPLED 4D REWARD FOR WORLD-CONSISTENT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

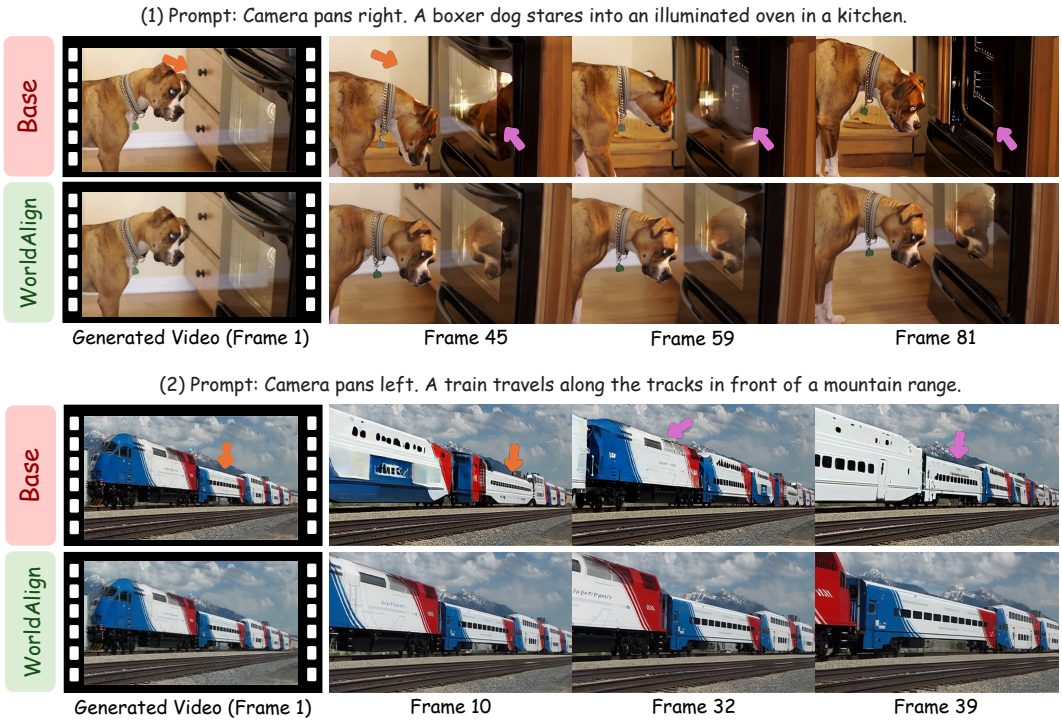

Figure 1: Comparison of WorldAlign and the base model (Wan2.1-I2V-14B-480P). The first case highlights improvements in static consistency: the base model loses the background cabinetry (orange arrows), while the oven handle disappears and its door deforms before vanishing (magenta arrows). The second case highlights improvements in dynamic consistency: the train cars abruptly change color and shape and move opposite to the locomotive's heading (see the cars marked by same-colored arrows). WorldAlign instead preserves static scene structure and plausible dynamic evolution. If the videos fail to play, **we strongly recommend using a multimedia-enabled PDF reader**. See the supplementary material for more examples.

ABSTRACT

Faithful visual world simulation requires generated videos to maintain 4D world consistency: static environments should preserve coherent 3D structure across viewpoints (static consistency), while dynamic subjects should exhibit plausible motion and consistent appearance over time (dynamic consistency). Existing approaches either modify the generator architecture to incorporate explicit geometry or improve pretrained models through geometry-aware post-training. The former may compromise the broad generalization acquired through large-scale pretraining. The latter preserves this valuable generalization capability but remains unreliable and limited in dynamic scenes. These methods either assume fully static scenes or accommodate dynamic scenes but risk conflating legitimate subject motion with undesired distortions in static regions when evaluating static consistency. Dynamic consistency is either overlooked or evaluated solely through appearance preservation. These limitations prevent effective improvements in world consis-

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.

tency in real-world dynamic scenes. In this paper, we introduce *WorldAlign*, a decoupled 4D reward framework that semantically separates static regions and dynamic subjects and aligns each with a world prior suited to its assumptions. For static regions, WorldAlign aligns static geometry with a geometric world prior through masked reprojection, excluding dynamic subjects while retaining undesired distortions in static regions for evaluation; an auxiliary camera-motion reward discourages nearly static solutions. For dynamic subjects, WorldAlign uses a strong vision-language model as a dynamic world prior, with sample-specific checklists assessing dynamicity, physical plausibility, shape, and texture consistency. This decoupled design enables online post-training without modifying the generator architecture or requiring human preference annotations. Across two pretrained image-to-video generators, WorldAlign jointly improves static and dynamic consistency over existing methods without suppressing overall or subject motion. These results demonstrate the effectiveness of decoupled world-prior alignment for advancing video generation toward faithful visual world simulation.

