Divide and Conquer: Decoupled Representation Alignment for Multimodal World Models

Junyuan Xiao[1], Dingkang Liang[2,†], Xin Zhou[2], Yixuan Ye[3], Tongtong Su[4], Guangmo Yi[1], Bin Xia[5], Qiang Lyu[6], Shurui Shi[1], Jun Huang[7], Jianlou Si[7,†], and Wenming Yang[1,⋆]

[1]Tsinghua University, China
[2]Huazhong University of Science and Technology, China
[3]CSU, China [4]ZJU, China [5]CUHK, Hong Kong SAR, China
[6]UCAS, China [7]Alibaba Group, China

xiao-jy24@mails.tsinghua.edu.cn, dkliang@hust.edu.cn

Abstract. Emerging multi-modal world models attempt to jointly generate videos across diverse modalities (*e.g.*, RGB, depth, and mask), yet they fail to fully exploit the priors of existing foundation models. We propose **M^2-REPA**, the first representation alignment method tailored for multi-modal video generation. Our key insight is that foundation models trained on different modality spaces naturally capture distinct domain-specific priors, acting as complementary "experts." Specifically, we first decouple modality-specific features from the diffusion model's intermediate representations, then align each with its corresponding expert foundation model. To this end, we design two synergistic objectives: a multi-modal representation alignment loss that enforces feature-to-expert matching, and a modality-specific decoupling regularization that encourages complementarity across different modalities. This design enables joint optimization, exploiting priors from multiple foundation models. Extensive experiments demonstrate that our method outperforms baselines in visual quality and long-term consistency.

Keywords: Multimodal World Model · Foundation Model · Video Generation · Representation Alignment

1 Introduction

World models enable agents to predict environmental dynamics and plan actions [11, 65]. Recent video diffusion models [1, 13, 22, 42, 53, 65] trained on large-scale datasets with structured inputs (*e.g.*, actions and camera movements) have emerged as promising world simulators [32, 51, 52], with applications in autonomous driving [10, 15, 36], embodied intelligence [7, 61], and interactive game engines [12, 29, 41, 57]. However, existing models operate solely on 2D RGB

† Project lead. ⋆ Corresponding author.

pixels, while the physical world is inherently 3D. This limitation poses significant challenges for 3D-aware modeling and applications requiring accurate depth estimation and multi-modal information.

Building on this motivation, a growing body of research has explored multi-modal world models [6, 17, 26, 61, 64], which can simultaneously predict videos across multiple modalities such as RGB, depth, surface normals, or optical flow. These multi-modal outputs can be directly leveraged to facilitate performance enhancement mechanisms or serve various downstream tasks [17, 26, 61]. Despite these promising advances, achieving robust and scalable multi-modal world models remains an open challenge.

Meanwhile, recent work [27, 31, 46, 56, 60] has demonstrated that leveraging high-quality representations extracted from large-scale pre-trained transformers, such as DINOv2 [28], CLIP [33], and VGGT [44], can substantially benefit video generation and downstream task performance, notably through Representation Alignment (REPA) [19, 39, 58]. REPA explicitly aligns early-stage diffusion features with clean image features from self-supervised foundation vision models, providing the generation process with useful priors. This alignment allows deeper layers to focus on high-frequency content, facilitating high-fidelity generation. However, existing methods are confined to single-modality alignment using individual foundation models, leaving multi-modal video generation largely unexplored.

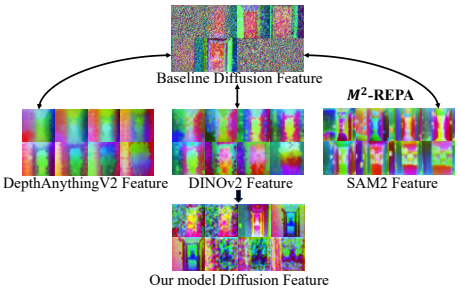

Fig. 1: Visualization of M^2-REPA. The features extracted from the backbones of DINOv2, DepthAnythingV2, and SAM2 exhibit pronounced differences, validating their distinct modality-specific characteristics. After applying our M^2-REPA, the extracted diffusion features align more closely with the semantic information from the foundation models.

A key insight driving our work stems from recognizing the fundamental nature of existing foundation models: they are inherently trained to model different modality-specific spaces. For instance, DINOv2 [28] serves as a general-purpose feature extractor trained in the RGB visual space; DepthAnythingV2 [50] specializes in depth estimation within the depth modality space; and SAM2 [34] models the mask modality space for segmentation tasks. From this perspective, each foundation model can be viewed as a modality expert with distinct domain gaps and prior knowledge specific to its training modality. The differences observed between features extracted from their backbones (see Fig. 1) provide visual evidence of this specialization, suggesting that complementary information likely exists among the prior representations from different modality-specific foundation models.

Table 1: Quantitative results for RGB-Depth video generation on RealEstate10K (8 and 200 frames). DINOv2 [28] and DepthAnythingV2 [50] excel in RGB and depth metrics, respectively. However, naively combining them for representation alignment yields marginal gains or even slight degradation compared to individual usage. **Bold** and underlined denote the **best** and second-best results.

Method	Frames	RGB modality				Depth modality	
		FVD↓	LPIPS↓	SSIM↑	PSNR↑	AbsRel↓	δ_1↑
RGB-D Baseline	8	105.00	0.3634	0.5823	19.9183	0.8747	0.2039
REPA (DINOv2 [28])	8	**81.38**	**0.2239**	0.7242	**20.1036**	0.8012	0.4239
REPA (DepthAnythingV2 [50])	8	84.12	0.2257	**0.7247**	20.1023	**0.7845**	**0.4339**
REPA (DINOv2+DepthAnythingV2)	8	84.56	0.2265	0.7236	20.0852	0.7768	0.4294
RGB-D Baseline	200	387.50	0.5222	0.3454	13.7027	1.0155	0.1094
REPA (DINOv2 [28])	200	322.25	0.4612	**0.4274**	**13.0289**	0.9930	0.1330
REPA (DepthAnythingV2 [50])	200	342.25	**0.4607**	0.4266	12.8878	**0.9813**	**0.1385**
REPA (DINOv2+DepthAnythingV2)	200	**317.75**	0.4633	0.4230	12.9625	0.9952	0.1314

This observation motivates our work to address how foundation models can be better leveraged to improve video generation performance in the presence of multi-modal outputs. We investigate ***whether simultaneously transferring complementary domain-specific priors from multiple foundation model experts across different modalities can benefit multi-modal world models.*** To validate this hypothesis, we conduct preliminary experiments on RGB-Depth (RGB-D) dual-modal video generation (Tab. 1). The results reveal a clear "modality advantage" phenomenon: models guided by DINOv2 [28] alignment exhibit superior performance in the RGB modality, while those guided by DepthAnythingV2 [50] alignment excel in depth generation. This confirms that foundation models of different modalities indeed capture unique modality-specific priors, creating opportunities for "multi-expert collaborative guidance." However, naively applying multiple foundation models simultaneously to existing REPA frameworks leads to feature conflicts that hinder training optimization, yielding negligible benefits (Tab. 1).

