DVD: Deterministic Video Depth Estimation with Generative Priors

Hongfei Zhang [1 †] **Harold H. Chen** [1 2 †] **Chenfei Liao** [1 †] **Jing He** [1 †] **Zixin Zhang** [1] **Haodong Li** [3] **Yihao Liang** [4] **Kanghao Chen** [1] **Bin Ren** [5] **Xu Zheng** [1] **Shuai Yang** [1] **Kun Zhou** [6] **Yinchuan Li** [7] **Nicu Sebe** [8] **Ying-Cong Chen** [1 2 ‡]

[1]HKUST(GZ) [2]HKUST [3]UCSD [4]Princeton University [5]MBZUAI [6]SZU [7]Knowin [8]UniTrento

[†]Equal Contribution [‡]Corresponding Author

Abstract

Existing video depth estimation faces a fundamental trade-off: *generative models* offer rich priors but suffer from scale drift and structural inconsistencies (often termed geometric hallucinations), while *discriminative models* provide stable predictions but demand massive labeled datasets to resolve semantic ambiguities. To mitigate this trade-off, we present `DVD`, to the best of our knowledge, the *first* framework to deterministically adapt pre-trained video diffusion models into single-pass depth regressors. `DVD` introduces three targeted strategies to effectively ground these generative priors: (***i***) repurposing the diffusion **timestep as a structural anchor** to balance global stability with high-frequency details; (***ii***) applying **latent manifold rectification (LMR)**, a parameter-free constraint to counteract regression-induced over-smoothing, preserving sharp boundaries and temporal structural integrity; and (***iii***) exploiting the model's inherent **global affine coherence** for lightweight overlap-based alignment, enabling seamless long-video inference. Extensive experiments demonstrate that `DVD` achieves *state-of-the-art* zero-shot performance across four standard video depth benchmarks. Crucially, by inheriting the robust spatio-temporal priors of video foundation models, this deterministic paradigm proves highly data-efficient, requiring only 367K task-specific downstream frames to adapt. We fully release our pipeline and training code to benefit the open-source community in `DVD Repo`.

1. Introduction

Depth estimation serves as a fundamental building block for 3D scene understanding, underpinning applications (Chen et al., 2025c; Charatan et al., 2024; Xu et al., 2025a) from autonomous driving to robotic manipulation. While image-based depth estimation has matured significantly (Bochkovskii et al., 2024; Yang et al., 2024c;b; Piccinelli et al., 2024; Yin et al., 2023; Fu et al., 2024), elevating this capability to the video domain remains a formidable challenge. The transition from static images to dynamic video is non-trivial; it demands not only precise geometric reasoning per frame but also rigorous temporal consistency. In real-world scenarios characterized by camera motion and dynamic objects, maintaining this consistency without sacrificing high-frequency geometric details is a persistent bottleneck.

Recent advances in video depth estimation have predominantly followed two paradigms, each constrained by inherent limitations that hinder their broader applicability, as shown in Fig. 1 (*Top*). **(I) Diffusion-based generative models** (Hu et al., 2025; Shao et al., 2025; Yang et al., 2024a) (*e.g.*, DepthCrafter) leverage pre-trained video foundation models to capture rich spatio-temporal priors, enabling impressive zero-shot generalization. However, their reliance on stochastic sampling introduces temporal uncertainties that limit their stability and reliability in real-world applications. Moreover, the generative nature of these models sometimes tends to prioritize visual plausibility over geometric accuracy, leading to *geometric hallucinations*, a failure to maintain precise and globally consistent geometry over time (Shao et al., 2025; Yang et al., 2024a; Hu et al., 2025). **(II) Discriminative ViT-based models** (Yang et al., 2024c; Chen et al., 2025b) (*e.g.*, Video Depth Anything, VDA), on the other hand, provide high inference efficiency and deterministic outputs. However, learning geometry strictly from dense annotations, they frequently suffer from *semantic ambiguity*, misinterpreting motion blur or textureless regions as structural boundaries. To overcome this ambiguity, discriminative paradigms rely on diversified and massive-scale downstream annotations (Chen et al., 2025b; Yang et al., 2024b;c; Birkl et al., 2023).

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

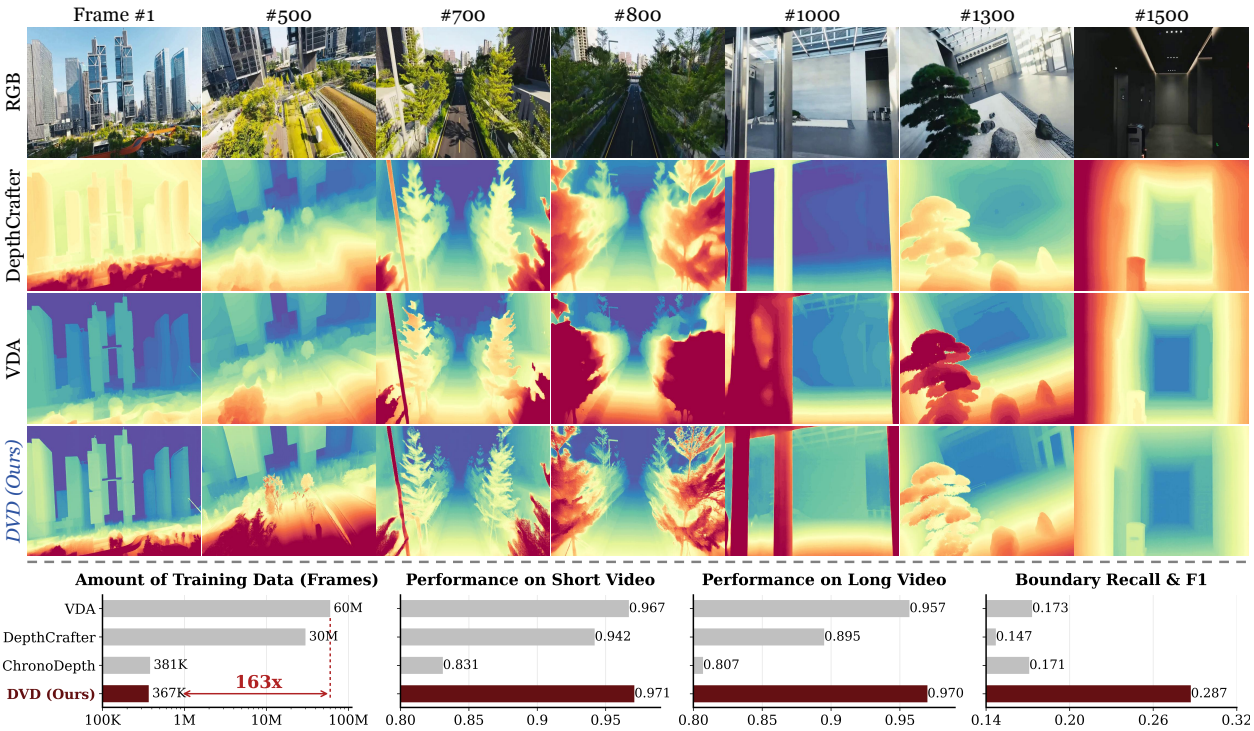

Figure 1. (***Top***) Illustration of the paradigm trade-off in video depth estimation. Representative generative models (*e.g.*, DepthCrafter (Hu et al., 2025)) often struggle with structural inconsistencies (hallucinations) and scale drift, whereas discriminative baselines (*e.g.*, VDA (Chen et al., 2025b)) can be susceptible to semantic ambiguities. As exemplified in this 1500-frame in-the-wild sequence, `DVD` effectively balances this trade-off, delivering consistent and high-fidelity geometry. (***Bottom***) `DVD` achieves superior performance on both short and long videos (Geiger et al., 2012; Dai et al., 2017; Palazzolo et al., 2019), successfully exploiting the rich priors implicit in video foundation models using minimal task-specific data ($< 1\%$ of VDA's setting).

This heavy data requirement not only raises barriers to scalability and reproducibility but also restricts their adaptability in broader, data-scarce scenarios. These challenges lead to our key research question:

Can we design a video depth estimation framework that effectively balances the structural stability of discriminative models and the rich spatio-temporal priors of generative approaches, while remaining efficient and scalable?

In response, we present `DVD`, a new framework that achieves deterministic video depth estimation with generative priors. Departing from the conventional stochastic generative paradigm, `DVD` explores a deterministic adaptation of pre-trained video diffusion models that learns a direct mapping from RGB latents to depth latents. This paradigm shift introduces a new design point: leveraging the backbone's rich semantic priors to reduce motion-induced ambiguity while enforcing a regression objective that predicts geometrically consistent depth, effectively mitigating generative hallucinations. However, directly extending previous deterministic adaptation from static images (Lee et al., 2024; He et al., 2025; Xu et al., 2024) to dynamic videos presents unique challenges: a naive regression is not merely prone to *blurring*, but suffers from *structural instability* and *scalability* issues (Hu et al., 2025; Shao et al., 2025).

