ICDepth: Taming Video Diffusion Models for Video Depth Estimation via In-Context Conditioning

Xuanhua He*, Jiaxin Xie*, Mingzhe Zheng, and Qifeng Chen†

The Hong Kong University of Science and Technology, Hong Kong SAR, China

Project Page: https://xuanhuahe.github.io/ICDepth/

Abstract. Monocular video depth estimation requires temporal consistency, geometric accuracy, and generalization across diverse scenarios—yet existing methods struggle to achieve all three simultaneously. Discriminative models excel at per-frame accuracy but suffer from temporal drift due to limited context windows, while generative methods improve consistency and generalization at the cost of extensive training data (10M+ samples) and lack of geometric precision. In response to these issues, we introduce **ICDepth**, a framework that adapts pre-trained text-to-video diffusion transformers for video depth estimation via In-Context Conditioning (ICC), leveraging their rich spatial-temporal priors. To address key challenges in transferring ICC from generation to dense prediction, we propose: (1) **SAND-Attention**, which ensures precise spatial-temporal alignment via shared RoPE and enforces unidirectional attention to prevent noise contamination; (2) **SRFM**, which injects DINOv2 semantic and resolution priors to enhance geometric precision. ICDepth achieves state-of-the-art results on multiple benchmarks with remarkable data efficiency, trained on only 0.8M frames (6–13× less than competing generative methods), while demonstrating strong zero-shot generalization to diverse domains.

Keywords: Video Depth Estimation · Monocular Depth Estimation · Video Diffusion Models

1 Introduction

Monocular video depth estimation, which predicts a dense depth map for each frame in a video, is a fundamental task in 3D computer vision. It is crucial for many applications, such as AR/VR, autonomous driving, and 3D reconstruction. While recent models like Depth Anything V2 [37] have made great progress in single-image depth estimation, video-based methods face three core challenges. The first is maintaining temporal consistency to ensure the depth maps are stable and flicker-free across frames. The second is effectively leveraging temporal information from adjacent frames to improve the accuracy and robustness of the

* Equal contribution. † Corresponding author.

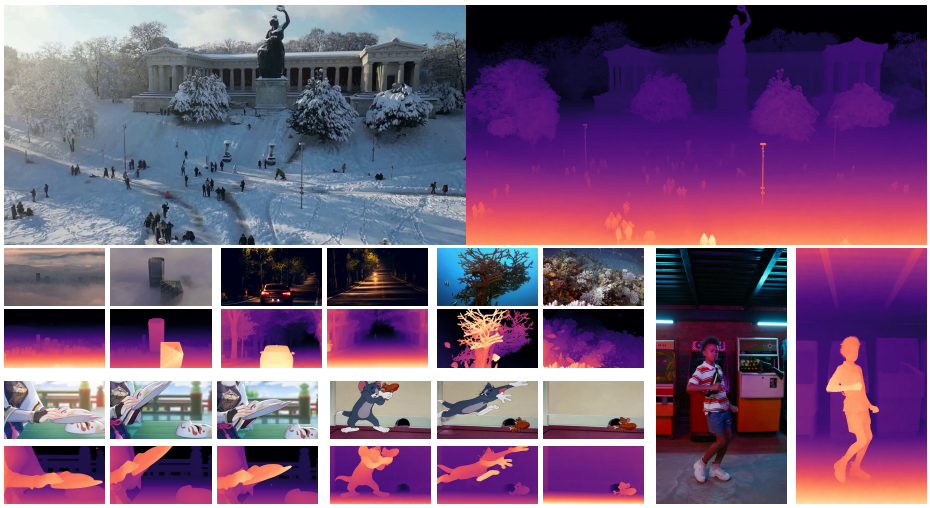

Fig. 1: Robust depth estimation across diverse scenarios. Our method performs high-resolution depth estimation (1080p) on videos with varying aspect ratios and demonstrates strong generalization across challenging conditions including foggy weather, nighttime scenes, underwater footage, and both 2D and 3D animated content.

prediction for the current frame. The third is generalization ability in diverse scenes. Overcoming these challenges, especially in long videos with complex motion, remains a primary goal in this field.

Current approaches to video depth estimation predominantly fall into two categories. Discriminative models, such as Video Depth Anything (VDA) [3], typically adapt powerful single-image backbones by appending a post-processing temporal head. However, this strategy is inherently limited; the temporal head usually operates over a narrow local window, struggling to capture long-range dependencies. This constrained temporal modeling capacity causes the consistency drift in long videos. The results of discriminative models are also overly smooth, lacking fine-grained details and their limited pre-training on diverse video datasets restricts generalization capability. On the other hand, generative models, like DepthCrafter [8] and Depth Any Video [36], leverage pre-trained video diffusion models, primarily Stable Video Diffusion (SVD) [1]. However, their performance is constrained by two factors. First, the foundational SVD model itself is a U-Net architecture with relatively weak prior [28]. Second, they employ a factorized spatio-temporal approach, interleaving 3D convolutional layers with separate 1D temporal attention modules. Adapting them for depth estimation often requires invasive architectural modifications and extensive training on massive datasets (more than 10M samples).

Discriminative methods excel at geometric accuracy but fail at generalization and cannot effectively utilize long-range information, while generative methods improve temporal consistency but sacrifice accuracy and demand prohibitive

data scale. Can we train a model that satisfies all three requirements with limited data?