To address these opportunities and challenges, we propose M^2-REPA, the first representation alignment method tailored for multi-modal video generation. M^2-REPA encourages models to jointly internalize domain-specific priors from diverse foundation model experts during training. To simultaneously align these priors while circumventing feature conflicts, we first decouple the diffusion latent features into distinct modality-specific representations via learnable decoupling layers, then align each representation with its corresponding expert foundation model. We design two synergistic regularization objectives: a multi-modal representation alignment loss that enforces feature-to-expert matching via cosine similarity, and a modality-specific decoupling regularization that explicitly encourages complementarity across different modalities via Centered Kernel Alignment (CKA) [23] similarity. This design is extensible and plug-and-play, enabling joint optimization that makes use of the priors of multi-modal experts.

Extensive experiments on challenging tri-modal (RGB-Depth-Mask) generation tasks validate our method. Whether on U-Net- or Transformer-based backbones, M^2-REPA consistently outperforms single-modality baselines and naive multi-expert extensions, achieving state-of-the-art fidelity in both short-term and long-range autoregressive generation.

2 Related Work

2.1 Video Generation Model for World Simulation

Interactive Long Video Generation. Recent autoregressive video generation methods [3, 18, 21, 38, 55] show promise for generating arbitrarily long sequences as world models. Diffusion Forcing [3] employs frame-wise independent noise scheduling for autoregressive generation, while DFoT [38] extends this with variable-length historical conditioning. The availability of large-scale video datasets with structured inputs (*e.g.*, actions and camera poses) [12, 43, 62, 63] enables training controllable world models [4, 8, 12, 29, 41, 47] conditioned on external commands. We introduce a novel training pipeline that integrates into existing frameworks while leveraging multi-modal generation capabilities.
Multimodal Video Generation. Since our physical world is inherently three-dimensional, deploying world simulation and its downstream applications often necessitates information from multiple modalities. Consequently, a growing body of research has explored multi-modal world models [4, 6, 17, 26, 61, 64], which generate videos beyond just RGB. For instance, TesserAct [61] learns a 4D world model by training on RGB-DN (RGB, Depth, and Normal) videos. Aether [64] develops an RGB-Depth model and leverages both modalities for reconstruction and planning. WorldWeaver [26] jointly models RGB, depth, and optical flow to enhance long video generation. While our work shares the goal of multi-modal video world modeling, it distinguishes itself by distilling strong priors from modality-specific foundation models into video representations, simultaneously improving generation quality across modalities.

2.2 Foundation Models for Video Generation

Foundation models such as CLIP [33], DINO [28, 37], VGGT [44], DepthAnythingV2 [50], and SAM [34] learn rich, transferable representations through self-supervised learning on large-scale data, capturing robust visual semantics and geometric priors [2,20,27,31,48,54,56,58]. Recent work shows that leveraging pre-trained foundation model features substantially improves diffusion-based generation quality and consistency [19,46,48,58,60]. The REPA family [19,58,60] aligns clean foundation model features with early-stage diffusion features via regularization, while Pixel-Perfect Depth [48] directly injects features into pixel-space diffusion blocks. However, prior works primarily focus on unimodal generation with a single foundation model. By contrast, our method exploits multiple foundation models to enhance generation quality across modalities.

3 Preliminaries

3.1 Problem Formulation

Given a single initial RGB image $\mathbf{X}_{\mathrm{RGB}}^{(0)} \in \mathbb{R}^{3\times H\times W}$, along with its corresponding depth map $\mathbf{X}_{\mathrm{D}}^{(0)} \in \mathbb{R}^{1\times H\times W}$, segmentation mask $\mathbf{X}_{\mathrm{M}}^{(0)} \in \mathbb{R}^{C\times H\times W}$, and a

sequence of structured control signals $\{\mathbf{S}^{(i)}\}_{i=1}^{T}$ (*e.g.*, camera poses or actions), our goal is to generate temporally coherent tri-modal video sequences in an autoregressive manner. We formulate this as learning the conditional distribution:

$$p\left(\{\mathbf{X}_{\mathrm{RGB}}^{(i)}, \mathbf{X}_{\mathrm{D}}^{(i)}, \mathbf{X}_{\mathrm{M}}^{(i)}\}_{i=1}^{T} \,\middle|\, \mathbf{X}_{\mathrm{RGB}}^{(0)}, \mathbf{X}_{\mathrm{D}}^{(0)}, \mathbf{X}_{\mathrm{M}}^{(0)}, \{\mathbf{S}^{(i)}\}_{i=1}^{T}\right), \tag{1}$$

where $\mathbf{X}_{\mathrm{RGB}}^{(i)}$, $\mathbf{X}_{\mathrm{D}}^{(i)}$, and $\mathbf{X}_{\mathrm{M}}^{(i)}$ denote the RGB frame, depth map, and segmentation mask at frame i, with C representing the number of mask classes.

Through joint modeling of all three modalities conditioned on structured inputs, this formulation facilitates controllable generation of pixel-aligned tri-modal videos, ensuring both cross-modal spatial consistency and temporal coherence throughout the sequence.

3.2 Autoregressive Video Diffusion Models

Autoregressive video generation extends video sequences by predicting future frames conditioned on past observations. To model the conditional distribution defined in Eq. (1), we adopt Flow Matching [24, 25], a deterministic framework that constructs continuous trajectories from noise to target video distributions.

Flow Matching. Let $\mathbf{x}^{(i)} = [\mathbf{X}_{\mathrm{RGB}}^{(i)}, \mathbf{X}_{\mathrm{D}}^{(i)}, \mathbf{X}_{\mathrm{M}}^{(i)}] \in \mathbb{R}^{(3+1+C)\times H\times W}$ denote the concatenated tri-modal representation for the i-th frame. Let $\mathbf{x}_1 \in \mathbb{R}^{T\times(3+1+C)\times H\times W}$ represent the entire clean tri-modal video sequence and $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the standard Gaussian noise. The continuous flow timestep is denoted as $t \in [0, 1]$. The linear interpolation from noise to data is defined as:

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \tag{2}$$

with the corresponding target velocity field $\mathbf{v}_t = \mathbf{x}_1 - \mathbf{x}_0$. The model parameter θ is optimized by minimizing the flow matching objective:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{x}_1, \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{y}, t\in[0,1]}\left[\left\|\mathbf{v}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \mathbf{v}_t\right\|_2^2\right], \tag{3}$$

where $\mathbf{y}$ represents all conditioning inputs, including the initial context frame $\mathbf{x}^{(0)}$ and structured signals $\{\mathbf{S}^{(i)}\}_{i=1}^{T}$. At inference, an ODE solver integrates the learned velocity field to generate samples from noise.