To address these bottlenecks, `DVD` introduces a video deterministic adaptation paradigm built upon three key mechanisms: ❶ **Timestep as a Structural Anchor:** Instead of treating the diffusion timestep t merely as a noise-level index, we empirically observe that for video diffusion priors, specific timesteps encode distinct spatio-temporal frequency preferences. Anchoring at an optimal state effectively balances low-frequency geometric stability with high-frequency spatial details. ❷ **Latent Manifold Rectification (LMR):** We then find that standard regression-induced over-smoothing is particularly detrimental to 3D video geometry, eroding physical boundaries and temporal structural integrity. To this end, we introduce a parameter-free differential supervision that effectively restores sharp structural details and coherent motion. ❸ **Global Affine Coherence:** We further uncover that our deterministic backbone inherently bounds inter-window divergence to global affine variations. This property enables a lightweight, affine-alignment strategy for long-video inference, bypassing the need for complex latent stitching (Hu et al., 2025; Shao et al., 2025). Ultimately, `DVD` provides a highly effective empirical paradigm to mitigate the ambiguity-hallucination trade-off, by adapting video generation models into deterministic regressors. Notably, `DVD` effectively unlocks the rich geometric priors embedded in

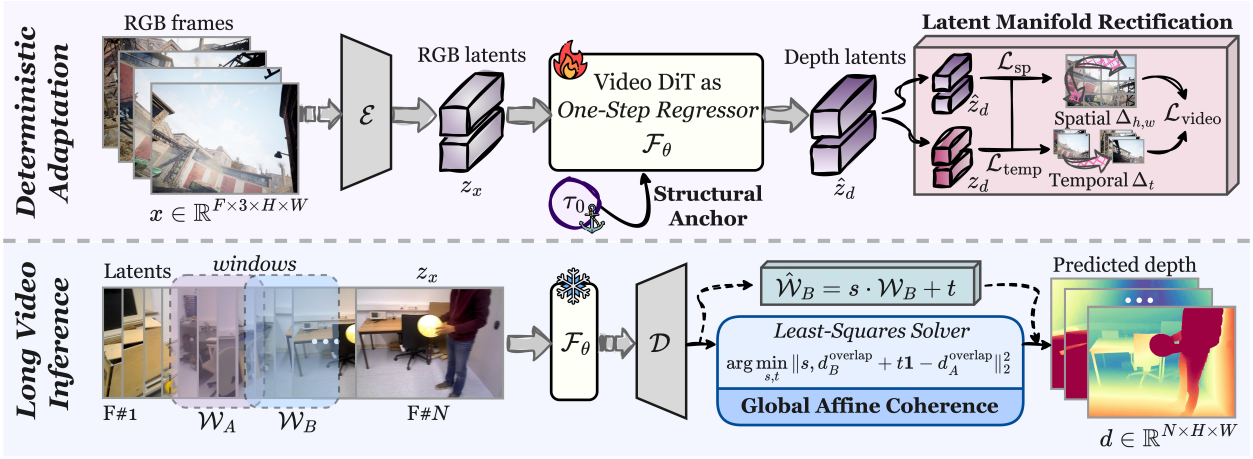

Figure 2. **Overview of DVD.** (***Top***) A video DiT ($\mathcal{F}_\theta$) performs single-pass depth regression, modulated by a structural anchor (τ_0). Latent manifold rectification (LMR) mitigates mean collapse via differential constraints. (***Bottom***) For long video depth estimation, overlapping windows ($\mathcal{W}_A, \mathcal{W}_B$) are seamlessly aligned using a closed-form least-squares solver that leverages global affine coherence.

foundation models using minimal task-specific data, while achieving *state-of-the-art* zero-shot video depth estimation. This establishes a highly efficient and scalable adaptation route for future 3D perception. In brief, our contributions are summarized as follows:

- ❏ **Bottleneck Identification.** Through analysis of existing video depth estimation paradigms, we identify key bottlenecks: geometric hallucinations in generative models and semantic ambiguity in discriminative models, which hinder scalability and practical deployment.

- ❏ **Our Solution.** We present DVD, to the best of our knowledge, the first framework to deterministically adapt pre-trained video diffusion models into single-pass depth regressors. To effectively ground these generative priors, DVD introduces three targeted designs based on our empirical observations: (***i***) repurposing the diffusion *timestep as a structural anchor* to balance geometric stability and detail precision; (***ii***) *latent manifold rectification (LMR)* to counteract regression-induced over-smoothing and preserve sharp boundaries; and (***iii***) exploiting inherent *global affine coherence* for lightweight overlap-based alignment in robust long-video inference.

- ❏ **Empirical Validation.** Extensive experiments across four real-world benchmarks demonstrate that DVD achieves ❶ **superior performance:** achieving *state-of-the-art* zero-shot geometric fidelity; ❷ **compelling efficiency:** effectively unlocking pre-trained world priors with minimal downstream data while maintaining comparable inference speed; and ❸ **robust scalability:** enabling robust inference on long videos and generalizing to unconstrained open-world domains.

2. Preliminary

2.1. Problem Formulation

We formalize video depth estimation as a mapping from an input RGB sequence $x \in \mathbb{R}^{F\times 3\times H\times W}$ to its corresponding depth sequence $d \in \mathbb{R}^{F\times H\times W}$, where F denotes the frame count. To exploit the rich semantic priors of large-scale pre-trained models, we operate within a compressed latent manifold. Specifically, a frozen variational autoencoder (VAE) encoder $\mathcal{E}(\cdot)$ projects both RGB and depth into a unified latent space:

$$z_x = \mathcal{E}(x) \in \mathbb{R}^{f\times c\times h\times w}, \quad z_d = \mathcal{E}(d) \in \mathbb{R}^{f\times c\times h\times w}, \tag{1}$$

where c, f, h, w represent the latent channels and downsampled dimensions, respectively. Our objective is to learn a deterministic mapping $\Phi : z_x \mapsto z_d$ that recovers the geometric structure directly in the latent space. The final depth $\hat{d}$ is reconstructed via the frozen VAE decoder $\hat{d} = \mathcal{D}(\hat{z}_d)$.

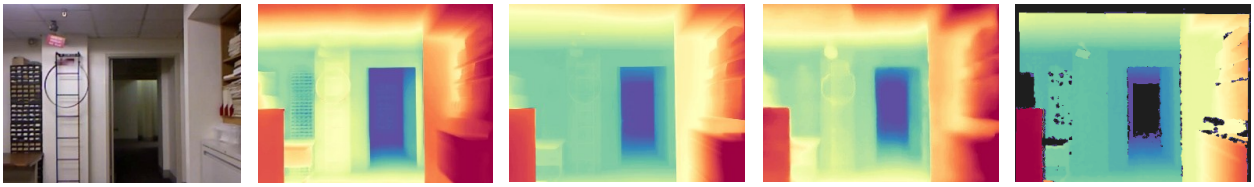

Figure 3. **Timestep as a structural anchor.** Visualizations on NYU (Nathan Silberman & Fergus, 2012) demonstrate a fidelity-stability trade-off. Late timestep ($\tau = 0.0$) recovers sharp boundaries but lacks global consistency, whereas early timestep ($\tau = 0.8$) causes detail loss (*e.g.*, blur). An optimal anchor ($\tau = 0.5$) balances these regimes, achieving a trade-off between detail recovery and metric accuracy. More detailed quantitative analyses are shown in Fig. 13.

2.2. Diffusion as Deterministic Regressor

Role of t in Rectified Flow. In traditional rectified flow (RF) (Liu et al., 2022; Lipman et al., 2022), the time variable $t \in [0, 1]$ explicitly parameterizes a noise interpolation trajectory between data distribution $z_0 \sim p_{\text{data}}$ and Gaussian noise $z_1 \sim \mathcal{N}(0, I)$. RF defines a linear interpolation trajectory $z_t = (1 - t)z_0 + tz_1$, where the scalar timestep $t \in [0, 1]$ explicitly parameterizes the corruption level. The network v_θ is trained to predict the velocity field of this flow by minimizing:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}[\|v_\theta(z_t, t) - (z_1 - z_0)\|^2]. \tag{2}$$

During standard generative inference, one produces samples by solving the ordinary differential equation (ODE) $dz_t/dt = v_\theta(z_t, t)$ via numerical integration from $t = 1$ to $t = 0$.

Deterministic Adaptation. Recent works in the image domain (He et al., 2024; 2025; Xu et al., 2025b) repurpose diffusion backbones as *one-step deterministic regressors*. Instead of iterative ODE integration over a noise trajectory, the network $\mathcal{F}_\theta$ performs a direct functional mapping. Formally, given the RGB latent z_x and a timestep condition t, depth is deterministically predicted in a single forward pass:

$$\hat{z}_d = \mathcal{F}_\theta(z_x, t). \tag{3}$$

Building upon this static-image formulation, Section §3 details how DVD extends this paradigm to videos, along with uncovering a crucial functional shift for t to preserve geometric consistency.