1 INTRODUCTION

Faithful visual world simulation requires more than frame-wise realism: the generated world must remain coherent over time. We refer to this property as world consistency, which comprises static and dynamic consistency over time. Static consistency requires static environments to maintain coherent 3D structure across viewpoints. Dynamic consistency requires subjects to move plausibly while preserving their appearance over time. Such consistency is essential when video generation models serve as world simulators for robotics (Shan et al., 2026; Zhang et al., 2026), autonomous driving (Hu et al., 2023; Wang et al., 2024; Gao et al., 2024), and interactive games (Bruce et al., 2024; Sun et al., 2025; Team et al., 2026). Despite substantial advances in high-fidelity, open-domain video synthesis (Brooks et al., 2024; Seedance et al., 2026; Team et al., 2025; Kong et al., 2024; Wan et al., 2025), large-scale pretrained video generation models remain insufficient for faithful world simulation. A key reason is that their pretraining objectives primarily optimize prediction in RGB or latent space, leaving 3D structure and physical dynamics to be learned only implicitly, without explicit world-consistency constraints. Consequently, static structures may slide, warp, or flicker across viewpoints, while dynamic subjects may exhibit implausible motion and appearance changes, as shown in Fig. 1. Achieving world consistency therefore remains a central challenge for faithful visual world simulation.

Existing efforts toward world-consistent video generation broadly follow two routes. The first modifies the generator architecture to integrate explicit geometric information, including point-conditioned generation (Ren et al., 2025; Yu et al., 2024; Liu et al., 2026a; Zhao et al., 2026) and joint RGB-geometry modeling (Zhang et al., 2025; Bai et al., 2026; Dai et al., 2025; Chen et al., 2025). Although effective, the additional modules and geometric supervision may compromise the broad generalization inherited from large-scale video pretraining. The second preserves the pretrained generator and improves consistency through post-training alignment, allowing expert knowledge or world priors to provide feedback while retaining its broad generative capability. However, handling dynamic content requires deciding where geometric constraints apply and how subject evolution should be evaluated. Epipolar-DPO, VideoGPA, and the geometric rewards of World-R1 rely on static-scene assumptions (Kupyn et al., 2025; Du et al., 2026; Wang et al., 2026). VGGRPO and GeoFlow instead use motion cues to filter regions or weight static consistency scores (An et al., 2026; Ackermann et al., 2026), but these cues can also exclude or downweight static content undergoing undesired distortion, weakening the static-consistency signal. Dynamic feedback remains limited: VGGRPO omits the filtered subjects, while GeoFlow only evaluates appearance consistency. Neither directly assesses whether subject motion is physically plausible. The challenge is therefore to retain geometric constraints on inconsistent static regions while evaluating dynamic subjects beyond appearance preservation.

Our design assigns geometric and dynamic criteria to the regions where their assumptions apply. Static regions are evaluated under rigid multi-view geometry, whereas dynamic subjects require judgments informed by their properties and scene context. We implement this principle in a decoupled 4D reward framework, *WorldAlign*, shown in Fig. 2. Specifically, WorldAlign semantically segments and tracks dynamic subjects to identify the static and dynamic regions of each generated

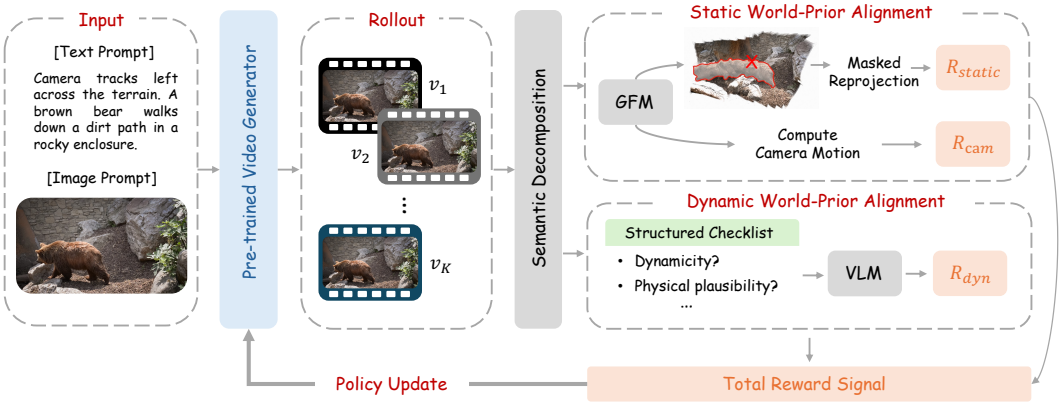

Figure 2: Overview of the proposed WorldAlign. Given generated videos, WorldAlign semantically segments and tracks dynamic subjects to decouple each video into static and dynamic regions. Static regions are aligned with a geometric world prior through masked reprojection and camera-motion rewards, while dynamic subjects are aligned with a VLM-based world prior through a structured *vlm-as-a-judge* reward. The reward signals are normalized within each same-prompt sample group and aggregated for online policy update.