Recent text-to-video diffusion transformer models, such as Wan 2.1 [32], offer a promising alternative. Through training on millions of diverse videos, these models have learned rich scene and spatial-temporal priors for generating videos. Unlike factorized models, VDiT employ native 3D attention that jointly processes all spatial and temporal tokens, naturally capturing complex spatio-temporal dependencies. This raises a question: *Can we transfer these temporal priors to depth estimation while adding the geometric precision that generative models lack?*

However, standard conditioning approaches are ill-suited for VDiT. Concatenating RGB and depth features along the channel dimension requires modifying the input projection layer and provides only weak feature interaction through separate channels. Instead, we adopt *In-Context Conditioning* (ICC) [7, 10], which treats RGB and depth as a unified token sequence $\mathbf{s} = [\mathbf{z}_{\text{depth}}; \mathbf{c}_{\text{RGB}}]$ processed by the VDiT's native attention. This design is non-invasive and provides rich cross-modal interactions through attention.

Transferring the ICC paradigm from generation to a dense prediction task like depth estimation introduces two non-trivial challenges, arising from ICC itself and the feature space of VDiT. First, in the vanilla ICC framework, the noisy depth latent $\mathbf{z}_{\text{depth}}$ would corrupt the clean RGB condition through bidirectional attention, degrading conditioning quality in deeper layers. Second, T2V models are trained for generation and lack both geometric precision and explicit resolution awareness needed for accurate multi-resolution depth prediction.

To address these challenges, we propose ICDepth, a specialized ICC framework for video depth estimation built upon two key innovations. First, we introduce SAND-Attention (Spatial-temporal Aligned, Noise-Decoupled Attention) to resolve noise contamination. This mechanism establishes precise spatial-temporal alignment between clean RGB conditions and noisy depth latents through RoPE Alignment, while ensuring strict unidirectional information flow by Decoupled Attention. The Decoupled Attention operates in a specialized two-stage process: noisy depth tokens can query clean RGB tokens for contextual guidance, while reverse attention is systematically blocked. Additionally, we zero-out timestep embeddings for RGB conditions, effectively shielding the clean conditioning signal from temporal noise interference.

Furthermore, to tackle geometric ambiguity, we introduce a Semantic- and Resolution-Aware Feature Modulation Block (SRFM). This module injects powerful external priors into the generative model, leveraging DINOv2 [24] features for rich semantic understanding and resolution embeddings for multi-scale inference. These innovations transform the generative model into a precise depth estimation foundation. Remarkably, our model achieves SOTA performance across diverse datasets despite training on a small-scale synthetic dataset (only 0.8M frames). As shown in Figure 1, our method produces highly accurate depth predictions while demonstrating strong generalization across diverse scenarios—including foggy weather, nighttime scenes, underwater footage, and

animated content—and flexibly handles varying resolutions and aspect ratios up to 1080p.

Our contributions are summarized as follows:

- We introduce the In-Context Conditioning (ICC) paradigm for video depth estimation, enabling a powerful text-to-video diffusion transformer to serve as a state-of-the-art depth estimation model.
- We propose a comprehensive solution to overcome the key challenges of applying ICC to this new domain. Our solution includes: a novel SAND-Attention mechanism that resolves the noise contamination issue via shared RoPE and one-way attention; and a semantic- and resolution-aware feature modulation block that injects DINOv2 and resolution priors to tackle the model's inherent geometric ambiguity.
- Our framework achieves state-of-the-art performance on multiple benchmarks with remarkable data efficiency.

2 Related Work

2.1 Video Depth Estimation

Video depth estimation aims to generate temporally consistent depth sequences [12, 17, 40]. Early approaches often relied on Test-Time Optimization, which fine-tunes a pre-trained single-image model for each new video [12, 17, 41]. While capable of producing consistent results, these methods have high inference overhead and are often dependent on precise camera poses or optical flow, limiting their applicability in the wild. Feed-forward methods predict depth directly from video clips. Some approaches introduce plug-and-play modules to stabilize off-the-shelf monocular depth predictions, such as NVDS [34]. Others leverage geometric priors like optical flow or poses to enforce consistency. However, the performance of these models is often susceptible to errors in the priors, and their generalization is frequently hampered by the scarcity of diverse, large-scale video depth training data. More recently, a new line of work leverages the powerful priors of pretrained model such as Depth Anything V2 [37] or SVD [1]. Concurrent works such as ChronoDepth [29], DepthCrafter [8], and Depth Any Video [36] repurpose these generative models to produce high-fidelity and temporally coherent depth sequences, while Video Depth Anything [3] designs a temporal head [13] to leverage the prior from Depth Anything V2.

2.2 In-Context Conditioning for Diffusion Models

The field has witnessed a paradigm shift toward leveraging sequence concatenation as a mechanism for In-Context Conditioning within diffusion transformers [9, 14, 18, 19, 23, 30, 31, 35]. Pioneering work such as Omnicontrol [31] established the foundational concept of unifying conditioning signals and noisy latent representations within a single sequential structure. Through the inherent

self-attention operations of Transformers, these combined sequences enable implicit learning of conditioning dependencies. The architectural simplicity of this methodology presents a compelling advantage—achieving sophisticated multi-modal control while preserving the original DiT framework, consequently enhancing both controllability and synthesis fidelity. This concatenation-based conditioning strategy has catalyzed a wave of follow-up investigations in image synthesis [14, 23, 30, 35]. More recently, the paradigm has been successfully adapted to address the inherently more complex video generation task, demonstrating remarkable efficacy [6, 7, 10, 15, 19–21, 38].