3. Methodology

3.1. Overall Framework

Existing video depth estimation methods are typically polarized: generative diffusion models offer rich spatio-temporal priors but suffer from stochastic geometric hallucinations, while discriminative regressors provide stable outputs but demand massive labeled datasets to resolve semantic ambiguities. To bridge this gap, we propose DVD, a novel framework that unites the generalization power of generative priors with the structural stability of deterministic regression, as shown in Fig. 2. Formally, given an input RGB video x, a VAE encoder $\mathcal{E}$ extracts the latent representation z_x. This latent sequence is then processed by a pre-trained video diffusion backbone $\mathcal{F}_\theta$. Instead of performing iterative stochastic denoising, DVD executes a single-pass deterministic mapping to predict the depth latent $\hat{z}_d$, modulated by a conditioning timestep τ:

$$\hat{z}_d = \mathcal{F}_\theta(z_x, \tau(t)). \tag{4}$$

To achieve high-fidelity depth estimation, DVD introduces three core designs tailored to the latent dynamics of video diffusion backbones. First, we repurpose the diffusion timestep τ as a *structural anchor* (§3.2) to govern the backbone's geometric operating regime, balancing low-frequency stability with high-frequency details. Then, we introduce *latent manifold rectification* (§3.3), a parameter-free supervision mechanism that enforces differential consistency to mitigate regression-induced mean collapse and sharpen spatio-temporal boundaries. Finally, we present *global affine coherence* (§3.4), an inherent property of our deterministic backbone that effectively bounds inter-window divergence, enabling seamless, affine-alignment inference for long-duration videos. We next detail the empirical observations and technical formulations that motivate these designs.

3.2. Timestep as Structural Anchor

In single-image deterministic adaptation (He et al., 2024; Xu et al., 2024; He et al., 2025), the diffusion timestep is often fixed at the terminal state ($t = 1$) or absorbed entirely. However, when applied to video backbones, we observe significant over-smoothing of geometric structures (Fig. 3). We attribute this to the spectral bias inherent in pre-trained diffusion priors (Kingma et al., 2021; Choi et al., 2022; Hang et al., 2025; 2023; Ho et al., 2022), where the timestep implicitly controls a frequency preference through the signal-to-noise ratio (SNR): higher timesteps favor low-frequency global structures, while lower timesteps encourage the recovery of high-frequency local details.

To exploit this property in deterministic adaptation, we replace the dynamic timestep t with a fixed conditioning state τ_0, forming a **structural anchor**:

$$\hat{z}_d = \mathcal{F}_\theta\left(z_x;\, \mathbf{e}_\phi(\tau_0)\right), \tag{5}$$

where $\mathbf{e}_\phi(\cdot)$ is the sinusoidal embedding.

Fidelity-Stability Trade-off. Our key finding is that the choice of τ_0 induces a strict *fidelity-stability trade-off* that persists even after fine-tuning converges. As shown in Fig. 3, early timestep (*e.g.*, $\tau = 0.8$) biases the model toward low-frequency global structures (stable but blurry), while late timestep (*e.g.*, $\tau = 0.0$) amplifies high-frequency details (sharp but unstable). Among the broadly similar embeddings in Fig. 4, anchoring at mid-range timestep (*e.g.*, $\tau = 0.5$) offers a low-variation conditioning region, better balancing global coherence with local sharpness (see Fig. 13 for more details). We further found that completely ablating this conditioning or re-initializing significantly degrades performance, confirming τ indexes an irreplaceable pre-trained geometric prior (Appendix §E).

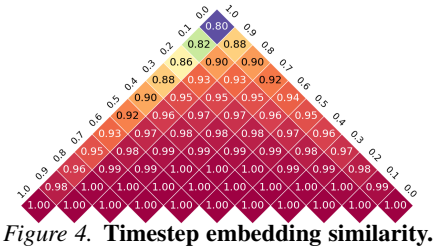

Figure 4. **Timestep embedding similarity.** Cosine similarity matrix of timestep embeddings ($t \in [0, 1]$, stride 0.1). While embeddings are broadly consistent, mid-range timesteps exhibit high similarity with a wider range of states.

3.3. Latent Manifold Rectification

Mean Collapse. Minimizing a pointwise loss to map RGB latents z_x to depth latents z_d inherently drives the predictor toward the conditional expectation $\mathbb{E}[z_d|z_x]$ (Ma et al., 2025; Song et al., 2020a;b; Liu et al., 2022). In ambiguous or occluded regions, this regression-to-the-mean collapses multi-modal geometric hypotheses (Papyan et al., 2020; Zhu et al., 2021), washing out high-frequency structural details. Notably, this degradation is further amplified under the spatio-temporal setting: the suppressed high-frequency differentials propagate and accumulate temporally, manifesting as progressive boundary erosion and severe motion flickering, as illustrated in Fig. 5.

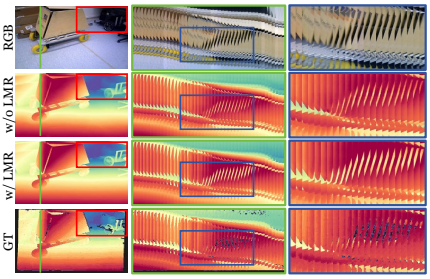

Figure 5. **LMR mitigates mean collapse.** Naive regression (*2nd Row*) exhibits mean collapse, losing high-frequency details. In contrast, our LMR (*3rd Row*) enforces differential constraints to rectify the latent manifold, recovering both sharp spatial boundaries and temporal coherence. Quantitative analyses are in Fig. 11.

Differential Manifold Constraints. While gradient constraints are standard in dense prediction, we uncover their indispensable role in deterministic video diffusion adaptation. To counteract regression-induced mean collapse, we introduce **latent manifold rectification (LMR)**. This parameter-free strategy preserves local structures by jointly enforcing spatial and temporal first-order consistency directly in the VAE latent space:

$$\mathcal{L}_{\text{sp}} = \frac{1}{F \cdot \Omega} \sum_{f=1}^{F} \sum_{\partial \in \{\nabla_h, \nabla_w\}} \|\partial \hat{z}_d^f - \partial z_d^f\|_1, \quad \mathcal{L}_{\text{temp}} = \frac{1}{(F-1) \cdot \Omega} \sum_{f=2}^{F} \|\nabla_t \hat{z}_d^f - \nabla_t z_d^f\|_1. \tag{6}$$

The overall objective integrates differential rectification with global consistency:

$$\mathcal{L}_{\text{video}} = \|\hat{z}_d - z_d\|_2 + \lambda_{\text{sp}} \mathcal{L}_{\text{sp}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}}. \tag{7}$$

Here, $\mathcal{L}_2$ anchors the global geometry, while LMR terms act as a vital safeguard, preserving latent high-frequency structures and temporal dynamics against the smoothing effects of deterministic regression, as shown in Fig. 11. Additional ablations supporting the role of LMR for deterministic regression are provided in Appendix §E.

Table 1. **Zero-shot video depth estimation.** The **best** and the second-best results are highlighted.

Method	Train Frames	KITTI		ScanNet		Bonn		Sintel	
		AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
DAv2-L (Yang et al., 2024c)	-	10.9	0.913	6.4	0.967	6.9	0.957	50.0	0.557
Marigold v1.1 (Ke et al., 2025c)	-	9.5	0.936	7.6	0.940	9.5	0.936	65.1	0.411
RollingDepth (Ke et al., 2025a)	-	9.8	0.912	5.8	0.964	5.9	0.966	43.7	0.500
ChronoDepth (Shao et al., 2025)	381K	15.2	0.775	17.1	0.818	16.8	0.901	52.8	0.504
DepthCrafter (Hu et al., 2025)	∼ 30M	9.9	0.907	7.1	0.960	5.9	0.959	**37.1**	0.664
VDA (Chen et al., 2025b)	60M	7.2	0.963	5.8	0.968	**4.7**	0.970	39.7	0.654
`DVD (Ours)`	367K	**6.7**	**0.967**	**5.5**	**0.974**	**4.7**	**0.971**	44.5	**0.667**

Table 2. **Zero-shot long video depth estimation.** Paradigm denotes the backbone type and the paradigm of each method, where "Diff.", "ViT", "D", and "G" denote Diffusion-based, ViT-based, Discriminative, and Generative, respectively.

Method	Paradigm	Bonn		ScanNet		KITTI	
		AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
DAv2-L (Yang et al., 2024c)	ViT.+D	8.7	0.952	9.5	0.940	11.9	0.879
Marigold v1.1 (Ke et al., 2025c)	Diff.+G	11.6	0.890	12.0	0.870	24.7	0.569
RollingDepth (Ke et al., 2025a)	Diff.+G	7.2	0.966	7.5	0.957	11.1	0.911
ChronoDepth (Shao et al., 2025)	Diff.+G	17.3	0.859	21.2	0.715	13.0	0.846
DepthCrafter (Hu et al., 2025)	Diff.+G	8.5	0.962	11.4	0.866	12.0	0.858
VDA (Chen et al., 2025b)	ViT+D	6.6	0.971	**7.3**	0.972	9.6	0.940
`DVD (Ours)`	Diff.+D	**5.3**	**0.978**	**7.3**	**0.977**	**7.6**	**0.956**

3.4. Global Affine Coherence

Long-video inference requires sliding-window processing due to memory constraints. For generative diffusion-based methods, independently sampled windows often follow different denoising trajectories, leading to stochastic scale drift and temporally inconsistent geometry. In contrast, given a fixed input window, DVD produces a deterministic prediction (*i.e.,* no sampling-induced variance).