video. For static regions, WorldAlign uses a geometry foundation model (GFM) (Wang et al., 2025) to assess whether different frames represent the same static 3D scene. The GFM reconstructs a shared point cloud from the frames and projects it back into each frame. When static structures remain consistent, the reprojected frames closely match the originals. When they slide, warp, or flicker, the point cloud cannot reproduce all frames accurately, resulting in a larger reprojection error. However, independently moving subjects can also increase this error, producing motion-induced artifacts that contaminate the static-consistency signal, as shown in Fig. 3. WorldAlign therefore constructs the shared point cloud using only points outside the tracked dynamic masks and computes the reprojection error over static regions. The resulting masked reprojection error measures static consistency and serves as the corresponding reward. Because regions are selected semantically rather than by their observed motion, undesired distortions in static regions remain part of the geometric evaluation. A camera-motion reward further prevents static collapse by preserving meaningful viewpoint changes. For dynamic subjects, a stable appearance can coexist with implausible motion. Assessing such failures requires knowledge of the subject and its environment, beyond frame-to-frame visual similarity. Humans draw on broad world knowledge to make these judgments. Inspired by this capability and recent advances in vision-language models (VLMs), WorldAlign uses a strong VLM as a dynamic world prior. To turn this prior into a training signal, we design a structured *vlm-as-a-judge* reward based on a checklist tailored to each sample. The checklist guides the VLM to assess physical motion and appearance consistency over time. The reward is the fraction of checklist items satisfied by the generated video, providing dedicated feedback for dynamic evolution. Together, these rewards enable WorldAlign to improve static and dynamic consistency jointly through post-training, without modifying the generator architecture or requiring human annotations.

To validate WorldAlign, we compare it with state-of-the-art post-training methods across two pretrained image-to-video generators using both reward-related and reward-independent metrics. Experiments demonstrate that WorldAlign improves static geometry and dynamic consistency over existing methods without suppressing camera or subject motion. These results support decoupled world-prior alignment as an effective approach to more faithful visual world simulation.

2 Related Work

2.1 Pretrained Video Generation Models

Large-scale pretrained video generation models (Brooks et al., 2024; Seedance et al., 2026; Team et al., 2025; Kong et al., 2024; Wan et al., 2025) have substantially advanced open-domain video synthesis. Their progress is driven by diffusion (Ho et al., 2020) and flow-matching objectives (Lip-

man et al., 2022; Liu et al., 2022), transformer-based architectures (Dosovitskiy et al., 2020), and large-scale video-text training. However, their pretraining objectives primarily optimize prediction in RGB or latent space, leaving 3D structure and physical dynamics to be learned only implicitly, without explicit world-consistency constraints. Consequently, world consistency remains unreliable across generations. WorldAlign therefore builds on these pretrained generators and improves their world consistency through post-training alignment.

2.2 World-Consistent Video Generation

Architecture Modifications for Geometry Integration. The first route modifies the generator architecture to integrate explicit geometric information. Point-conditioned methods augment pretrained video generators with modules that process 3D points reconstructed from input or generated frames (Ren et al., 2025; Yu et al., 2024; Liu et al., 2026a; Zhao et al., 2026). Joint RGB-geometry methods instead extend the generator to predict RGB appearance together with geometric representations such as depth or 3D coordinates (Zhang et al., 2025; Bai et al., 2026; Dai et al., 2025; Chen et al., 2025). Although effective, these methods introduce geometry-specific modifications that may compromise the broad generalization acquired through large-scale pretraining, which is costly to reproduce and important to retain. WorldAlign instead leaves the pretrained generator architecture unchanged, preserving its broad generative capability.