3 Method

In this section, we first briefly introduce the concept of video diffusion model, then we dive into the details of our proposed method.

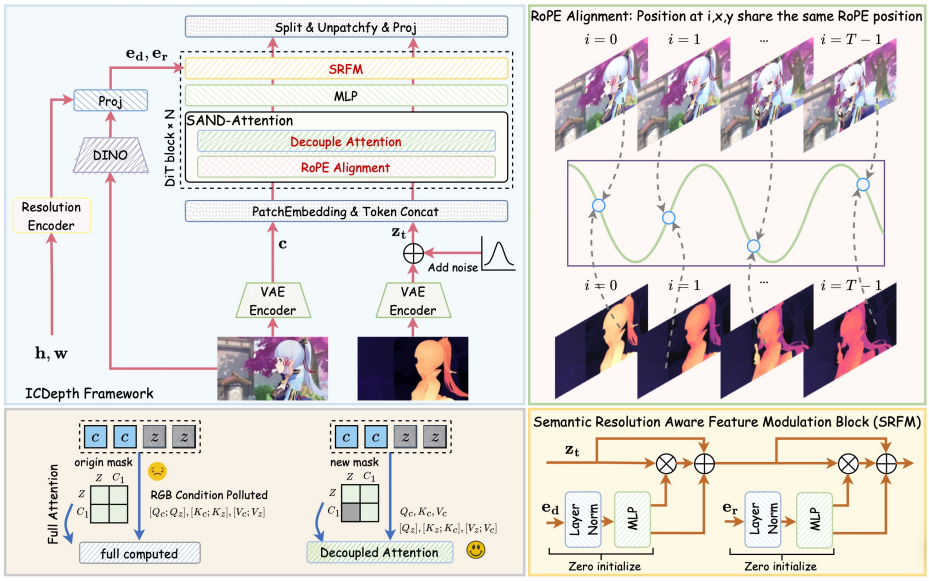

Fig. 2: The framework of our proposed ICDepth, which contains two components: SRFM and SAND-Attention. The RoPE alignment ensures that corresponding spatial-temporal positions in the video-depth paired data share identical RoPE positional encodings.

3.1 Preliminary

video diffusion transformer (VDiT) Our approach builds upon the VDiT, which leverages a pure transformer architecture for video generation. The model

u_Θ is composed of stacked transformer blocks that jointly implement **spatial-temporal self-attention** and **cross-attention**, followed by **MLPs**.

This design provides a unified architecture for capturing complex, long-range spatio-temporal dependencies, as opposed to hybrid CNN-attention models like UNet. The self-attention mechanism globally relates all patches in the spatio-temporal volume, while the cross-attention seamlessly integrates conditional information at every layer.

Following the Flow Matching framework [16], we train the model to predict the velocity field $\mathbf{v}_t$ through the objective:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1,\mathbf{c}} \| u_\Theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{v}_t \|^2 \tag{1}$$

where u_Θ represents our VDiT backbone, $\mathbf{z}_t$ denotes the noisy latent video representation at time t, and $\mathbf{c}$ encompasses conditional inputs.

During inference, we generate video samples by solving the corresponding ODE defined by the learned velocity field using numerical solvers:

$$\frac{d\mathbf{z}(t)}{dt} = u_\Theta(\mathbf{z}_t, t, \mathbf{c}) \tag{2}$$

This formulation enables efficient sampling while maintaining temporal coherence across video frames.

3.2 Overview

The framework of our proposed ICDepth is shown in Figure 2. ICDepth builds upon the pretrained text -to-video generation model Wan [32] with two core designs for video depth estimation.

Given an input RGB video V_I and its corresponding depth video V_D, we first encode each modality into the latent space using a pretrained VAE encoder. This yields the clean latent representation $\mathbf{z}_0$ from V_D and the conditioning latent $\mathbf{c}$ from V_I. We then corrupt $\mathbf{z}_0$ by adding Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to obtain the noisy latent $\mathbf{z}_t$ at diffusion time step t. Both $\mathbf{z}_t$ and $\mathbf{c}$ are patchified and reshaped into tensors of shape $\mathbb{R}^{n\times c}$, then concatenated along the token dimension to form the unified sequence $\mathbf{s}_t = [\mathbf{z}_t; \mathbf{c}] \in \mathbb{R}^{2n\times c}$.

To incorporate prior knowledge into the denoising model, we introduce both semantic and resolution information to enrich the generation process. For semantic conditioning, we extract DINOv2 features [24] from the input video V_I, obtaining feature maps f_d that are subsequently pooled and reshaped into semantic embeddings $\mathbf{e}_d$. For resolution conditioning, the spatial dimensions (h, w) are encoded to $(\mathbf{e}_h, \mathbf{e}_w)$ using sinusoidal positional embeddings, which are concatenated and processed through MLP to yield the resolution embedding $\mathbf{e}_r$.