Global Affine Coherence. Specifically, we empirically observe that the discrepancy between adjacent overlapping windows is not arbitrary. Since the latent predictions are produced by the same deterministic mapping, the remaining mismatch is mainly induced by context-dependent decoding and normalization effects in the VAE decoder. In typical overlapping-window inference, we empirically find that the dominant inter-window discrepancy can often be well approximated by a global affine transformation. We refer to this property as **global affine coherence**. As shown in Fig. 6, DVD exhibits much lower overlap alignment error than generative baselines. Additional analysis in Appendix §C further shows that unaligned overlapping predictions form a near-linear relationship and that affine alignment yields well-bounded residuals, validating the affine assumption.

Long-Video Inference. Exploiting this well-bounded, affine-invariant property, we propose a lightweight, parameter-free *affine-alignment* strategy for sliding-window inference. Let $\mathcal{W}_A$ and $\mathcal{W}_B$ denote the decoded depth tensors for the preceding and current windows, respectively. `DVD` extracts the flattened depth predictions $\mathbf{d}_A^{\text{overlap}}, \mathbf{d}_B^{\text{overlap}} \in \mathbb{R}^N$ from their N overlapping pixels. To align $\mathcal{W}_B$ to the canonical scale of $\mathcal{W}_A$, `DVD` estimates a global scale s and shift t by minimizing the least-squares objective over the overlap:

$$\arg\min_{s,t} \|s\mathbf{d}_B^{\text{overlap}} + t\mathbf{1} - \mathbf{d}_A^{\text{overlap}}\|_2^2. \tag{8}$$

This yields a closed-form solution:

$$s = \frac{\text{Cov}(\mathbf{d}_A^{\text{overlap}}, \mathbf{d}_B^{\text{overlap}})}{\text{Var}(\mathbf{d}_B^{\text{overlap}})}, \quad t = \mu_A - s\mu_B, \tag{9}$$

where μ_A and μ_B denote the mean values of the overlapping regions. The estimated transformation is applied to the entire window ($\hat{\mathcal{W}}_B = s \cdot \mathcal{W}_B + t$), followed by linear blending in the overlap region. This simple calibration enables seamless long-video inference without requiring feature matching, optical flow, or recurrent temporal modules (Hu et al., 2025; Yang et al., 2024a; Shao et al., 2025).

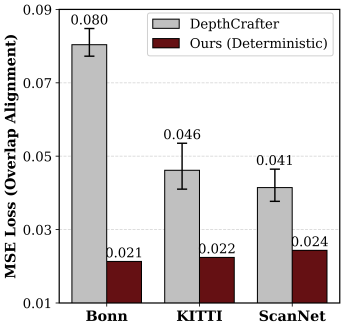

Figure 6. **Inter-window overlap consistency** (Geiger et al., 2012; Palazzolo et al., 2019; Dai et al., 2017). Unlike generative baselines with high alignment error and variance, our deterministic regression yields minimal MSE and zero variance, validating our global affine coherence that bounds inter-window discrepancies to linear transformations.

3.5. Image-Video Joint Training

Training exclusively on video data often compromises spatial sharpness, whereas sequential fine-tuning (image → video) may risk forgetting per-frame details. To bypass this trade-off, we optimize `DVD` via an *image-video joint training*

Table 3. **Zero-shot boundary metrics.** Higher values indicate sharper boundaries and finer details.

Method	Bonn		ScanNet		KITTI	
	B-Recall ↑	B-F1 ↑	B-Recall ↑	B-F1 ↑	B-Recall ↑	B-F1 ↑
ChronoDepth (Shao et al., 2025)	0.221	0.319	0.144	0.204	0.049	0.090
DepthCrafter (Hu et al., 2025)	0.282	0.185	0.115	0.173	0.082	0.044
VDA (Chen et al., 2025b)	0.223	0.325	0.147	0.210	0.047	0.088
DVD (Ours)	**0.336**	**0.422**	**0.208**	**0.259**	**0.217**	**0.285**

Table 4. **Zero-shot single-image depth estimation results.**

Method	KITTI		DIODE		NYUv2	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
ChronoDepth (Shao et al., 2025)	-	-	80.3	0.549	21.2	0.767
DepthCrafter (Hu et al., 2025)	11.0	0.877	53.3	0.592	17.1	0.868
VDA (Chen et al., 2025b)	8.3	0.933	27.0	0.730	**4.7**	**0.977**
DVD (Ours)	**8.1**	**0.944**	**23.1**	**0.738**	5.5	0.969

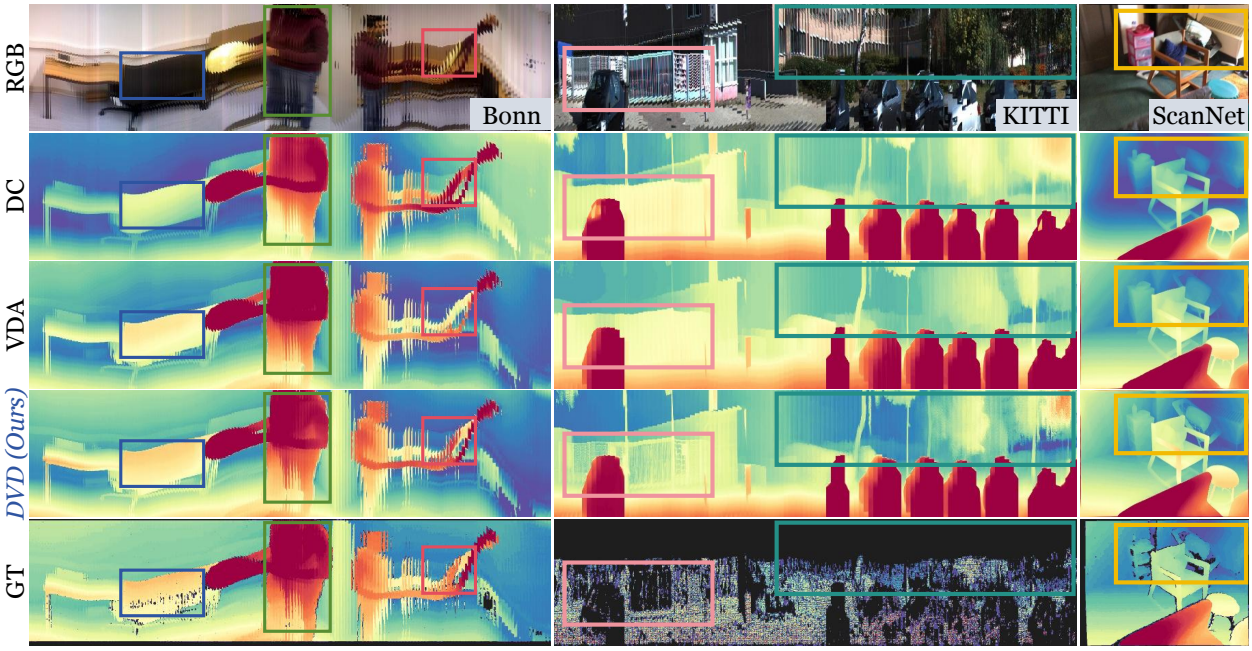

Figure 7. **Qualitative comparison on indoor and outdoor scenes.** DVD consistently produces higher fidelity depth with noticeably sharper structural boundaries.

strategy. By constructing batches comprising both static images ($F = 1$) and dynamic video sequences, the images act as high-frequency spatial anchors while the videos enforce temporal coherence. The unified objective is simply formulated as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{video}} + \lambda_{\text{image}} \mathcal{L}_{\text{image}}. \tag{10}$$

This simple yet effective strategy enables DVD to maintain the spatial quality of diffusion priors while achieving robust temporal stability.

4. Experiments

4.1. Experimental Setup

Implementation Details. We adopt Wan2.1-1.3B (Wan et al., 2025) as DVD's backbone, fine-tuned via LoRA following (He et al., 2025). We train our models strictly using public synthetic datasets: video clips from TartanAir (Wang et al., 2020) and Virtual KITTI (Gaidon et al., 2016), alongside static images from Hypersim (Roberts et al., 2021) and Virtual KITTI. The entire framework converges in under 36 hours on 8 H100 GPUs, demonstrating higher training efficiency than prior works (Chen et al., 2025b; Hu et al., 2025). More details are provided in Appendix §D.