Geometry-Aware Post-Training. The second route preserves the pretrained generator and introduces geometric feedback through post-training alignment. Epipolar-DPO uses epipolar consistency to construct preference pairs (Kupyn et al., 2025). VideoGPA derives preference pairs from reprojection errors computed using predicted geometry and camera poses, while filtering out pairs with insufficient camera motion (Du et al., 2026). Both methods are designed under a static-scene assumption and therefore provide geometric feedback only for static scenes. World-R1 combines reconstruction, meta-view, and trajectory-alignment terms into a 3D-aware reward, alongside a general visual reward (Wang et al., 2026). To prevent overfitting to static rigidity, it periodically disables the 3D-aware reward and optimizes dynamic prompts using only the general visual reward. Although this strategy preserves motion diversity, the general reward does not explicitly evaluate dynamic consistency. VGGRPO accommodates dynamic scenes by using scene flow to filter out moving regions before applying geometric rewards to the remaining static regions (An et al., 2026). Because this filtering is motion-based rather than semantic, it may also remove inconsistent static content exhibiting warping or flickering, thereby weakening the static-consistency signal. Moreover, the filtered dynamic regions receive no consistency reward. Similarly, GeoFlow (Ackermann et al., 2026) uses optical-flow residuals to distinguish static and dynamic regions. However, regions identified as dynamic may include both legitimate subject motion and undesired distortions in static regions. Downweighting these regions during depth-consistency evaluation therefore weakens the static-consistency signal, leaving such distortions insufficiently penalized. Its dynamic feedback measures DINO-based appearance consistency across the whole scene after flow alignment, without explicitly assessing motion plausibility.

WorldAlign differs in both region decomposition and evaluation criteria: tracked semantic masks determine where static geometry is assessed, and subject-specific checklists evaluate dynamicity, physical plausibility, shape, and texture consistency.

3 Method

3.1 Method Overview

We consider image-to-video generation[1], where a pretrained generator produces an N-frame video $v = \{I_n\}_{n=1}^{N}$ conditioned on an image I^c and a text prompt y. Our goal is to improve the generator's world consistency through post-training alignment without modifying its architecture.

Real-world scenes contain both static environments and dynamic subjects, which satisfy different consistency assumptions. WorldAlign therefore separates them and aligns each with a suitable world

[1]WorldAlign can be extended to text-to-video generation by obtaining subject masks from generated videos and constructing checklists from text prompts.

prior, as illustrated in Fig. 2. A subject mask M^c on the conditioning image initializes a tracker that propagates the subject throughout the generated video, producing dynamic masks $M = \{M_n\}_{n=1}^{N}$. Their complements $\overline{M}_n$ define the static regions. For static regions, WorldAlign uses a masked reprojection reward R_{static} together with an auxiliary camera-motion reward R_{cam}. For dynamic subjects, it uses a structured *vlm-as-a-judge* reward R_{dyn}.

3.2 Static World-Prior Alignment

Masked Reprojection Reward. Static regions approximately satisfy rigid multi-view geometry. We therefore use a geometry foundation model (GFM) (Wang et al., 2025) to measure their consistency through cross-view reprojection. Given a generated video v, the GFM predicts a pixel-aligned 3D point map P_n, camera intrinsics K_n, and camera extrinsics C_n for each frame I_n.

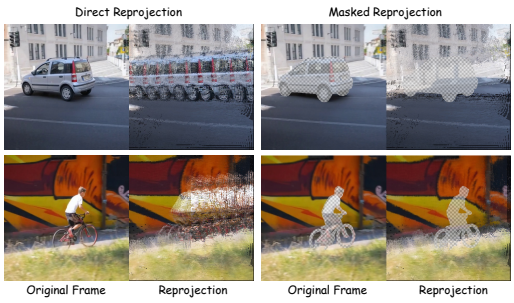

Figure 3: Motion-induced artifacts of Direct Reprojection. Our masked reprojection can isolate static geometry from subject motion, yielding a more reliable static-consistency reward.

However, directly aggregating the predicted point maps also fuses the dynamic subject at its successive positions into the shared point cloud, forming a 3D trace of its motion. Reprojecting this point cloud into each frame produces motion-induced artifacts. Pixels affected by these artifacts yield large reprojection errors regardless of whether the corresponding static regions are geometrically consistent, making direct reprojection unreliable for static-consistency evaluation, as shown in Fig. 3.

To remove this interference, WorldAlign constructs the shared point cloud using only static regions. Because P_n is pixel-aligned, the tracked mask M_n can be applied directly to exclude 3D points corresponding to the dynamic subject:

$$\mathcal{P}_{\text{static}} = \left\{ P_n(x) \mid x \in \overline{M}_n,\ n = 1, \ldots, N \right\}. \tag{1}$$

The resulting point cloud excludes the subject's motion trajectory and retains only the 3D points assigned to static regions. Because dynamic regions are defined by tracked subject masks rather than motion-based thresholds, undesired distortions in static regions remain subject to geometric evaluation wherever reprojection is valid. WorldAlign then projects $\mathcal{P}_{\text{static}}$ into each frame using the predicted camera parameters: $\hat{I}_n = \Pi\left(\mathcal{P}_{\text{static}}; K_n, C_n\right)$, where Π denotes the reprojection operation. We exclude dynamic-subject pixels from the comparison because they would otherwise compare the reprojected static background with the foreground subject visible in I_n. We therefore define the valid comparison region as

$$\Omega = \left\{ (n, x) \mid x \in \overline{M}_n,\ \hat{I}_n(x) \text{ is valid} \right\}. \tag{2}$$