Per-Frame Noise Scheduling. To enable flexible autoregressive generation, we adapt the per-frame noise scheduling strategy from Diffusion Forcing [3] into the Flow Matching framework. Instead of a global timestep t, we assign independent continuous flow timesteps $t_i \in [0, 1]$ to each frame i. For the video sequence $\mathbf{x}_{1:T} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$, the intermediate state of the i-th frame at its specific timestep t_i is constructed as:

$$\mathbf{x}_{t_i}^{(i)} = t_i\mathbf{x}_1^{(i)} + (1-t_i)\mathbf{x}_0^{(i)}, \tag{4}$$

where $\mathbf{x}_1^{(i)}$ is the clean data for frame i and $\mathbf{x}_0^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is its corresponding independent standard Gaussian noise. This frame-wise formulation allows

the model to condition on frames at arbitrary noise levels, supporting stable generation beyond the training horizon.

Autoregressive Sampling. At inference, we define a timestep vector $\mathbf{t} = [t_1, \ldots, t_T]$. We initialize the clean context frames with $t_i = 1$ (data) for $i \leq N_{\text{ctx}}$, and the future frames to be generated with $t_j = 0$ (noise) for $j > N_{\text{ctx}}$. Sampling proceeds by solving the empirical ODE:

$$\mathrm{d}\mathbf{x} = \mathbf{v}_\theta\left(\mathbf{x}_\mathbf{t}, \mathbf{y}, \mathbf{t}\right) \mathrm{d}\mathbf{t}, \tag{5}$$

solved iteratively via the Euler method from $\mathbf{t} = \mathbf{0}$ to $\mathbf{t} = \mathbf{1}$ (for generated frames), extending to long-term video generation and interactive simulation.

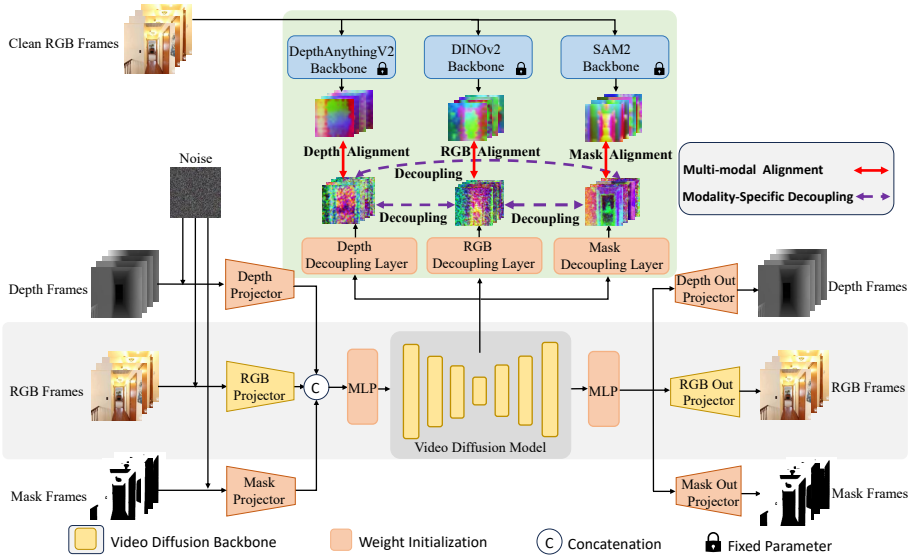

Fig. 2: Overview of M^2-REPA. To mitigate feature conflicts, M^2-REPA decouples multi-modal features into modality-specific representations and aligns them with corresponding expert foundation models. The framework is optimized by two synergistic objectives: (1) a cosine similarity-based multi-modal alignment loss for joint representation alignment, and (2) a CKA [23] similarity-driven modality-specific decoupling loss to explicitly encourage cross-modal complementarity.

4 Method

4.1 Method Overview

Motivation. Recent works [46, 48, 58, 60] have explored leveraging features from pre-trained foundation encoders, such as DINOv2 [28], CLIP [33], and VGGT [44], to ground the generation process and improve temporal consistency in video diffusion models. However, these methods are still largely confined to RGB-only generation and typically rely on isolated foundation models, leaving

multi-modal scenarios underexplored. We observe that existing foundation models are trained in distinct modality spaces: DINOv2 [28] captures rich RGB visual representations, DepthAnythingV2 [50] excels at geometric depth reasoning, and SAM2 [34] provides strong segmentation knowledge. These models therefore serve as modality experts with unique and complementary priors. The distributional differences among their backbone features (see Fig. 1) further reveal their specialization. This insight motivates our key question: can a multi-modal world model benefit from simultaneously transferring domain-specific priors from multiple foundation model experts across different modalities?

Observation. As a pilot study, we first construct a dual-modality video diffusion model for joint RGB-depth generation by concatenating both modalities, applying MLP projection layers, and fine-tuning with modality-specific heads atop a pre-trained backbone [38]. We conduct experiments using the REPA alignment method with two foundation models: DINOv2 (RGB expert) and Depth Anything V2 (depth expert). As shown in Tab. 1, a clear **"modality advantage"** emerges: the model guided by the RGB expert (DINOv2) achieves superior generation quality in the RGB modality, while the model guided by the depth expert (DepthAnythingV2) exhibits stronger performance in the depth modality. This demonstrates that foundation models trained on different modalities capture distinct modality-specific prior knowledge. **This finding reveals a key opportunity**: by simultaneously transferring knowledge from multiple expert foundation models into the video diffusion model, we can improve high-fidelity generation across multiple modalities, thereby achieving multi-expert guided generation.

Therefore, we propose a representation alignment-based approach that first disentangles modality-specific features within the video diffusion model, then aligns each with its corresponding expert foundation model. This encourages the network to jointly internalize diverse prior knowledge during training. In the following sections, we detail our M^2-REPA framework, which bridges the gap between these decoupled diffusion features and heterogeneous foundation priors via two carefully designed regularization objectives.