Evaluation. To assess both temporal consistency and per-frame accuracy, we conduct comprehensive evaluations across two settings: ❶ **video datasets:** KITTI (Geiger et al., 2012), ScanNet (Dai et al., 2017), Bonn (Palazzolo et al., 2019), and Sintel (Butler et al., 2012); and ❷ **image datasets:** KITTI (Geiger et al., 2012), DIODE (Vasiljevic

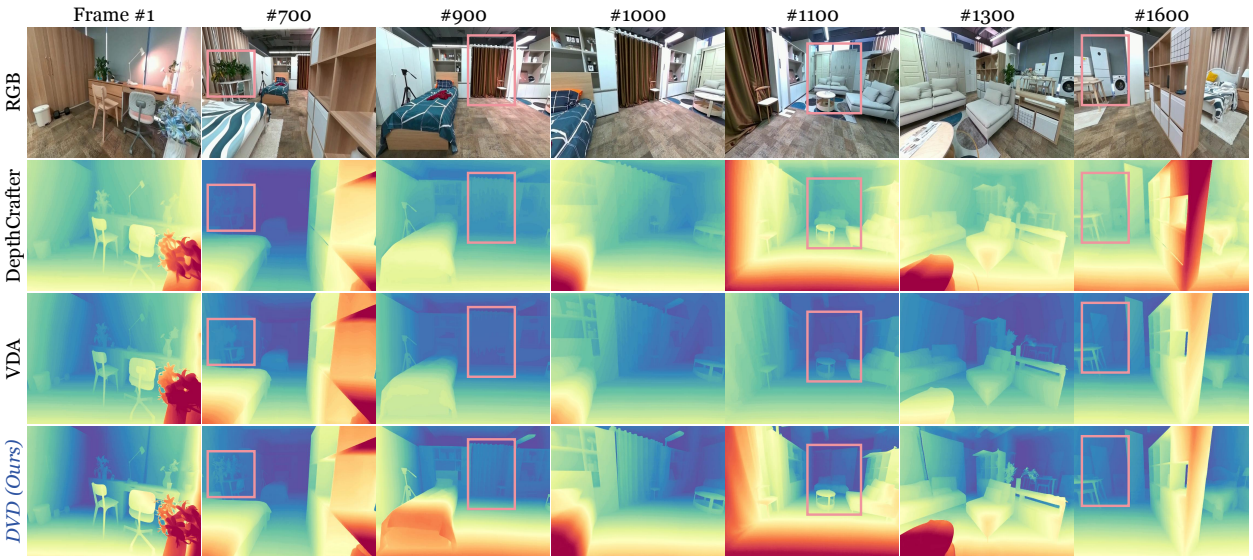

Figure 8. **Qualitative results on long-horizon indoor navigation.** DVD leverages global affine coherence to better preserve sharp boundaries and globally coherent geometry across thousands of frames compared to prior methods.

et al., 2019), and NYUv2 (Nathan Silberman & Fergus, 2012). We report standard metrics including absolute relative error (AbsRel) and threshold accuracy (δ_1), as well as boundary metrics like boundary F1-score (B-F1) and boundary Recall (B-Recall) (Bochkovskii et al., 2024). For video benchmarks, metrics are aggregated over entire sequences instead of independently evaluating sampled frames, allowing temporal inconsistencies and accumulated geometric drift to affect the final scores directly.

Baselines. We compare DVD with representative state-of-the-art video depth estimation methods from two paradigms: ❶ **Diffusion-based generative:** ChronoDepth (Shao et al., 2025) and DepthCrafter (Hu et al., 2025); ❷ **ViT-based discriminative:** Video Depth Anything (VDA) (Chen et al., 2025b); along with image-based Depth Anything V2 (DAv2) (Yang et al., 2024c), Marigold (Ke et al., 2025c), and RollingDepth (Ke et al., 2025a) for reference. Following standard protocols (Ke et al., 2025c; Yang et al., 2024c; Chen et al., 2025b; Hu et al., 2025; He et al., 2024; Birkl et al., 2023; Yang et al., 2024b; Ke et al., 2024; Yang et al., 2024a), "zero-shot" denotes direct evaluation on target benchmarks without any domain-specific fine-tuning.

4.2. Main Results

This section provides empirical evidence of DVD's effectiveness. We evaluate DVD across short/long video depth benchmarks (Tables 1, 2), fine-grained boundary metrics (Table 3), and single-image generalization (Table 4). Supported by qualitative (Figs. 1, 7, 8) and efficiency analyses (Fig. 12), our key observations are summarized as follows:

Obs.❶ DVD achieves overall superior geometric fidelity. Across standard benchmarks (Table 1), DVD matches or outperforms leading generative (*e.g.*, DepthCrafter) and discriminative (*e.g.*, VDA) baselines, achieving the lowest AbsRel on ScanNet (5.5) and KITTI (6.7). This advantage extends to long-video scenarios (Table 2), yielding a substantial margin over DepthCrafter on Bonn (5.3 *vs.* 8.5 AbsRel). Beyond global metrics, DVD better preserves fine-grained details (Table 3, Fig. 7), boosting the ScanNet B-F1 score to 0.259 (*vs.* 0.210 for VDA). Importantly, it also retains competitive single-image generalization (Table 4).

Obs.❷ DVD exhibits compelling data efficiency and comparable inference speed. By inheriting priors from video diffusion models, DVD requires significantly fewer task-specific annotations. Trained on just 367K frames, it surpasses VDA (60M frames) in zero-shot performance (Fig. 12 (*Left*)). Moreover, by avoiding the overhead of iterative generative sampling, DVD achieves inference speeds comparable to efficient discriminative models while delivering superior accuracy (Fig. 12 (*Middle*)).

Obs.❸ DVD exhibits robust scalability to long videos. As visualized in both in-the-wild (Fig. 1) and complex indoor sequences (Fig. 8), generative baselines suffer from scale drift, whereas discriminative ones are susceptible to semantic ambiguities. In contrast, DVD maintains consistent global scale across temporal windows. This structural stability is

quantitatively validated in Fig. 12 (*Right*): as sequence length increases, baseline performance degrades noticeably, while DVD remains highly stable. More examples and failure cases are in Appendix §F.

Due to space limitations, we provide framework analysis and ablation studies in the appendix.

5. Conclusion

In this work, we present DVD, the first framework to deterministically adapt pre-trained video diffusion priors into a single-pass depth regressor. By eliminating stochastic sampling, DVD successfully mitigates the ambiguity-hallucination trade-off, uniting the semantic richness of generative models with the structural stability of discriminative regressors. Our three core designs: (***i***) a timestep-driven structural anchor, (***ii***) latent manifold rectification (LMR) against spatio-temporal mean collapse, and (***iii***) global affine coherence for seamless long-video inference, collectively establish a robust zero-shot solution. Crucially, by effectively grounding these generative priors, DVD achieves state-of-the-art geometric fidelity and temporal coherence while requiring minimal task-specific downstream data. This establishes a highly scalable and efficient paradigm for dynamic 3D scene understanding.

Impact Statement

As a fundamental advancement in dynamic 3D scene understanding, our DVD framework carries broader societal implications. On the positive side, by significantly lowering the downstream data barrier, our highly efficient adaptation paradigm democratizes access to high-fidelity video depth estimation. This facilitates progress in beneficial applications such as autonomous driving, robotic navigation, and augmented reality, particularly for researchers and practitioners in resource-constrained or data-scarce domains.

However, the capability to robustly extract precise geometric structures from unconstrained RGB videos presents potential dual-use risks. It could inadvertently enhance unauthorized 3D surveillance systems or be exploited to generate more physically convincing manipulated media (*e.g.*, 3D-aware deepfakes). Furthermore, because DVD inherently grounds the generic priors of pre-trained video foundation models, it may inherit the dataset biases present in their massive, uncurated pre-training corpora. This could lead to disparate geometric accuracy across different geographic locations, cultural environments, or demographic characteristics. We encourage the community to actively audit these foundational geometric priors for fairness and to establish rigorous safety guidelines prior to real-world deployment in sensitive downstream applications.

References

Bhat, S. F., Alhashim, I., and Wonka, P. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., and Müller, M. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Birkl, R., Wofk, D., and Müller, M. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

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*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S. R., and Koltun, V. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012.

Charatan, D., Li, S. L., Tagliasacchi, A., and Sitzmann, V. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., and 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, 2025b.

Chen, Y., Wang, Y., Zhao, W., Shen, G., Deng, T., and Wang, J. Guided diffusion-based generation of adversarial objects for real-world monocular depth estimation attacks. *arXiv preprint arXiv:2512.24111*, 2025c.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.

Chou, G., Xian, W., Yang, G., Abdelfattah, M., Hariharan, B., Snavely, N., Yu, N., and Debevec, P. Flashdepth: Real-time streaming video depth estimation at 2k resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9638–9648, 2025.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., and Long, X. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2016.

Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

Hang, T., Gu, S., Bao, J., Wei, F., Chen, D., Geng, X., and Guo, B. Improved noise schedule for diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4796–4806, 2025.