The masked reprojection error and its corresponding static reward are

$$E_{\text{rep}} = \frac{1}{|\Omega|} \sum_{(n,x)\in\Omega} \left\| I_n(x) - \hat{I}_n(x) \right\|_2^2, \quad R_{\text{static}} = -E_{\text{rep}}. \tag{3}$$

Camera-Motion Reward. The masked reprojection reward admits a trivial solution: suppressing viewpoint changes can reduce the reprojection error without requiring consistent 3D structure. To prevent this collapse, WorldAlign reuses the camera extrinsics predicted by the GFM to construct an auxiliary reward that penalizes insufficient camera motion.

Let $C_n = (R_n, t_n)$ denote the predicted camera extrinsics, with rotation R_n and translation t_n. We define the camera-motion magnitude as $m(C_{1:N}) = \bar{d}_{\text{trans}} + \bar{d}_{\text{rot}}$, where $\bar{d}_{\text{trans}}$ and $\bar{d}_{\text{rot}}$ denote the mean adjacent-frame translation $\|t_{n+1} - t_n\|_2$ and rotation angle $\theta(R_{n+1}R_n^\top)$, respectively. The camera-motion reward applies a hinge penalty when the predicted motion falls below a threshold:

$$R_{\text{cam}} = -\max\left(0, \tau - m(C_{1:N})\right), \tag{4}$$

where τ is the lower-bound camera-motion threshold. This reward penalizes videos with insufficient viewpoint changes, preventing the generator from reducing the reprojection error solely through nearly static viewpoints.

3.3 Dynamic World-Prior Alignment

WorldAlign uses a strong VLM as a dynamic world prior and translates its world knowledge into a structured *vlm-as-a-judge* reward through subject-specific binary questions.

Before training, we use a VLM to construct a sample-specific checklist $\mathcal{Q} = \{q_j\}_{j=1}^{J}$ from each conditioning pair (I^c, y). The checklist consists of subject-centric binary questions tailored to the subject and its intended motion, covering four complementary aspects of dynamic consistency:

- **Dynamicity**: whether the subject exhibits the intended motion rather than remaining static.
- **Physical plausibility**: whether the motion follows physically plausible dynamics given the subject's intrinsic properties and the surrounding environment.
- **Shape**: whether the subject preserves its structure over time while allowing physically plausible articulation.
- **Texture**: whether the subject preserves its defining visual attributes, such as color, patterns, and surface details.

Together, the shape and texture criteria prevent appearance drift that could make the subject appear as a different entity. The sample-specific checklist accounts for subject- and context-dependent dynamics, while its binary format decomposes an open-ended judgment into explicit criteria for more reliable evaluation.

During training, the judging VLM receives the full generated video v and the checklist $\mathcal{Q}$. The full video allows it to consider the surrounding scene, while the evaluation instruction directs it to focus on the dynamic subject and answer each question with YES or NO. Let $a_j \in \{0, 1\}$ denote the answer to question q_j, with YES mapped to 1. We define the *vlm-as-a-judge* reward as

$$R_{\mathrm{dyn}} = \frac{1}{J} \sum_{j=1}^{J} a_j. \tag{5}$$

Thus, R_{dyn} measures the fraction of checklist criteria satisfied by the generated video.

3.4 Online Policy Update

We optimize the pretrained generator using online GRPO following Flow-GRPO (Liu et al., 2026b). For each conditioning pair (I^c, y), the generator samples a group of K videos $\{v_i\}_{i=1}^{K}$. For every v_i, WorldAlign computes the three reward components R_{static}, R_{cam}, and R_{dyn}.

To obtain group-relative signals while balancing their different scales, we normalize each reward component separately within the group:

$$\widehat{R}_k(v_i) = \frac{R_k(v_i) - \mu_k}{\sigma_k + \epsilon}, \qquad k \in \{\mathrm{static}, \mathrm{cam}, \mathrm{dyn}\}, \tag{6}$$

where μ_k and σ_k are the group mean and standard deviation of reward k, and ϵ is a small constant for numerical stability. The normalized components are aggregated as the advantage:

$$A_i = \lambda_s \widehat{R}_{\mathrm{static}}(v_i) + \lambda_c \widehat{R}_{\mathrm{cam}}(v_i) + \lambda_d \widehat{R}_{\mathrm{dyn}}(v_i), \tag{7}$$

where λ_s, λ_c, and λ_d control the contributions of the three reward components. The resulting advantage is used with the standard clipped GRPO objective to update the generator. The tracker, GFM, and VLM remain frozen and are used only for reward computation.