These conditioning signals are fed into the VDiT model, which is trained to estimate the vector field that characterizes the transformation between the noise distribution and the data distribution. The overall objective is to minimize the flow matching loss. During inference, depth estimation is achieved by solving the ODE defined by the learned vector field, starting from Gaussian noise and

conditioned on the input RGB video with its derived semantic and resolution features, generating accurate and temporally consistent depth maps through numerical integration.

3.3 Key Component

In-Context Conditioning To adapt a pretrained video generation model for depth estimation, we need to effectively incorporate the RGB video condition. Existing methods [8,11,36] achieve this by concatenating $\mathbf{z}$ and $\mathbf{c}$ along the channel dimension, followed by fusion through projection layers. However, this approach has limitations: it causes information loss and underutilizes the attention mechanisms of the pretrained model, as channel concatenation and projection cannot establish strong relationships between the two types of tokens. Instead, we leverage the ICC framework, where the condition $\mathbf{c}$ is provided as context tokens without information loss. The model learns the relationships between depth and RGB features within the attention layers. Specifically, by concatenating $\mathbf{z}$ and $\mathbf{c}$ along the token dimension to form $\mathbf{s} = [\mathbf{z}; \mathbf{c}]$, the self-attention mechanism naturally models their interactions. This framework enables the model to fully exploit the pretrained attention capabilities, leading to improved performance.

SAND-Attention Traditional ICC framework processes $\mathbf{s} = [\mathbf{z}; \mathbf{c}]$ through full self-attention, suffering from three key issues: (1) RoPE creates artificial sequential relationships between aligned depth-RGB tokens; (2) Bidirectional attention allows noise from $\mathbf{z}$ to degrade $\mathbf{c}$; (3) Uniform timestep embedding $\mathbf{e}_t$ injects unnecessary variance into stable RGB features.

To address these limitations, we propose **SAND-Attention** (Spatial-temporal Aligned and Noise-Decoupled Attention). We first separate queries, keys, and values for depth and conditioning tokens:

$$Q_{\mathbf{z}}, Q_{\mathbf{c}} = \text{split}(Q_{\mathbf{s}}), \quad K_{\mathbf{z}}, K_{\mathbf{c}} = \text{split}(K_{\mathbf{s}}), \quad V_{\mathbf{z}}, V_{\mathbf{c}} = \text{split}(V_{\mathbf{s}}) \tag{3}$$

We then apply RoPE with shared positional indices $\mathcal{P}$ to ensure spatial-temporal alignment:

$$\begin{aligned} Q'_{\mathbf{z}}, Q'_{\mathbf{c}} &= \text{RoPE}(Q_{\mathbf{z}}, \mathcal{P}), \text{RoPE}(Q_{\mathbf{c}}, \mathcal{P}) \\ K'_{\mathbf{z}}, K'_{\mathbf{c}} &= \text{RoPE}(K_{\mathbf{z}}, \mathcal{P}), \text{RoPE}(K_{\mathbf{c}}, \mathcal{P}) \end{aligned} \tag{4}$$

The core innovation is our noise-decoupled attention mechanism. We compute self-attention within clean conditioning tokens:

$$O_{\mathbf{c}} = \text{softmax}\left(Q'_{\mathbf{c}} {K'_{\mathbf{c}}}^{T}\right) V_{\mathbf{c}} \tag{5}$$

Followed by unidirectional cross-attention from noisy depth to clean conditioning:

$$O_{\mathbf{z}} = \text{softmax}(Q'_{\mathbf{z}}[K'_{\mathbf{z}}; K'_{\mathbf{c}}]^T)[V_{\mathbf{z}}; V_{\mathbf{c}}] \tag{6}$$

Finally, we concatenate outputs $O_{\mathbf{s}} = [O_{\mathbf{z}}; O_{\mathbf{c}}]$ and modify timestep conditioning by zeroing $\mathbf{e}_t$ for $\mathbf{c}$ tokens, preserving RGB feature stability across diffusion steps. This design maintains compatibility with flash attention while ensuring clean conditioning information flows unidirectionally to guide depth generation.

Semantic-Resolution Aware Feature Modulation Although the pretrained VDiT model can benefit from the ICC framework and our SAND-Attention, certain limitations remain. While the T2V model is trained on extremely large-scale datasets, its learned representations are not always optimal for perception tasks such as depth estimation. Moreover, depth maps are highly correlated with image resolution and aspect ratio. During inference, different test videos have varying aspect ratios, and using a fixed resolution leads to performance degradation.

To address these issues, we enhance the VDiT model using Semantic-Resolution Aware Feature Modulation Block (SRFM). First, we incorporate DINOv2 [24] features as prior prompts to enrich the semantic feature representation. Second, we provide explicit resolution information to improve performance across different training and inference resolutions, particularly for non-standard aspect ratios.

Given feature s^l_{mlp} after the MLP layer in a VDiT block, we inject DINOv2 and resolution priors. As described in Section 3.2, we have embeddings e_d and e_r representing DINOv2 and resolution priors, respectively. We project them using two MLP-based functions:

$$e^l_d, e^l_r = \Theta_d(e_d), \Theta_r(e_r) \tag{7}$$

where the feature channels are doubled through this projection. We then split the features along the channel dimension to obtain scale and shift parameters for modulation:

$$e^{\text{scale}}_d, e^{\text{shift}}_d = \text{split}(e^l_d) \qquad e^{\text{scale}}_r, e^{\text{shift}}_r = \text{split}(e^l_r) \tag{8}$$

To inject this information into $\mathbf{z}$, we split $\mathbf{s}^l$ into $\mathbf{z}^l$ and $\mathbf{c}^l$, then apply sequential modulation:

$$\begin{aligned} \mathbf{z}^l_d = \mathbf{z}^l \cdot (1 + e^{\text{scale}}_d) + e^{\text{shift}}_d, \quad \mathbf{z}^l_r = \mathbf{z}^l_d \cdot (1 + e^{\text{scale}}_r) + e^{\text{shift}}_r \\ s^{l+1} = [\mathbf{z}^l_r; \mathbf{c}^l] \end{aligned} \tag{9}$$

This design improves model performance across diverse scenes and resolutions.