He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., and Chen, Y.-C. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

He, J., Li, H., Sheng, M., and Chen, Y.-C. Lotus-2: Advancing geometric dense prediction with powerful image generative model. *arXiv preprint arXiv:2512.01030*, 2025.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., and Shen, S. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., and 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.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., and Schindler, K. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ke, B., Narnhofer, D., Huang, S., Ke, L., Peters, T., Fragkiadaki, K., Obukhov, A., and Schindler, K. Video depth without video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Ke, B., Narnhofer, D., Huang, S., Ke, L., Peters, T., Fragkiadaki, K., Obukhov, A., and Schindler, K. Video depth without video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7233–7243, 2025b.

Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., and Schindler, K. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025c.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Kong, H., Yang, X., Zheng, X., and Wang, X. Worldwarp: Propagating 3d geometry with asynchronous video diffusion. *arXiv preprint arXiv:2512.19678*, 2025.

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.

Lee, H.-Y., Tseng, H.-Y., and Yang, M.-H. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7861–7871, 2024.

Li, H., Wang, C., Lei, J., Daniilidis, K., and Liu, L. Stereodiff: Stereo-diffusion synergy for video depth estimation. *arXiv preprint arXiv:2506.20756*, 2025.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ma, C., Obuchi, T., and Tanaka, T. Neural collapse in cumulative link models for ordinal regression: An analysis with unconstrained feature model. *arXiv preprint arXiv:2506.05801*, 2025.

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Palazzolo, E., Behley, J., Lottes, P., Giguère, P., and Stachniss, C. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. 2019. URL https://www.ipb.uni-bonn.de/pdfs/palazzolo2019iros.pdf.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Pham, D.-H., Do, T., Nguyen, P., Hua, B.-S., Nguyen, K., and Nguyen, R. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17060–17069, 2025.

Piccinelli, L., Yang, Y.-H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., and Yu, F. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Piccinelli, L., Wandel, T., Sakaridis, C., Abbeloos, W., and Van Gool, L. Video depth propagation. *arXiv preprint arXiv:2512.10725*, 2025.

Ranftl, R., Bochkovskiy, A., and Koltun, V. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., and 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.

Sobko, I., Riemenschneider, H., Gross, M., and Schroers, C. Stabledpt: Temporal stable monocular video depth estimation. *arXiv preprint arXiv:2601.02793*, 2026.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Z., Wang, Z., Li, B., Zhang, H., Zhu, R., Liu, L., Jiang, P.-T., and Zhang, T. Depthmaster: Taming diffusion models for monocular depth estimation. *arXiv preprint arXiv:2501.02576*, 2025.

Team, G. Mochi 1. https://github.com/genmoai/models, 2024.

Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F. Z., Daniele, A. F., Mostajabi, M., Basart, S., Walter, M. R., and Shakhnarovich, G. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. URL http://arxiv.org/abs/1908.00463.

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.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., and Yang, J. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5271, 2025.

Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., and Scherer, S. Tartanair: A dataset to push the limits of visual slam. 2020.

Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., and Shen, C. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.

Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., and Pollefeys, M. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025a.

Xu, S., Wei, S., Wei, Q., Geng, Z., Li, H., Shen, L., Sun, Q., Han, S., Ma, B., Li, B., et al. Diffusion knows transparency: Repurposing video diffusion for transparent object depth and normal estimation. *arXiv preprint arXiv:2512.23705*, 2025b.

Xu, T.-X., Gao, X., Hu, W., Li, X., Zhang, S.-H., and Shan, Y. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6632–6644, 2025c.

Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., and He, T. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024b.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., and Shen, C. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9043–9053, 2023.

Zhang, Z., Yang, L., Yang, T., Yu, C., Guo, X., Lao, Y., and Zhao, H. Stabledepth: Scene-consistent and scale-invariant monocular depth. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7069–7078, 2025.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A. Limitations and Future Work

While DVD establishes a robust paradigm for video depth estimation, we acknowledge several limitations that highlight promising avenues for future research.

Boundary Conditions of Long Videos. In unconstrained long videos, extreme dynamics, such as prolonged occlusions, rapid illumination shifts, or erratic camera motions, can introduce local non-linear distortions that temporarily overpower our global affine assumption, leading to scale inconsistencies. Our current evidence for global affine coherence is primarily empirical: although it holds consistently across the evaluated datasets and enables simple long-video inference, it may not characterize all forms of inter-window discrepancy, particularly when the overlap is very small or the scene content changes drastically. To provide a more comprehensive understanding, we further examine this assumption from multiple perspectives in the appendix: Appendix §C analyzes the residual distribution after affine alignment, Appendix §E studies the effect of different overlap sizes, and Appendix §F discusses challenging cases with large motion and abrupt scene transitions where the global affine coherence assumption fails. A more principled characterization of when this approximation holds remains an important direction for future work. Future extensions could explore larger temporal context windows or non-linear latent tracking to better handle these edge cases.

Constraints on Edge Devices and Real-Time Deployment. Although avoiding stochastic sampling significantly accelerates inference, DVD still relies on a large video DiT backbone (*e.g.*, Wan2.1-1.3B (Wan et al., 2025)). Achieving true real-time inference (*e.g.*, $\geq$ 10 Hz) for latency-critical on-device applications remains challenging. Moreover, while DVD uses substantially fewer task-specific training frames, its data efficiency should be interpreted primarily as downstream annotation efficiency rather than overall pre-training, since it builds upon a large pre-trained video foundation model. Promising future directions include architectural distillation and efficient linear-complexity sequence models to transfer these generative priors into lightweight networks.

B. Related Work

Monocular Depth Estimation. Modern approaches have evolved from handcrafted features to data-driven deep learning (Bhat et al., 2021; Eigen et al., 2014), broadly categorized into two dominant paradigms: **(I) Discriminative Regression:** This paradigm leverages ViTs and large-scale supervision to learn direct depth mappings (Hu et al., 2024; Wang et al., 2025; Birkl et al., 2023; Sobko et al., 2026; Piccinelli et al., 2025; Zhang et al., 2025). Foundation models like Depth Anything V1/V2 (Yang et al., 2024b;c; Pham et al., 2025) demonstrate robust zero-shot generalization by scaling up unlabeled pre-training. To recover metric scale, approaches such as Metric3D (Yin et al., 2023), UniDepth (Piccinelli et al., 2024), and Depth Pro (Bochkovskii et al., 2024) focus on resolving focal length ambiguities and preserving high-frequency details. In the video domain, methods like Video Depth Anything (Chen et al., 2025b) extend these backbones with temporal modules or flow-based refinement. While efficient and deterministic, discriminative methods typically lack the generative priors necessary to resolve semantic ambiguities in textureless or motion-blurred regions. **(II) Generative Diffusion:** To incorporate rich geometric priors, recent works repurpose pre-trained diffusion models for depth estimation (Song et al., 2025; Kong et al., 2025; Li et al., 2025; Ranftl et al., 2021; Bhat et al., 2023; Yang et al., 2024a; Xu et al., 2025c). Image-based methods, such as Marigold (Ke et al., 2025c) and Lotus (He et al., 2024; 2025), fine-tune latent diffusion models to achieve superior structural detail compared to discriminative baselines. Video-specific approaches, including ChronoDepth (Shao et al., 2025), DepthCrafter (Hu et al., 2025), and RollingDepth (Ke et al., 2025b), further adapt these priors to model temporal dynamics. However, their reliance on stochastic multi-step sampling inherently introduces high latency and geometric hallucinations, a bottleneck that DVD mitigates via deterministic adaptation.

Video Diffusion Models. The field has witnessed a paradigm shift from adapting 2D U-Nets (Ronneberger et al., 2015) to scalable diffusion transformers (DiT) (Peebles & Xie, 2023). Early pioneering works (Blattmann et al., 2023a;b; Ho et al., 2022; Guo et al., 2023; Wang et al., 2023) primarily extended pre-trained image architectures by inserting temporal attention or 3D convolutions. Recently, a paradigm shift has been driven by spacetime patchified sequence modeling (Brooks et al., 2024) and continuous flow matching (Lipman et al., 2022). By scaling DiT with advanced 3D VAEs, modern foundation models (Team, 2024; Seedance et al., 2025; Gao et al., 2025; Chen et al., 2025a), such as CogVideoX (Hong et al., 2022), HunyuanVideo (Kong et al., 2024), and Wan (Wan et al., 2025), have demonstrated unprecedented capabilities in simulating physical dynamics and maintaining strict 3D consistency. These