4 Experiments

4.1 Experimental Setup

Datasets. WorldAlign focuses on real-world scenes in which static environments and dynamic subjects coexist. Existing geometry-alignment datasets, such as DL3DV (Ling et al., 2024), contain

only static scenes and therefore cannot support dynamic consistency alignment. We thus construct training and evaluation data from DAVIS (Pont-Tuset et al., 2017) and COCO (Lin et al., 2014). Each sample comprises a conditioning image, a subject mask for the conditioning image, an image-to-video prompt, and a sample-specific dynamic-consistency checklist. The resulting dataset contains 1,389 samples, of which 100 are randomly held out for evaluation and the remaining 1,289 are used for training. Further details on data selection, prompt generation, and checklist construction are provided in the supplementary material.

Baselines. We evaluate WorldAlign on Wan2.2-TI2V-5B and Wan2.1-I2V-14B-480P (Wan et al., 2025), using each pretrained generator as a base model, and compare against Epipolar-DPO (Kupyn et al., 2025), VideoGPA (Du et al., 2026), World-R1 (Wang et al., 2026), and GeoFlow (Ackermann et al., 2026). For a controlled comparison, all methods are trained on the same data with the same base generators and matched generation settings. Our World-R1 baseline is a reproduction adapted to the trajectory-free setting shared by all methods. It retains the reconstruction, meta-view, and general visual rewards, while omitting camera-aware noise warping and trajectory-alignment rewards because they require explicit camera trajectories. VGGRPO (An et al., 2026) is not included because its official implementation has not been publicly released.

Evaluation Metrics. We evaluate generated videos along three dimensions: static consistency, dynamic consistency, and general video quality. For static and dynamic consistency, we pair reward-related and reward-independent metrics to verify gains beyond the optimized objective. General video quality is assessed using third-party VBench metrics (Huang et al., 2024), none of which is used during training.

① *Static Consistency.* We report **Masked Epipolar Error (M.Epi.)** and **Masked Reprojection Error (M.Reproj.)**. M.Epi. is the reward-independent metric: it measures epipolar violations between static-region correspondences. M.Reproj. measures the discrepancy between original and reprojected static regions. Both exclude tracked dynamic subjects. Lower values are better. Because suppressing viewpoint changes can trivially reduce both errors, we interpret them jointly with Dynamic Degree.

② *Dynamic Consistency.* We report **Tracked-Subject DINOv3 Similarity (Subject DINO)** and the **Checklist "Yes" Rate**. Tracked-Subject DINOv3 Similarity (Siméoni et al., 2025) is the reward-independent metric and computes the mean cosine similarity of DINOv3 features in tracked subject regions across adjacent frames. Checklist "Yes" Rate measures the proportion of dynamic-consistency criteria satisfied by each video. Higher values are better. Because a frozen subject can trivially produce high DINO similarity, we interpret it jointly with Dynamicity "Yes" Rate and Dynamic Degree.

③ *General Video Quality.* Following VBench (Huang et al., 2024), we report **Dynamic Degree**, **Motion Smoothness**, and **Aesthetic Quality**, with higher values being better. Dynamic Degree measures the overall amount of motion and helps identify motion suppression, while Motion Smoothness and Aesthetic Quality assess motion continuity and frame-wise visual quality, respectively.

Implementation details are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

We compare WorldAlign with pretrained base models and geometry-aware alignment methods on static consistency, dynamic consistency, and motion preservation. Complementary **qualitative comparisons and human evaluation are provided in the supplementary material** due to space limitations; both further support the quantitative findings in Tab. 1.

As shown in Tab. 1, the pretrained base models retain moderate motion but exhibit high geometric errors and limited dynamic consistency, confirming that world consistency remains an unreliable capability of pretrained video generators. The alignment methods differ in their ability to improve consistency while preserving motion. Epipolar-DPO and VideoGPA report deceptively low geometric errors, accompanied by substantial decreases in both Dynamic Degree and Dynamicity "Yes" Rate. Their apparent geometric gains are therefore largely attributable to the suppression of camera and subject motion rather than real improvements in world consistency. VideoGPA exhibits less severe motion suppression because it filters candidates with insufficient camera motion during preference construction, whereas Epipolar-DPO has no such mechanism. Our World-R1 repro-

Table 1: Comparison with base models and state-of-the-art alignment methods on two image-to-video generators. For the Checklist "Yes" Rate, parentheses report the Dynamicity "Yes" Rate over all test videos. M.Reproj. and Checklist "Yes" Rate are reward-related; M.Epi. and Subject DINO are reward-independent. Bold and underlined entries denote the best and second-best results, respectively. Consistency metrics are interpreted jointly with Dynamic Degree because motion suppression can yield deceptively favorable scores, shown in gray. WorldAlign improves both reward-related and reward-independent consistency metrics while preserving overall and subject motion.