3.4 Training Pipeline

Given an input RGB video V_I and its corresponding depth video V_D, we encode V_D into the clean depth latent z_0 using the VAE encoder, while V_I is encoded as the conditioning latent c. For the corresponding depth video $V_D \in \mathbb{R}^{f \times h \times w}$,

we first compute a binary mask $M \in \{0,1\}^{f\times h\times w}$ by thresholding against the maximum depth value $D_{\max}$:

$$M_{i,j,k} = \begin{cases} 1 & \text{if } V_D(i,j,k) < D_{\max} \\ 0 & \text{otherwise} \end{cases}$$

where the mask identifies valid depth regions.

For relative depth estimation training, we convert depth V_D to disparity and normalize it to range $[-1,1]$ by per-video min-max normalization. The mask M is adapted to the VAE latent resolution.

We train the model using the flow matching objective, where the loss is computed only in valid regions ($M = 1$):

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1,\mathbf{c},\mathbf{e}_r,\mathbf{e}_d}\left[M \odot \|\mathbf{u}_\Theta(\mathbf{z}_t,t,\mathbf{c},\mathbf{e}_r,\mathbf{e}_d) - \mathbf{v}_t\|^2\right] \tag{10}$$

where $\odot$ denotes element-wise multiplication.

4 Experiment

4.1 Dataset and Benchmark

Compared to DepthCrafter [8], we train our model on a more compact dataset comprising: (1) VKITTI [2], (2) subsets of TartanAir [33] and TartanGround [27] (using single-direction cameras), and (3) the synthetic subset of OmniWorld [42]. The total training datasets contain around 0.8M frames. For relative depth estimation, we compare against state-of-the-art video and image depth estimation methods, including both discriminative and generative models: ChronoDepth [29], DepthCrafter [8], Depth Any Video [36], Video Depth Anything [3], and Depth Anything V2 [37]. We evaluate on diverse benchmarks encompassing both synthetic and real-world scenarios, with video lengths ranging from 50 to over 100 frames. Specifically, we use Sintel [22], KITTI [5], ScanNet [4] and Bonn [26] for quantitative evaluation. Additionally, we provide qualitative comparisons on challenging domains including animation, gaming, and underwater scenes to demonstrate our method's generalization capability. We evaluate our results using three standard metrics: Root Mean Square Error (RMSE), threshold accuracy ($\delta 1$), and Absolute Relative error (AbsRel).

It is worth noting that the reported performance of these baseline methods often varies across different publications due to discrepancies in their adopted evaluation protocols. To ensure a fair comparison, we conduct inference for all baseline methods and evaluate their predictions using the official evaluation script provided by DepthCrafter [8]. Consequently, our evaluated metrics for these baselines may differ from those reported in their original papers, appearing either lower or higher depending on the specific method and dataset.

4.2 Implementation Details

We build our model upon the pretrained Wan2.1 1.3B T2V model [32]. Training is conducted on 4 H800 GPUs for 8 epochs with a batch size of 1 per GPU, a learning rate of 2×10^{-4}, and gradient accumulation over 32 steps. We employ multi-resolution training to preserve the original aspect ratios of videos while ensuring spatial dimensions are divisible by 32. The temporal length varies from 21 to 77 frames depending on the spatial resolution of each video. To maintain a reasonable computational budget, we use $H \times W \times T = 672 \times 384 \times 77$ as our token budget baseline, adjusting temporal length inversely with spatial resolution. Detailed configurations of this strategy are provided in the supplementary material. During inference, we set diffusion sampling steps to 5.

4.3 Relative Depth Comparison with SOTA Methods

Quantitative Comparison We evaluate different methods through zero-shot video depth estimation experiments, with quantitative results presented in Table 1. Our method achieves state-of-the-art performance on Sintel, KITTI, and Bonn datasets, obtaining the best results in both AbsRel and δ_1 metrics. Specifically, we achieve significant improvements of 16.0% in AbsRel and 10.1% in δ_1 on Sintel compared to the second-best method; Sintel is the benchmark featuring highly dynamic scenes and complex motion patterns. On ScanNet, our method achieves competitive second-best performance, closely matching the top results while using substantially less training data. Notably, our approach accomplishes this superior performance with only 0.8M training samples, representing 6-13× reduction in data requirements compared to other state-of-the-art video depth estimation methods, demonstrating remarkable data efficiency without compromising estimation accuracy.

Table 1: Quantitative comparison of zero-shot video depth estimation with state-of-the-art methods on the Sintel, ScanNet, KITTI and Bonn dataset. The **best** and second-best results are highlighted in bold and underlined, respectively. ↓: lower is better; ↑: higher is better.