4.2 MultiModal Representation Alignment

Challenge. Introducing priors from multiple modality-specific foundation models into Video Diffusion Models (VDMs) is non-trivial. Existing strategies fall into two categories: direct injection [48] and implicit alignment [19,39,46,58,60]. Direct injection suffers from the absence of video inputs at inference and inherent incompatibility between pixel-space foundation features and VAE latent spaces; injecting features from multiple modalities further risks feature redundancy and training collapse. REPA [58] offers a compelling alternative by regularizing alignment between diffusion and foundation model features during training, improving generation quality without inference-time overhead. However, existing REPA methods [19, 39, 58, 60] are confined to single-modality RGB generation with a single foundation model. Moreover, naively extending alignment to multiple foundation models via feature similarity alone introduces feature conflicts, yielding negligible gains (see Tab. 1).

These limitations collectively indicate that directly applying standard REPA is inadequate for enhancing multi-modal world model generation. This requires a different strategy that simultaneously incorporates priors from multiple modality-specific foundation models while actively preventing feature conflicts arising from concurrent multi-modal alignment.

To address these challenges, we propose **M^2-REPA** (see Fig. 2), which extracts multi-modal features from intermediate diffusion representations, decouples them through modality-specific decoupling layers, and aligns each with its corresponding "single-modality expert" foundation model. We introduce two complementary regularization objectives: (1) *Multi-modal Representation Alignment* and (2) *Modality-specific Feature Decoupling*, which jointly optimize the alignment between diffusion latent features and multiple pre-trained foundation models, making use of their complementary modality-specific priors.

Model Architecture. To accommodate diverse generation scenarios, we construct a unified tri-modal (RGB-Depth-Mask) video diffusion baseline. Without loss of generality, we detail the pixel-space instantiation built upon the autoregressive UViT backbone [14] below, while deferring the structural adaptations for the latent-space DiT variant to the supplementary material. As illustrated in Fig. 2, we introduce separate projection layers for each modality. Specifically, we duplicate the pre-trained RGB input and output projection layers to serve as dedicated encoders and decoders for the newly incorporated Depth and Mask modalities. For each modality $m \in \{\mathrm{RGB}, \mathrm{D}, \mathrm{M}\}$, we first encode the input data X_t^m as $e_t^m = \mathrm{Enc}^m(X_t^m)$. The tri-modal embeddings $e_t^m \in \mathbb{R}^d$ are then concatenated along the channel dimension and projected through a learnable MLP layer to obtain a unified joint representation $e_t = \mathrm{MLP}_{\mathrm{fuse}}([e_t^{\mathrm{RGB}}; e_t^{\mathrm{D}}; e_t^{\mathrm{M}}])$. The fused embeddings are subsequently processed by the diffusion backbone f_θ, conditioned on context frames and task-specific signals y (*e.g.*, camera poses), to predict the velocity field $\tilde{v}_t = f_\theta(e_t, y, t)$. The output is projected via another MLP and split channel-wise to recover modality-specific representations $[\hat{e}_t^{\mathrm{RGB}}; \hat{e}_t^{\mathrm{D}}; \hat{e}_t^{\mathrm{M}}] = \mathrm{Split}(\mathrm{MLP}_{\mathrm{split}}(\tilde{v}_t))$. Finally, each token is decoded to produce the final tri-modal velocity field $\hat{v}_t = [\hat{v}_t^{\mathrm{RGB}}, \hat{v}_t^{\mathrm{D}}, \hat{v}_t^{\mathrm{M}}]$, allowing us to minimize the objective in Eq. (3).

Multi-modal Representation Alignment. While prior REPA-based methods [19, 60] focus on RGB modality alignment using a single vision foundation model, we introduce Multi-modal Representation Alignment (M^2-REPA), the first alignment approach for multi-modal video generation that simultaneously leverages multiple foundation models. By exploiting complementary strengths of diverse pre-trained models, our method improves generation quality and semantic consistency.

We enforce alignment between the diffusion intermediate features $h_t = f_\theta(x_t)$ extracted from the denoising network on noisy frames x_t and the target representations $y_*^{(k)} = f^{(k)}(x)$ extracted from K pre-trained foundation models on clean frames x, where $k \in \{1, ..., K\}$ denotes the model index and K is the total number of foundation models. In this work, we employ three complementary

"modality experts" ($K = 3$): DINOv2 [28] for RGB, DepthAnythingV2 [50] for depth, and SAM2 [34] for mask segmentation.

For each modality k, features are reshaped to $\mathbb{R}^{BT\times N\times D}$ where B is batch size, T is frame count, and N is spatial tokens. After L2 normalization, the multi-modal alignment loss is defined as:

$$\mathcal{L}_{\text{M}^2\text{-REPA}} = -\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\cos(\hat{y}^{(k)}_{*,[n]}, \hat{h}^{(k)}_{\phi,[n]})\right], \tag{6}$$

where $\hat{h}^{(k)}_{\phi} = g^{(k)}_{\phi}(h_t)$ denotes the projected features obtained through a lightweight MLP $g^{(k)}_{\phi}$(detailed in the next subsection), $\cos(\cdot,\cdot)$ represents the cosine similarity, and the loss maximizes token-wise cosine similarity across all modalities. During training, only the MLP projectors and the diffusion backbone are optimized while all foundation models remain frozen.

Modality-specific Feature Decoupling. However, naively aligning diffusion intermediate features with multiple foundation models simultaneously can paradoxically degrade generation performance compared to using a single foundation model (see Tab. 1). This stems from the inherent conflicts between different modality-specific priors: directly forcing the shared diffusion features to match heterogeneous foundation model representations leads to feature entanglement and optimization difficulties. To address this challenge, we introduce Modality-specific Feature Decoupling, which explicitly disentangles modality-specific information from the diffusion backbone before alignment, enabling effective multi-modal joint optimization.

Concretely, we introduce modality-specific decoupling layers $\{g^{(k)}_{\phi}\}_{k=1}^{K}$ implemented as lightweight MLPs, which project the shared diffusion features h_t into modality-specific subspaces: $\hat{h}^{(k)}_{\phi} = g^{(k)}_{\phi}(h_t)$. Each decoupled representation $\hat{h}^{(k)}_{\phi}$ is then aligned with its corresponding foundation model target $y^{(k)}_{*}$ via Eq. 6. This design is simple yet highly effective, allowing each modality expert to guide its dedicated feature subspace without interfering with others. Empirical analysis and further discussions motivating the choice of MLPs for feature decoupling are deferred to the supplementary material.