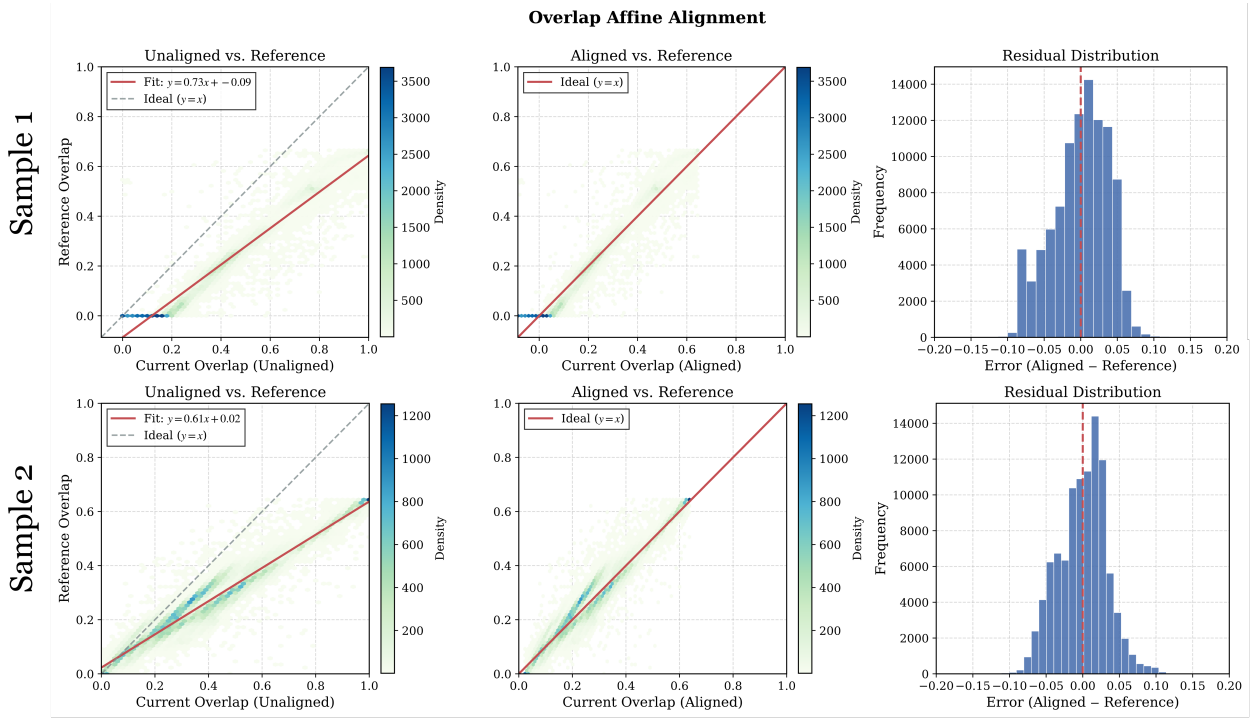

Figure 9. Visualization of inter-window affine alignment. (***Left***) Joint pixel density distribution between the current and reference windows in the overlapping region. The strictly linear correlation (red line) confirms that the inter-window discrepancy is predominantly affine. (***Middle***) After applying our calculated affine transformation, the predictions tightly cluster along the ideal $y = x$ diagonal. (***Right***) The histogram of pixel-wise residuals (Aligned minus Reference) exhibits a zero-mean, tightly bounded distribution with minimal variance, demonstrating the high precision of our global affine coherence strategy.

foundation models effectively function as world simulators, encoding rich geometric and dynamic priors that DVD repurposes for deterministic depth regression.

More Video Depth Methods. Beyond the generative and discriminative paradigms for relative video depth evaluated in our main text, recent advancements have diversified video depth estimation into several specialized tracks. One prominent direction optimizes for **(I) real-time streaming efficiency**, with methods like FlashDepth (Chou et al., 2025) and VeloDepth (Piccinelli et al., 2025) employing lightweight architectures for latency-critical applications. Other parallel tracks include **(II) metric geometry recovery**, where GeometryCrafter (Xu et al., 2025c) alters the target representation to unbounded point maps to facilitate downstream 3D/4D reconstruction. While these works make significant strides in their respective settings, their primary objectives diverge fundamentally from our problem formulation, where DVD explores a more general direction for video depth. Consequently, these works also serve as valuable complementary approaches to the field, rather than direct baselines for our core setting.

C. More Details of Global Affine Coherence

To further validate the core assumption underlying our global affine coherence, we visualize the pixel-wise depth relationship within overlapping temporal windows. As shown in Fig. 9 (*Left*), the joint density distribution of unaligned predictions exhibits an approximately linear trend. This observation supports our hypothesis that inter-window discrepancies are mainly dominated by global scale and shift factors, rather than complex non-linear distortions. After applying the estimated affine transformation, the predictions closely align with the ideal $y = x$ diagonal (*Middle*). Moreover, the post-alignment residual distribution (*Right*) is approximately zero-centered with small variance. These results suggest that our lightweight linear alignment strategy can effectively stitch long sequences while introducing limited geometric error.

D. More Implementation Details

Training Details. To facilitate reproducibility, we detail the architectural configurations and training hyperparameters of DVD in this section. To adapt this model for deterministic regression while strictly preserving its pre-trained world priors, we freeze the original weights and employ Low-Rank Adaptation (LoRA) exclusively on the attention blocks. The detailed settings for the VAE compression, LoRA configuration, optimization schedule, and joint-training loss weights are systematically summarized in Table 5.

Table 5. Hyperparameter configurations for DVD.

Setting	Hyperparameter	Value
	Model Architecture & Adaptation	
Backbone	pre-trained model	Wan2.1-1.3B
VAE	spatial compression	$8\times$
	temporal compression	$4\times$
LoRA	target modules	$W_q, W_k, W_v, W_o, W_{ffn}$
	rank (r) / alpha (α)	$512/512$
	Training & Inference Settings	
Optimization	optimizer	AdamW
	base learning rate	1×10^{-4}
	LR schedule	Constant
	hardware	$8\times$ NVIDIA H100 GPUs
Data	global batch size (video)	16
	global batch size (image)	128
	spatial resolution	480×640
	window size	45
	stride (inference)	9
	Objective Weights	
Loss Weights	spatial rectification (λ_{sp})	0.5
	temporal rectification (λ_{temp})	0.5
	image joint loss (λ_{image})	1.0

Experimental Details. We provide more details here regarding our fine-grained evaluation protocols and inference efficiency benchmarking to ensure fair comparison and reproducibility. ❶ **Video Length Partitioning.** Aggregating metrics across an entire dataset often masks the severe scale drift generative models suffer on extended sequences. To rigorously assess temporal stability, we partition the test sets by duration: *short videos* (50–200 frames) and *long videos* ($>$ 200 frames). This explicit decoupling effectively isolates short-term geometric fidelity from long-term structural persistence. ❷ **Inference Efficiency Benchmarking.** Latency and FPS are evaluated on a single NVIDIA RTX A6000 GPU under identical environments. To reflect practical deployment, we implement two streamlined optimizations: (***i***) merging LoRA weights into the base backbone to eliminate modular overhead, and (***ii***) using non-tiled VAE decoding to bypass spatial slicing bottlenecks. While our single-pass paradigm is inherently fast, integrating advanced accelerators (*e.g.*, TensorRT) remains a promising future direction for real-time edge applications.

E. More Analysis

In this section, we analyze the core design choices of DVD. Unless otherwise specified, all ablation experiments are conducted on ScanNet (Dai et al., 2017). More studies are provided in Appendix §E.

Deterministic Adaptation *vs.* Generative Sampling. To isolate our methodological gains and ensure fair comparisons with generative baselines, we evaluate our single-pass regression against standard multi-step stochastic sampling using the identical backbone (Fig. 11 (*Middle*)). Deterministic adaptation significantly improves geometric accuracy, dropping AbsRel from 9.7 ($T = 10$) to 7.3 ($T = 1$). This controlled setting supports our hypothesis that, in this setting, iterative stochastic sampling can introduce variance that affects structural consistency, whereas our deterministic mapping directly predicts the geometric structure, ensuring superior stability.

Table 6. Cross-backbone generalization on KITTI. We deploy `DVD` on CogVideoX-5B (Hong et al., 2022). Consistent with our default Wan2.1-1.3B backbone, the mid-range timestep ($\tau = 0.5$) serves as the optimal structural anchor.