Method	Sintel (50 frames)		ScanNet (90 frames)		KITTI (110 frames)		Bonn (110 frames)		Venue	Data size
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑		
Depth Anything V2	0.403	0.547	0.123	0.852	0.102	0.910	0.084	0.947	NeurIPS'24	62.62M
Video Depth Anything	0.383	0.629	**0.075**	**0.954**	0.078	0.950	**0.053**	0.975	CVPR'25	1.35M
ChronoDepth	0.587	0.486	0.159	0.783	0.167	0.759	0.100	0.911	CVPR'25	-
DepthCrafter	0.313	0.680	0.142	0.803	0.105	0.898	0.066	0.971	CVPR'25	10.5M
Depth Any Video	0.300	0.643	0.119	0.865	0.098	0.925	0.063	0.963	ICLR'25	6M
Ours	**0.250**	**0.749**	0.076	0.952	**0.061**	**0.968**	**0.053**	**0.979**	-	0.8M

In addition to full-video evaluation, we also report per-frame metrics, where we perform inference on complete videos but calculate metrics (including scale

Table 2: Per-frame depth estimation accuracy on Sintel. Metrics are computed on individual frames with per-frame scale-shift alignment.

Method	AbsRel↓	RMSE↓	δ_1 ↑
Depth Anything V2	0.226	5.121	0.740
Video Depth Anything	0.209	4.857	0.754
ChronoDepth	0.431	8.565	0.603
DepthCrafter	0.194	4.694	0.784
Depth Any Video	0.287	5.109	0.723
Ours	**0.186**	**4.227**	**0.787**

Table 3: Temporal consistency and long-sequence evaluation on 500-frame ScanNet++ sequences. TAE denotes Temporal Alignment Error.

Method	TAE ↓	AbsRel ↓	RMSE ↓	δ_1 ↑
DepthCrafter	3.07	0.204	0.658	0.589
Depth Any Video	5.69	0.148	0.483	0.814
Ours	**2.61**	**0.131**	**0.441**	**0.832**

and shift alignment) on individual frames to assess per-frame depth estimation accuracy. The results in Table 2 demonstrate that our method outperforms other video depth estimation models in per-frame accuracy, even surpassing Depth Anything V2. This improvement can be attributed to the additional temporal information provided by video context. To further quantify temporal consistency on long sequences, we evaluate 500-frame ScanNet++ [39] videos using Temporal Alignment Error (TAE). As shown in Table 3, ICDepth achieves the lowest TAE and the best long-sequence depth accuracy.

To further assess robustness under unseen distribution shifts, we construct a low-light variant of Sintel by reducing illumination and adding sensor noise. Such low-light videos are not included in our depth training data. As shown in Table 4, ICDepth exhibits the smallest relative δ_1 degradation compared with clean Sintel, suggesting better robustness under unseen low-light conditions.

Qualitative Comparison Figure 3 presents qualitative comparisons across five diverse domains (3D games, 2D cartoons, underwater, indoor, and driv-

Table 4: OOD robustness on the unseen low-light Sintel setting. We report the relative δ_1 degradation from clean Sintel to low-light Sintel. Lower is better.

Method	Rel. δ_1 Drop ↓
Depth Anything V2	13.2%
DepthCrafter	8.4%
Depth Any Video	13.2%
ICDepth	**4.0%**

ing scenes). Our method demonstrates superior capability in producing sharper depth boundaries and recovering fine-grained details consistently missed by competitors. For instance, in underwater scenes (Row 3), we precisely separate fish and coral boundaries, while in driving scenes (Row 5), we uniquely reconstruct bicycles and crowds with clean spatial separation. Temporal analysis (green boxes) further confirms our exceptional consistency without flickering artifacts, validating advantages in boundary precision, temporal coherence, and detail reconstruction. Figure 4 further evaluates depth estimation under challenging low-light conditions. Despite adverse illumination and low visibility in nighttime scenes, our method produces more accurate depth maps.