	Static Consistency		Dynamic Consistency		VBench		
Method	M.Epi. ↓	M.Reproj. ↓	Subject DINO ↑	Checklist "Yes" Rate ↑	Dynamic Degree↑	Motion Smooth. ↑	Aesthetic Quality ↑
Wan2.2-TI2V-5B							
Base	0.463	0.040	0.922	0.697 (0.63)	0.63	0.957	0.479
Epipolar-DPO	**0.268**	**0.022**	**0.961**	0.646 (0.43)	0.31	0.975	0.481
VideoGPA	0.308	0.024	0.948	0.652 (0.56)	0.54	0.971	0.476
World-R1	0.432	0.038	0.945	0.657 (0.61)	0.56	0.965	0.459
GeoFlow	0.367	0.031	0.934	0.736 (0.74)	0.67	0.971	0.478
WorldAlign (Ours)	0.321	0.026	0.938	**0.781 (0.85)**	**0.72**	**0.979**	**0.488**
Wan2.1-I2V-14B-480P							
Base	0.469	0.037	0.918	0.760 (0.81)	0.62	0.963	0.481
Epipolar-DPO	**0.215**	**0.019**	**0.967**	0.594 (0.53)	0.33	0.976	**0.494**
VideoGPA	0.243	0.022	0.963	0.608 (0.61)	0.42	0.970	0.489
World-R1	0.428	0.034	0.951	0.636 (0.72)	0.53	0.962	0.440
GeoFlow	0.336	0.028	0.936	0.807 (0.83)	0.65	0.975	0.485
WorldAlign (Ours)	0.264	0.024	0.939	**0.856 (0.92)**	**0.74**	**0.978**	0.491

duction provides only marginal improvements in static consistency while still suppressing subject motion, as reflected by its lower Dynamic Degree and Dynamicity "Yes" Rate. Its reconstruction and meta-view terms evaluate the reconstructed scene through perceptual and semantic similarity without isolating dynamic subjects. In dynamic scenes, subject-motion discrepancies can therefore dominate the reward, diluting the static-consistency signal and biasing optimization toward subject-motion suppression. GeoFlow improves static and dynamic consistency over the base models while increasing both Dynamic Degree and Dynamicity "Yes" Rate. However, its Checklist "Yes" Rate remains 4.5 and 4.9 percentage points below WorldAlign on Wan2.2 and Wan2.1, respectively.

Compared with both the pretrained base models and GeoFlow, WorldAlign achieves lower reward-related M.Reproj. and reward-independent M.Epi. Dynamic improvements similarly hold for the reward-related Checklist "Yes" Rate and reward-independent Subject DINO, while Dynamicity "Yes" Rate and Dynamic Degree also increase. Thus, the gains transfer beyond the optimized rewards without relying on overall or subject motion suppression. Its gains in Motion Smoothness and Aesthetic Quality further indicate the resulting dynamics remain coherent and visually plausible.

4.3 Ablation Study

Tab. 2 evaluates the reward components. Masked reprojection alone reduces geometric errors but suppresses motion; adding the Camera Reward raises Dynamic Degree from 0.42 to 0.70. The Dynamic Reward further improves Subject DINO and Checklist "Yes" Rate while retaining the geometric gains. Direct reprojection instead penalizes legitimate subject motion, creating conflicting objectives and yielding poorer static and dynamic consistency than the full masked design. The complete ablation analysis is provided in the supplementary material.

4.4 Comparison of Dynamic Region Identification

Fig. 4 compares scene-flow-based, optical-flow-based, and semantic dynamic region identification on the same videos. The motion cues used by VGGRPO (An et al., 2026) and GeoFlow (Ackermann et al., 2026) can conflate undesired background distortion with legitimate subject motion, as illustrated by the red overlays at the arrow-marked locations. Filtering these regions out of geo-

Table 2: Ablation of the proposed WorldAlign on Wan2.2-TI2V-5B. Camera and dynamic rewards are incrementally added to both the direct- and masked-reprojection branches. The results show that masked reprojection avoids imposing a static-scene assumption on the entire video and prevents the suppression of dynamic subject motion. The camera reward alleviates the static collapse observed with masked reprojection alone, while the dynamic reward further improves dynamic consistency.