To further encourage the decoupled features to capture complementary information across modalities, we design a Modality-specific Decoupling Regularization Loss. We employ Centered Kernel Alignment (CKA) [23] as the similarity metric, which is particularly well-suited for quantifying representational similarity due to its invariance to isotropic scaling and orthogonal transformations. Specifically, we minimize the pairwise CKA similarity among all K decoupled modality-specific features to encourage orthogonality and complementarity. The modality-specific decoupling regularization loss is defined as:

$$\mathcal{L}_{\text{decouple}} = \frac{2}{K(K-1)}\sum_{i=1}^{K}\sum_{j=i+1}^{K}\text{CKA}(\hat{h}^{(i)}_{\phi}, \hat{h}^{(j)}_{\phi}), \tag{7}$$

Table 2: Quantitative comparison on the RealEstate10K dataset for both short-term (8-frame) and long-term (200-frame) video generation. Our M^2-REPA achieves the best performance across most metrics.

Method	Frames	RGB modality				Depth modality		Mask modality
		FVD↓	LPIPS↓	SSIM↑	PSNR↑	AbsRel↓	δ_1↑	mIoU↑
DFoT [38]	8	110.88	0.2471	0.6691	18.3858	–	–	–
Geometry Forcing [46]	8	106.00	0.2292	0.7008	18.8756	–	–	–
RGB-DM Baseline	8	102.44	0.3449	0.5906	19.5264	0.9076	0.1724	0.8839
REPA (DINOv2 [28])	8	100.00	0.2372	0.7142	19.6205	1.0329	0.2987	0.8467
REPA (SAM2 [34])	8	97.63	0.2344	0.7172	19.5104	0.8431	0.2866	0.8801
REPA (DepthAnythingV2 [50])	8	96.81	0.2364	0.7165	19.5978	0.8301	0.3090	0.8787
M^2-REPA (ours)	8	**81.13**	**0.2123**	**0.7439**	**20.4845**	**0.8224**	**0.4449**	**0.9045**
DFoT [38]	200	401.50	0.5218	0.3536	11.1614	–	–	–
Geometry Forcing [46]	200	371.50	0.4897	0.3791	11.2901	–	–	–
RGB-DM Baseline	200	407.50	0.5334	0.3391	12.7900	0.9991	0.1238	0.6385
REPA (DINOv2 [28])	200	308.50	0.4682	**0.4417**	**12.9255**	1.0007	0.1240	0.6191
REPA (SAM2 [34])	200	301.25	0.4696	0.4319	12.5578	0.9958	0.1247	0.6283
REPA (DepthAnythingV2 [50])	200	319.00	0.4668	0.4390	12.8776	0.9988	0.1240	0.6446
M^2-REPA (ours)	200	**262.50**	**0.4592**	0.4374	12.8918	**0.9813**	**0.1408**	**0.6591**

where $\hat{h}_\phi^{(k)}$ denotes the decoupled feature for the k-th modality (e.g., RGB, Depth, and Mask when $K = 3$). By minimizing $\mathcal{L}_{\text{decouple}}$, we explicitly penalize redundant information sharing across modalities, ensuring that each modality-specific subspace captures distinct and complementary semantic structures.

4.3 MultiModal Video Diffusion Models

The complete training objective integrates flow matching, multi-modal alignment, and decoupling regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{align}}\mathcal{L}_{\text{M}^2\text{-REPA}} + \lambda_{\text{decouple}}\mathcal{L}_{\text{decouple}}, \tag{8}$$

where λ_{align} and $\lambda_{\text{decouple}}$ balance the contributions of alignment and decoupling objectives. During training, only the diffusion backbone and MLP projectors $\{g_\phi^{(k)}\}_{k=1}^K$ are trainable, while all foundation models remain frozen to preserve their pre-trained knowledge. This design enables our M^2-REPA framework to harness diverse foundation model priors without feature conflicts, improving multi-modal video generation quality. Hyperparameter sensitivity analysis is deferred to the supplementary material.

5 Experiments

Experimental Setup. To evaluate M^2-REPA, we build our multi-modal baseline upon DFoT [38], following its experimental protocol. We consider two representative scenarios: **(1) Real-world scene generation** on RealEstate10K [62] under camera-pose conditioning, and **(2) Dynamic environment generation** in Minecraft [49] under action conditioning. For multi-modal data construction, depth pseudo-labels are estimated via DepthAnythingV2 [50], and segmentation masks are extracted using SAM2 [34], retaining the top-3 highest-confidence masks per frame. We adopt UViT [14] as the diffusion backbone

Table 3: Ablation study of M^2-REPA components. †D+S+DA denotes the naive combination of DINOv2, SAM2, and DepthAnythingV2.

Method	Frames	RGB modality				Depth modality		Mask modality
		FVD↓	LPIPS↓	SSIM↑	PSNR↑	AbsRel↓	δ_1↑	mIoU↑
RGB-DM Baseline	8	102.44	0.3449	0.5906	19.5264	0.9076	0.1724	<u>0.8839</u>
REPA (DINOv2 [28])	8	100.00	0.2372	0.7142	<u>19.6205</u>	1.0329	0.2987	0.8467
REPA (SAM2 [34])	8	97.63	<u>0.2344</u>	<u>0.7172</u>	19.5104	0.8431	0.2866	0.8801
REPA (DepthAnythingV2 [50])	8	96.81	0.2364	0.7165	19.5978	<u>0.8301</u>	<u>0.3090</u>	0.8787
REPA (D+S+DA)†	8	<u>96.38</u>	0.2377	0.7133	19.4952	0.8384	0.3095	0.8648
M^2-REPA (cos^2 loss)	8	99.56	0.2381	0.7147	19.4227	0.8822	0.3103	0.8637
M^2-REPA (CKA loss, Ours)	8	**81.13**	**0.2123**	**0.7439**	**20.4845**	**0.8224**	**0.4449**	**0.9045**
RGB-DM Baseline	200	407.50	0.5334	0.3391	12.7900	0.9991	0.1238	0.6385
REPA (DINOv2 [28])	200	308.50	0.4682	<u>0.4417</u>	<u>12.9255</u>	1.0007	0.1240	0.6191
REPA (SAM2 [34])	200	<u>301.25</u>	0.4696	0.4319	12.5578	0.9958	0.1247	0.6283
REPA (DepthAnythingV2 [50])	200	319.00	0.4668	0.4390	12.8776	0.9988	0.1240	0.6446
REPA (D+S+DA)†	200	328.75	0.4689	0.4340	12.7331	<u>0.9940</u>	<u>0.1275</u>	0.6470
M^2-REPA (cos^2 loss)	200	319.75	<u>0.4652</u>	**0.4420**	**12.9756**	0.9992	0.1239	<u>0.6490</u>
M^2-REPA (CKA loss, Ours)	200	**262.50**	**0.4592**	0.4374	12.8918	**0.9813**	**0.1408**	**0.6591**

for RealEstate10K and DiT [30] for Minecraft, spanning both U-Net-based and Transformer-based architectures to evaluate our approach across diverse backbone designs.