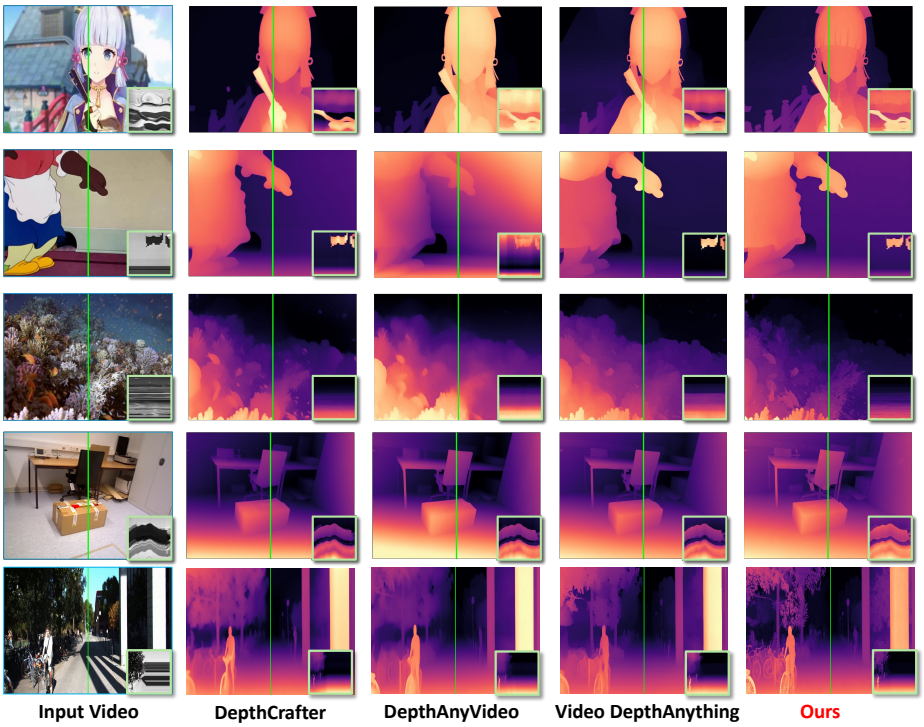

Fig. 3: We compare against representative video depth estimation baselines. For temporal consistency visualization, we show depth profiles (green boxes) extracted along the temporal dimension at the green line locations, illustrating the temporal stability of each method.

4.4 Ablation Studies

To validate the effectiveness of our design, we conduct ablation studies on the Sintel dataset to answer three key questions: **Q1:** *Is ICC superior to stan-*

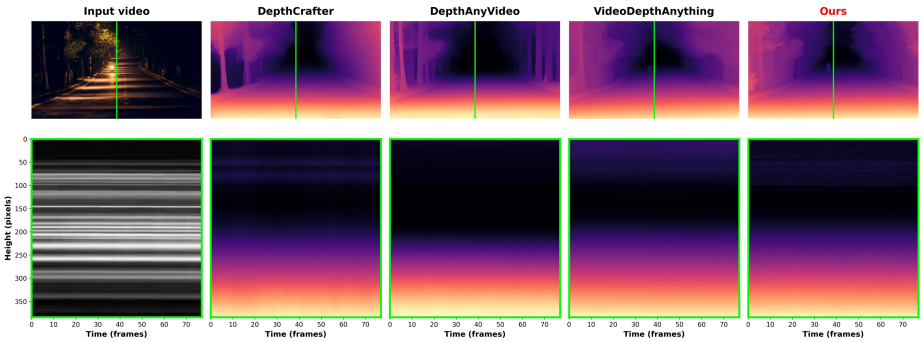

Fig. 4: Qualitative comparison under challenging nighttime conditions. We compare our method with state-of-the-art approaches on nighttime scenes. The second row displays spatio-temporal plots along the green scanlines. Our method demonstrates superior depth accuracy and temporal stability, producing flicker-free results even under adverse illumination and visibility.

dard channel concatenation? **Q2:** *How to adapt ICC framework to VDiT-based depth estimation model?* **Q3:** *Are task-specific priors essential for bridging the generative-perceptive gap?* Results are presented in Table 5.

Q1: ICC vs. Channel Concatenation. We compare our ICC framework against the standard adaptation approach using Wan2.1 Control [25], where RGB and depth features are concatenated along the channel dimension and processed through a modified input projection layer. Table 5 shows ICC dramatically outperforms channel concatenation. This validates that ICC's token-level conditioning is crucial: the VDiT's attention mechanism can directly model RGB-depth correspondences at matching spatial-temporal locations, rather than relying on implicit late-stage fusion through separate channels.

Q2: Critical Modifications to ICC framework. We conduct ablation studies to identify essential modifications for adapting the ICC framework to VDiT-based depth estimation, evaluating three variants of our proposed SAND-Attention mechanism.

Naive full attention. Replacing SAND-Attention with standard full attention (lacking both RoPE Alignment and Decoupled Attention) results in complete performance degradation, demonstrating the fundamental inadequacy of the original ICC framework for this task.

RoPE Alignment. Removing only the shared RoPE alignment also causes severe failure, confirming that establishing precise spatial-temporal correspondence between RGB and depth tokens at shared (x, y, t) coordinates is paramount. Without this explicit alignment, the model cannot learn proper cross-modal relationships.

Decoupled Attention. Removing the attention decoupling mechanism yields moderate but significant degradation (AbsRel: 0.262 vs. 0.250). The substantial drop in δ_1 (0.710 vs. 0.749) particularly highlights how bidirectional attention

Table 5: Ablation study of our proposed components conducted on Sintel dataset.

Method Variant	AbsRel ↓	RMSE ↓	δ_1 ↑
Full model (Ours)	**0.250**	**5.155**	**0.749**
In-Context Conditioning (ICC)			
Replace with channel concat	0.367	5.956	0.654
SAND-Attention			
Replace with full attention	0.413	6.078	0.443
w/o RoPE Alignment	0.410	5.999	0.450
w/o Decoupled Attention	0.262	5.196	0.710
Semantic-Resolution Aware feature modulation (SRFM)			
w/o SRFM	0.306	6.325	0.696
w/o DINOv2 features	0.269	5.730	0.709
w/o Resolution Embedding	0.264	5.565	0.728

flow corrupts confident predictions through noise contamination, though this effect proves less critical than spatial misalignment.

Q3: Necessity of Task-Specific Priors. We conduct ablation studies to investigate the necessity of semantic and resolution information injection, evaluating three variants of our SRFM module.

Full SRFM removal. Complete removal of SRFM causes substantial performance degradation, demonstrating that T2V generative features alone are insufficient for precise depth estimation. The magnitude of this degradation exceeds that of any individual SRFM component, revealing synergistic effects between DINOv2 semantic priors and resolution embeddings.