	Static Consistency		Dynamic Consistency		VBench		
Variant	M.Epi. ↓	M.Reproj. ↓	Subject DINO ↑	Checklist "Yes" Rate ↑	Dynamic Degree ↑	Motion Smooth. ↑	Aesthetic Quality ↑
Base	0.463	0.0396	0.922	0.697 (0.63)	0.63	0.957	0.479
w/ Direct R. Reward	0.298	0.0223	**0.968**	0.628 (0.18)	0.24	**0.992**	0.482
+ Camera Reward	0.324	0.0272	0.963	0.632 (0.22)	0.64	0.982	0.479
+ Dynamic Reward	0.338	0.0291	0.954	0.670 (0.55)	0.69	0.974	0.477
w/ Masked R. Reward	**0.289**	**0.0215**	0.956	0.667 (0.55)	0.42	0.986	0.480
+ Camera Reward	0.317	0.0262	0.927	0.711 (0.64)	0.70	0.968	0.475
+ Dynamic Reward (Full Design)	0.321	0.0260	0.938	**0.781 (0.85)**	**0.72**	0.979	**0.488**

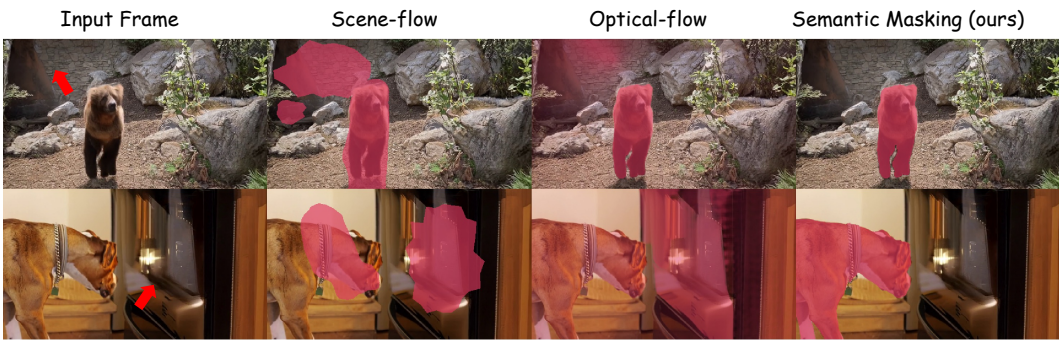

Figure 4: Qualitative comparison of dynamic region identification. The first column shows only the first frame of each video; red arrows mark background regions that distort over time and should remain subject to static consistency evaluation. Red overlays indicate regions identified as dynamic. Scene-flow and optical-flow estimates also mark distorted background regions, whereas our semantic masks more cleanly isolate the subjects.

metric evaluation or downweighting their depth-consistency contribution weakens feedback on the very background errors that should be corrected. Our semantic masks retain these regions for static evaluation while separating the dynamic subjects, supporting the choice of semantic rather than motion-based region assignment.

5 Conclusion

We introduced WorldAlign, a decoupled 4D reward framework that separates static regions and dynamic subjects and aligns each with a world prior suited to its assumptions. For static regions, a masked reprojection reward leverages a geometry foundation model to assess static consistency without interference from subject motion, while an auxiliary camera-motion reward prevents nearly static solutions. For dynamic subjects, a structured *vlm-as-a-judge* reward uses sample-specific checklists to assess physical motion and appearance consistency. These rewards enable online post-training without architectural changes or human preference annotations. Across two pre-trained image-to-video generators, WorldAlign jointly improves static and dynamic consistency on both reward-related and reward-independent metrics without suppressing overall or subject motion. These results demonstrate the effectiveness of decoupled world-prior alignment for advancing video generation toward more faithful visual world simulation.

AI USE STATEMENT

Generative AI tools were used for manuscript language polishing. The authors verified the cited references, reviewed all AI-assisted edits, and take full responsibility for the final manuscript.

REFERENCES

Jan Ackermann, Shengqu Cai, Boyang Deng, Zhengfei Kuang, Songyou Peng, and Gordon Wetzstein. GeoFlow: Enforcing implicit geometric consistency in video generation. *arXiv preprint arXiv:2605.18365*, 2026.

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *Advances in Neural Information Processing Systems*, 38:57602–57622, 2026.

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.

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 *Forty-first International Conference on Machine Learning*, 2024.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

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

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *IEEE Transactions on Image Processing*, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026b.

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.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Ziyu Shan, Zhenyu Wu, Xiaofeng Wang, Zheng Zhu, and Ziwei Wang. Dvg-wm: Disentangled video generation enables efficient embodied world model for robotic manipulation. *arXiv preprint arXiv:2606.32028*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y Chen, Zhiyuan He, et al. World-r1: Reinforcing 3d constraints for text-to-video generation. *arXiv preprint arXiv:2604.24764*, 2026.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, et al. Physisforcing: Physics reinforced world simulator for robotic manipulation. *arXiv preprint arXiv:2606.28128*, 2026.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21685–21695, 2025.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026.