Implementation Details. Guided by the architecture in Sec. 4.2, we evaluate our method on the challenging joint RGB-Depth-Mask generation task. For camera view-conditioned video generation [62], the model is trained in pixel space using 8-frame video clips at 256×256 resolution for 2,500 iterations. For action-conditioned video generation [49], we adopt a latent-diffusion paradigm where videos are pre-encoded into a latent space via a pre-trained VAE [35]; training is conducted on 50-frame latent sequences for 5,000 iterations. Both setups share a learning rate of 8×10^{-6} and a batch size of 8. During inference, we perform auto-regressive roll-outs conditioned on the first frame, encompassing both short-range (8 frames) and long-range (200 frames) horizons. To balance the multi-objective optimization, we empirically set $\lambda_{\text{align}} = 0.5$ and $\lambda_{\text{decouple}} = 0.05$. All experiments are executed on a cluster of 8 NVIDIA H20 GPUs.

Evaluation Metrics. For the RGB modality, we adopt standard video generation metrics to assess visual quality, including FVD [40], PSNR, SSIM [45] and LPIPS [59]. For the depth modality, following DepthCrafter [16], we first align predicted depth maps with ground truth via uniform scale-shift normalization across the entire video, then evaluate geometric accuracy using AbsRel and δ_1 accuracy [16, 50]. For segmentation masks, we compute mIoU [9] by performing greedy one-to-one matching between predicted and ground-truth masks, averaging matched IoU scores at the frame level, then across all frames. Additional metric details are provided in the supplementary material.

5.1 Quantitative comparisons

Real-world Scene Generation. We conduct comprehensive quantitative evaluations on the RealEstate10K dataset [62], covering both short-term (8-frame) and long-term (200-frame) video generation tasks. As shown in Tab. 2, we compare M^2-REPA against several strong baselines, including DFoT [38], Geometry

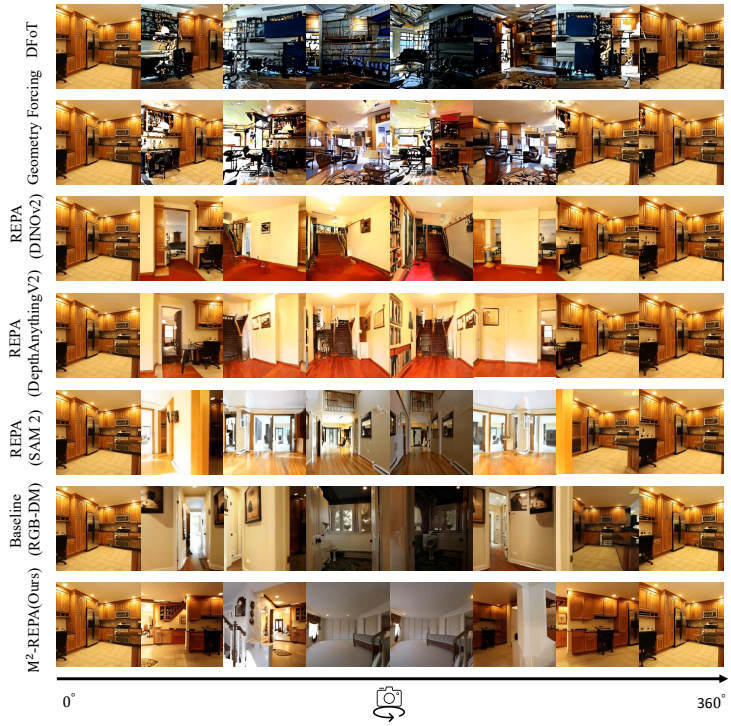

Fig. 3: Qualitative comparison of camera view-conditioned video generation under fullcircle rotation. Videos are generated from a input frame and corresponding per-frame camera poses simulating a full 360° rotation.

Forcing [46], our RGB-DM baseline, and single-modality REPA variants that leverage DINOv2 [28], SAM2 [34], and DepthAnythingV2 [50]. From the metrics, our method achieves substantial improvements across all modalities. M^2-REPA consistently outperforms all baselines across multiple evaluation metrics spanning three modalities, including FVD, δ_1, and mIoU, demonstrating the best generation performance. These results validate that our approach harnesses the complementary priors from multiple foundation models, enabling improved joint generation and optimization across modalities.

Dynamic Environment Generation. To evaluate the generalizability of our framework across diverse data distributions and architectures, we extend our tri-modal baseline to the dynamic and visually varied Minecraft environment [49]. For this task, we employ DiT [30] as the underlying backbone to further validate our method's backbone-agnostic nature. As summarized in Tab. 4, M^2-REPA yields notable performance gains in long-term (150-frame) video generation. These results indicate that our method can be integrated into various video diffusion frameworks, delivering consistent gains regardless of the specific data domain or network architecture.

Table 4: Evaluation on dynamic environment generation in Minecraft. †D+S+DA denotes the naive combination of DINOv2, SAM2, and DepthAnythingV2.

Method	Frames	FVD↓	LPIPS↓	SSIM↑	PSNR↑	AbsRel↓	δ_1↑	mIoU↑
RGB-DM Baseline	150	191.25	0.5561	0.4515	12.8364	1.0467	0.0228	0.5962
REPA(DINOv2)	150	129.50	0.5300	0.4482	12.9883	1.0505	0.0228	0.6371
REPA(SAM2)	150	132.12	0.5386	0.4474	12.9483	1.0539	**0.0229**	0.6455
REPA(DepthAnythingV2)	150	133.50	0.5390	0.4474	12.9510	1.0545	**0.0229**	0.6451
REPA (D+S+DA)†	150	144.75	0.5388	0.4459	12.9494	1.0562	**0.0229**	0.6369
M^2-REPA (Ours)	150	**81.81**	**0.5258**	**0.4566**	**13.0081**	**1.0159**	0.0228	**0.6612**

5.2 Qualitative comparisons

Fig. 3 presents qualitative comparisons of our M^2-REPA against baseline methods on the RealEstate10K dataset. Each video is generated conditioned on the first frame and future camera poses, simulating a 360° rotation. Examining the four intermediate frames, we observe that both DFoT and Geometry Forcing suffer from severe scene collapse, while the RGB-DM baseline and single-modality REPA variants exhibit varying degrees of distortion and blurry artifacts. In contrast, our method generates RGB frames with sharper details and more realistic textures. Additionally, Fig. 4 visualizes the generation results across all three modalities (RGB-Depth-Mask) under the long video generation setting (200 frames). While the baseline and single-modality REPA methods exhibit memory drift of scene elements such as the sofa during long-horizon generation, M^2-REPA maintains consistent temporal coherence throughout the sequence, demonstrating substantial joint quality improvements across all modalities.