DINOv2 features. Removing DINOv2 features results in moderate performance decline. This validates that DINOv2's rich semantic understanding significantly enhances geometric reasoning capabilities.

Resolution Embeddings. The ablation of resolution embeddings yields a measurable performance drop, confirming their role in providing complementary spatial context. This information proves particularly critical on datasets with unconventional resolutions, such as KITTI, where explicit resolution conditioning enhances the model's robustness and generalization capability to diverse and unseen spatial dimensions.

4.5 Efficiency Analysis

We compare the computational overhead of our method against both discriminative and generative video depth estimation models in Table 6. While discriminative feed-forward models such as Video Depth Anything are naturally faster, ICDepth remains efficient within the generative diffusion paradigm. Compared with Depth Any Video, ICDepth achieves comparable inference speed while substantially reducing GPU memory consumption, obtaining stronger depth estimation accuracy.

Table 6: Efficiency at resolution 480×640, 53 video frames.

Method	Type	Time (s)↓	FPS↑	Mem (GB)↓
Video Depth Anything	Discr.	1.82	29.1	8.5
Depth Any Video	Gen.	11.19	4.7	33.0
Ours	Gen.	11.85	4.5	**11.0**

Moreover, as shown in Table 7, ICDepth achieves a strong accuracy–efficiency balance with only 3–5 sampling steps.

Table 7: Quantitative results with different sampling steps on Sintel.

Steps	**AbsRel**↓	**RMSE**↓	δ_1↑
3	0.262	5.330	0.746
5	0.250	5.154	**0.749**
10	0.229	4.584	0.735
20	**0.217**	**4.469**	0.721

5 Conclusion

We introduce ICDepth, a novel framework adapting pre-trained T2V diffusion transformers via In-Context Conditioning for video depth estimation. Through SAND-Attention and SRFM (tackling geometric ambiguity), our method preserves pre-trained spatial-temporal priors while adding perceptual precision.

ICDepth achieves state-of-the-art results on multiple benchmarks with remarkable data efficiency: trained on only 0.8M frames (6-13× less than competing generative methods), it delivers superior temporal consistency, per-frame accuracy, and strong zero-shot generalization to diverse domains. This demonstrates that foundation model priors can be effectively transferred to dense prediction tasks through careful architectural design.

Acknowledgments

The work was supported by the Research Grants Council of HKSAR under grant number AoE/E-601/24-N.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

2. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
3. 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 Computer Vision and Pattern Recognition Conference. pp. 22831–22840 (2025)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
6. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long context tuning for video generation. arXiv preprint arXiv:2503.10589 (2025)
7. He, X., Liu, Q., Ye, Z., Ye, W., Wang, Q., Wang, X., Chen, Q., Wan, P., Zhang, D., Gai, K.: Fulldit2: Efficient in-context conditioning for video diffusion transformers. arXiv preprint arXiv:2506.04213 (2025)
8. 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 Computer Vision and Pattern Recognition Conference. pp. 2005–2015 (2025)
9. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
10. Ju, X., Ye, W., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Xu, Q.: Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907 (2025)
11. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
12. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1611–1621 (2021)
13. Kuang, Z., Zhang, T., Zhang, K., Tan, H., Bi, S., Hu, Y., Xu, Z., Hasan, M., Wetzstein, G., Luan, F.: Buffer anytime: Zero-shot video depth and normal from image priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17660–17670 (2025)
14. Li, Z.Y., Du, R., Yan, J., Zhuo, L., Li, Z., Gao, P., Ma, Z., Cheng, M.M.: Visualcloze: A universal image generation framework via visual in-context learning. arXiv preprint arXiv:2504.07960 (2025)
15. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061 (2025)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
17. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (ToG) **39**(4), 71–1 (2020)
18. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable video generation: A survey. arXiv preprint arXiv:2507.16869 (2025)

19. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4d creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025)
20. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
21. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
22. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016)
23. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. arXiv preprint arXiv:2504.16915 (2025)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
25. PAI, A.: Wan2.1-fun-1.3b-control. `https://huggingface.co/alibaba-pai/Wan2.1-Fun-1.3B-Control` (2024)
26. Palazzolo, E., Behley, J., Lottes, P., Giguere, P., Stachniss, C.: Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7855–7862. IEEE (2019)
27. Patel, M., Yang, F., Qiu, Y., Cadena, C., Scherer, S., Hutter, M., Wang, W.: Tartanground: A large-scale dataset for ground robot perception and navigation. arXiv preprint arXiv:2505.10696 (2025)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
29. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22841–22852 (2025)
30. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
31. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
32. Wan Team, 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)
33. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
34. Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., Xian, K., Lin, G.: Neural video depth stabilizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9466–9476 (2023)

35. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160 (2025)
36. Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., He, T.: Depth any video with scalable synthetic data. In: International Conference on Learning Representations. vol. 2025, pp. 97335–97349 (2025)
37. 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)
38. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025)
39. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
40. Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y.: Exploiting temporal consistency for real-time video depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1725–1734 (2019)
41. Zhang, Z., Cole, F., Tucker, R., Freeman, W.T., Dekel, T.: Consistent depth of moving objects in video. ACM Transactions on Graphics (ToG) **40**(4), 1–12 (2021)
42. 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)