5.3 Ablation studies

Effect of Naive Multi-Modal Alignment. As shown in Tab. 3, while individual single-modality REPA variants consistently outperform the RGB-DM baseline, naively aligning all three foundation models simultaneously without modality-specific decoupling severely degrades long-video generation. This phenomenon corroborates our core hypothesis (see Sec. 4.2): directly forcing shared representations to match heterogeneous foundation models inevitably triggers severe feature conflicts.

Effectiveness of Modality-Specific Decoupling Regularization. The results demonstrate that incorporating modality-specific decoupling regularization yields substantial performance improvements compared to the naive direct alignment of multiple foundation models. To further assess the importance of our modality-specific decoupling loss $\mathcal{L}_{\text{decouple}}$, we compare two regularization strategies: (1) **cos^2 loss**, which minimizes $\frac{1}{3}\sum_{i,j}(\cos(\hat{h}_\phi^{(i)}, \hat{h}_\phi^{(j)}))^2$ to encourage orthogonality, and (2) **CKA loss** [23] (our method), as defined in Equation 7. The empirical results reveal a clear advantage of the CKA-based formulation. These findings underscore the importance of employing CKA as a similarity metric, which exhibits superior robustness to feature distribution variations and orthogonal transformations compared to naive cosine-based penalties.

Feature Layer Selection and Sensitivity. For foundation models, we strictly follow the original settings from REPA series works [39,58]: DINOv2 uses the 4th

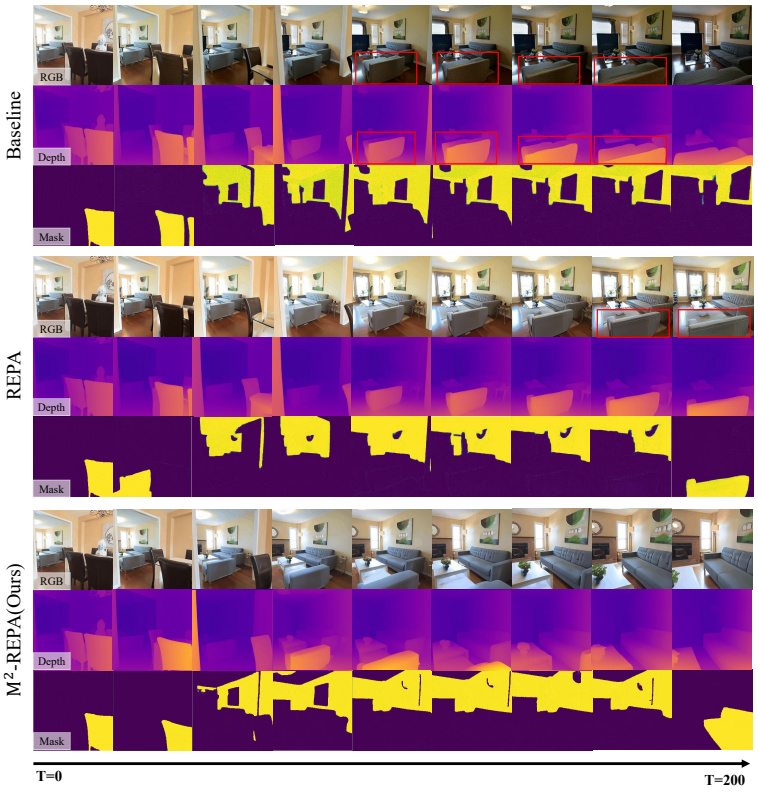

Fig. 4: Qualitative comparison in ablation study. Under the 200-frame long video setting, we compare our method (M^2-REPA) against baseline and REPA methods.

layer output, DepthAnythingV2 uses the last layer of the encoder, and SAM2 uses the 11th layer of the image encoder. Additionally, as shown in Fig. 5, we ablate the choice of alignment layer across the 7-layer U-ViT backbone [14] (3 downsampling, 1 bottleneck, and 3 upsampling layers) on RealEstate10K, finding that aligning with earlier blocks yields superior performance.

Robustness to Supervision Bias. To preclude potential "self-fulfilling" effects, where the model might merely distill the same teacher used for label generation, we conduct a cross-source robustness study. Specifically, we train the baseline using depth labels from Video-Depth-Anything [5], while our M^2-REPA is guided by DepthAnythingV2 [50] and DINOv2 [28]. As shown in Tab. 5, our method consistently yields substantial gains over the baseline on the RealEstate10K dataset despite this source mismatch. This confirms that M^2-REPA's efficacy transcends simple label-space distillation; rather, it enhances generation by regularizing internal latent-level representations with structural priors that are orthogonal to the specific choice of pseudo-labels.

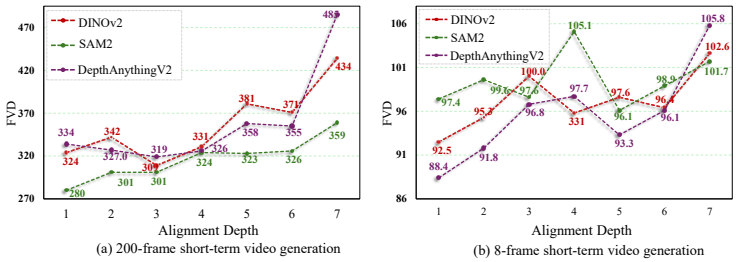

Fig. 5: Ablation on alignment layer depth. FVD-200 and FVD-8 scores across different alignment layers of the diffusion backbone.

Table 5: Cross-source robustness analysis on RealEstate10K. M^2-REPA consistently outperforms the baseline under mismatched depth supervision.

Method	Frames	RGB modality				Depth modality		Mask modality
		FVD↓	LPIPS↓	SSIM↑	PSNR↑	AbsRel↓	δ_1↑	mIoU↑
Video-Depth-Anything Baseline	8	124.00	0.3120	0.6690	**19.1000**	0.4530	0.5300	**0.8810**
M^2-REPA (DINOv2+DepthAnythingV2)	8	**118.00**	**0.2950**	**0.6800**	18.7000	**0.4170**	**0.5720**	0.8800
Video-Depth-Anything Baseline	200	346.75	0.5332	0.3916	12.0521	0.5184	0.3754	0.6109
M^2-REPA (DINOv2+DepthAnythingV2)	200	**291.00**	**0.5046**	**0.4223**	**12.3226**	**0.5099**	**0.3989**	**0.6365**

5.4 Limitations and Broader Impact

M^2-REPA's robustness is tied to the quality of the employed foundation experts, which may contain implicit biases. Furthermore, multi-expert alignment requires additional training compute, albeit without inference overhead. Broadly, the dual-use nature of high-fidelity synthesis highlights the importance of developing robust media authentication tools.

6 Conclusion

We presented M^2-REPA, the first representation alignment framework for multi-modal video generation that simultaneously integrates complementary priors from diverse foundation experts. By explicitly decoupling modality-specific features and enforcing joint alignment, our method resolves inherent representation conflicts while making use of heterogeneous pre-trained knowledge. Empirical results demonstrate M^2-REPA's consistent improvements over competitive baselines across all modalities, providing a strong baseline and a promising direction for multi-modal world modeling.

Acknowledgements

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No. KJZD20231023094700001) and the Shenzhen-Tsinghua Special Project for Fundamental & Frontier Research in Artificial Intelligence (No. AI2026018).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical AI. Proceedings of the 3rd International Workshop on Rich Media With Generative AI, ACM (2025)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)
4. Chen, J., Zhu, H., He, X., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Fu, Z., Pang, J., et al.: DeepVerse: 4D autoregressive video generation as a world model. arXiv preprint arXiv:2506.01103 (2025)
5. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video Depth Anything: Consistent depth estimation for super-long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22831–22840 (2025)
6. Chen, Z., Liu, T., Zhuo, L., Ren, J., Tao, Z., Zhu, H., Hong, F., Pan, L., Liu, Z.: 4DNeX: Feed-forward 4D generative modeling made easy. arXiv preprint arXiv:2508.13154 (2025)
7. Cherian, A., Corcodel, R., Jain, S., Romeres, D.: LLMPhy: Complex physical reasoning using large language models and world models. arXiv preprint arXiv:2411.08027 (2024)
8. Decart, E., McIntyre, Q., Campbell, S., Chen, X., Wachen, R.: Oasis: A universe in a transformer. URL: https://oasis-model.github.io (2024), accessed: 2026-06-29
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
10. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
11. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
12. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-Game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
14. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. In: International Conference on Machine Learning. pp. 13213–13232. PMLR (2023)
15. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: GAIA-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
16. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: DepthCrafter: Generating consistent long depth sequences for open-world videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2005–2015 (2025)

17. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
18. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
19. Hwang, S., Jang, H., Kim, K., Park, M., Choo, J.: Cross-frame representation alignment for fine-tuning video diffusion models. arXiv preprint arXiv:2506.09229 (2025)
20. Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: DINO-Foresight: Looking into the future with DINO. arXiv preprint arXiv:2412.11673 (2024)
21. Kim, J., Kang, J., Choi, J., Han, B.: FIFO-Diffusion: Generating infinite videos from text without training. Advances in Neural Information Processing Systems **37**, 89834–89868 (2024)
22. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
23. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMLR (2019)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
26. Liu, Z., Deng, X., Chen, S., Wang, A., Guo, Q., Han, M., Xue, Z., Chen, M., Luo, P., Yang, L.: WorldWeaver: Generating long-horizon video worlds via rich perception. arXiv preprint arXiv:2508.15720 (2025)
27. Mei, Z., Shorinwa, O., Majumdar, A.: Geometry meets vision: Revisiting pretrained semantics in distilled fields. arXiv preprint arXiv:2510.03104 (2025)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Parker-Holder, J., Fruchter, S., et al.: Genie 3: A new frontier for world models. URL: https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/ (2025), accessed: 2026-06-29
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
31. Qian, Z., Chi, X., Li, Y., Wang, S., Qin, Z., Ju, X., Han, S., Zhang, S.: WristWorld: Generating wrist-views via 4D world models for robotic manipulation. arXiv preprint arXiv:2510.07313 (2025)
32. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., et al.: WorldSimBench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072 (2024)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)

34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
36. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: GAIA-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
38. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
39. Tian, Y., Chen, H., Zheng, M., Liang, Y., Xu, C., Wang, Y.: U-REPA: Aligning diffusion U-Nets to ViTs. arXiv preprint arXiv:2503.18414 (2025)
40. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
41. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
42. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
43. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L.Z., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., et al.: SpatialVID: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025)
44. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
46. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)
47. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: WorldMem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369 (2025)
48. Xu, G., Lin, H., Luo, H., Wang, X., Yao, J., Zhu, L., Pu, Y., Chi, C., Sun, H., Wang, B., et al.: Pixel-perfect depth with semantics-prompted diffusion transformers. arXiv preprint arXiv:2510.07316 (2025)
49. Yan, W., Hafner, D., James, S., Abbeel, P.: Temporally consistent transformers for video generation. In: International Conference on Machine Learning. pp. 39062–39098. PMLR (2023)
50. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
51. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114 (2023)

52. Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., Schuurmans, D.: Video as the new language for real-world decision making. arXiv preprint arXiv:2402.17139 (2024)
53. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
54. Yin, T., Mei, Z., Sun, T., Zha, L., Zhou, E., Bao, J., Yamane, M., Sho, O., Majumdar, A.: WoMAP: World models for embodied open-vocabulary object localization. In: RSS 2025 Workshop: Mobile Manipulation: Emerging Opportunities & Contemporary Challenges (2025)
55. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
56. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
57. Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., Liu, X.: GameFactory: Creating new games with generative interactive videos. arXiv preprint arXiv:2501.08325 (2025)
58. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
60. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: VideoREPA: Learning physics for video generation through relational alignment with foundation models. arXiv preprint arXiv:2505.23656 (2025)
61. Zhen, H., Sun, Q., Zhang, H., Li, J., Zhou, S., Du, Y., Gan, C.: TesserAct: learning 4D embodied world models. arXiv preprint arXiv:2504.20995 (2025)
62. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
63. Zhou, Y., Wang, Y., Zhou, J., Chang, W., Guo, H., Li, Z., Ma, K., Li, X., Wang, Y., Zhu, H., et al.: OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. arXiv preprint arXiv:2509.12201 (2025)
64. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8535–8546 (2025)
65. Zhu, Z., Wang, X., Zhao, W., Min, C., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., Zhang, C., et al.: Is Sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520 (2024)