Metric	Wan2.1-1.3B	CogVideoX-5B (by τ)				
	`DVD`	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
AbsRel ↓	6.7	7.8	7.6	7.4	7.4	9.5
δ ↑	0.967	0.931	0.934	0.938	0.935	0.898

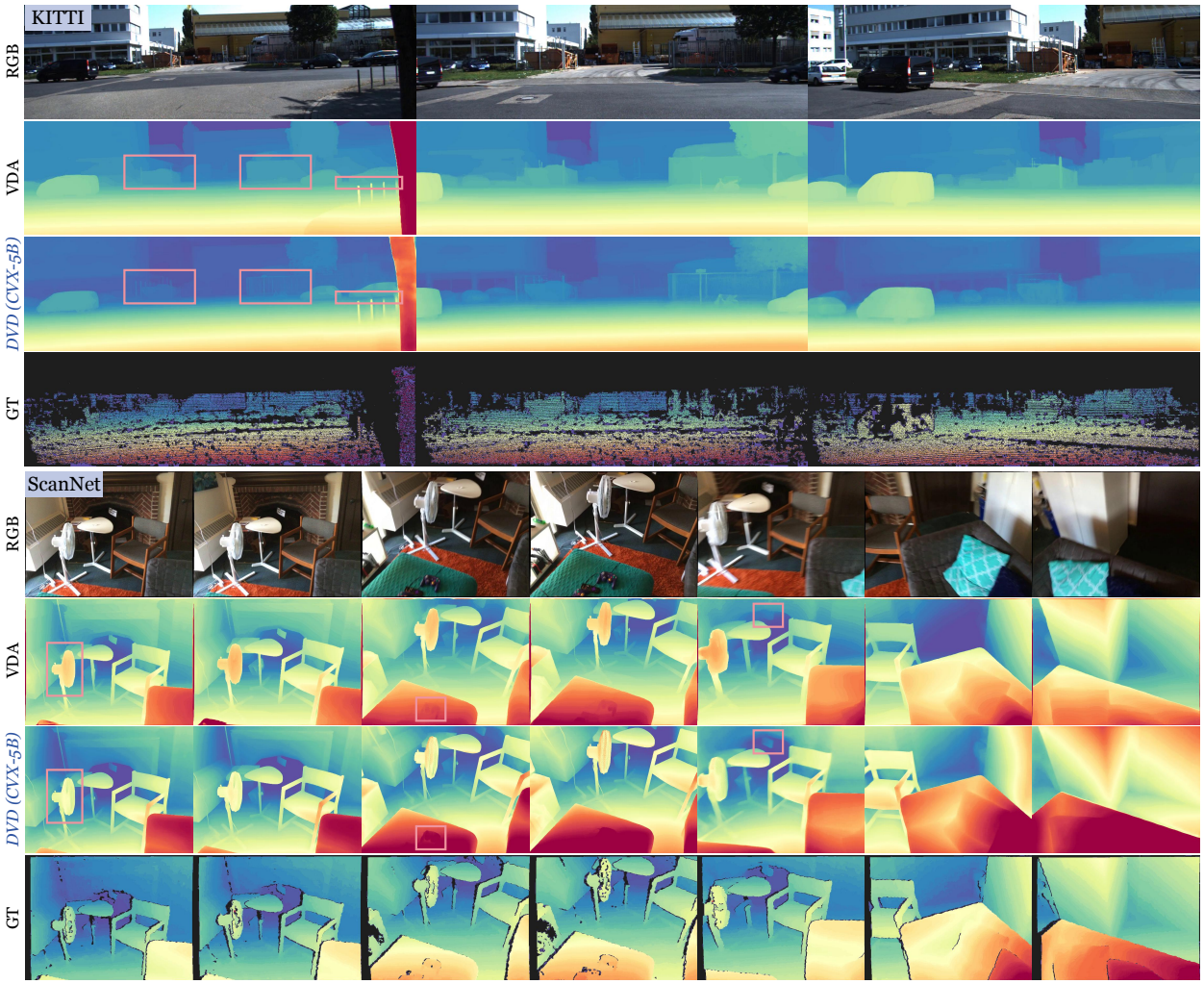

Figure 10. Qualitative results of `DVD` on CogVideoX (Hong et al., 2022). Despite employing a different foundation architecture, our deterministic paradigm still preserves significantly sharper high-frequency details (highlighted in red boxes) compared to the leading baseline.

Role of Timestep Conditioning. We further investigate the impact of the structural anchor τ. Fig. 13 illustrates the impact of the structural anchor τ, revealing a clear fidelity-stability trade-off where $\tau = 0.5$ is optimal. Outdoor scenes (KITTI) are significantly more sensitive to this anchor, with AbsRel fluctuating drastically from 13.8 ($\tau = 0.0$) down to 6.7 ($\tau = 0.5$). Conversely, extreme high values ($\tau \geq 0.9$) trigger a severe performance decrease across datasets, as low-frequency bias completely washes out essential details. This confirms τ effectively dictates the backbone's pre-trained geometric operating regime.

Impact of Latent Manifold Rectification. Evaluating our LMR module (Fig. 11 (*Left*)) shows that adding spatial ($\mathcal{L}_{\text{sp}}$) and temporal ($\mathcal{L}_{\text{temp}}$) differential constraints progressively improves global accuracy (AbsRel drops from 8.5 to 7.3) and boundary precision (B-F1 rises from 0.210 to 0.259). This demonstrates that explicitly enforcing differential

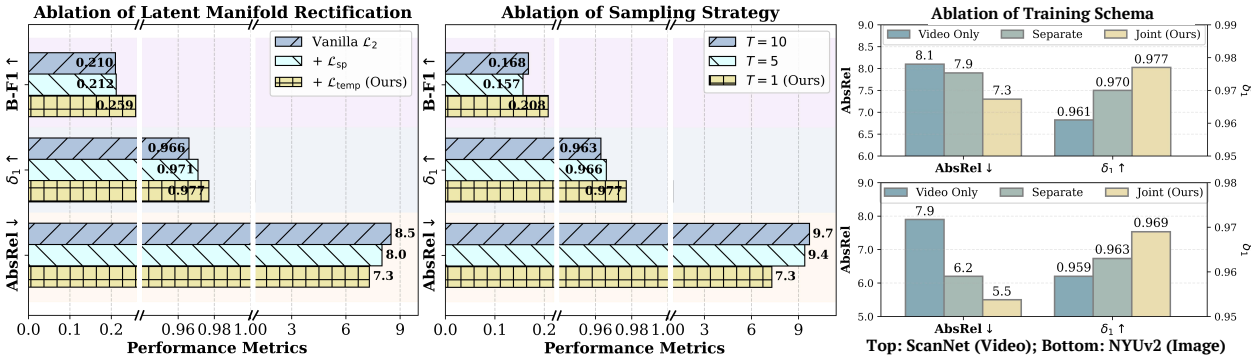

Figure 11. (***Left***) Ablation of latent manifold rectification. (***Middle***) Ablation of the sampling strategy. (***Right***) Ablation of training scheme. The results demonstrate that latent manifold rectification, deterministic adaptation, and joint image-video training effectively enhance geometric accuracy.

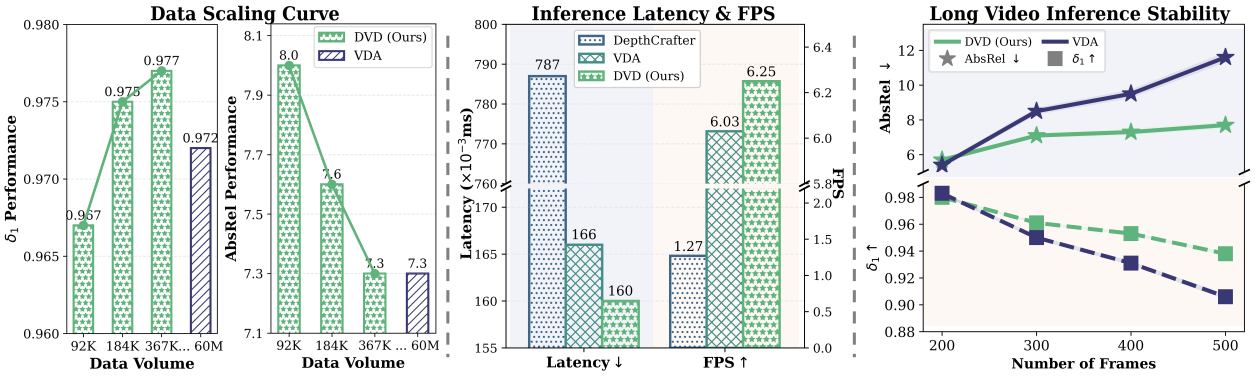

Figure 12. (***Left***) Data scaling curve of long video performance on ScanNet. (***Middle***) Inference latency & FPS comparisons. (***Right***) Stability of long video inference on KITTI. These results demonstrate DVD's remarkable data efficiency, competitive inference speed, and consistent long-video stability.

constraints successfully mitigates the mean collapse inherent in single-pass regression, restoring sharp structural boundaries and coherent motion.

Effectiveness of Image-Video Joint Training. Fig. 11 (*Right*) further analyzes our training strategy. Training exclusively on videos underfits spatial details, while separate sequential training suffers from catastrophic forgetting on single-frame tasks. Our joint strategy achieves the best performance across both domains, *i.e.*, maximizing ScanNet video accuracy ($\delta_1 = 0.977$) while sharply reducing NYUv2 single-image AbsRel to 5.5. This demonstrates that images and videos offer complementary supervision: images act as high-frequency spatial anchors, while videos enforce temporal consistency.

Cross-Backbone Generalization. Finally, to ensure our deterministic adaptation paradigm is not restricted to a specific architecture, we further apply DVD to the CogVideoX-5B (Hong et al., 2022) backbone. As detailed in Appendix §E, DVD demonstrates consistent fidelity-stability behavior and successfully extracts sharp high-frequency geometries, confirming the broad generalizability of our empirical findings across diverse video foundation models.

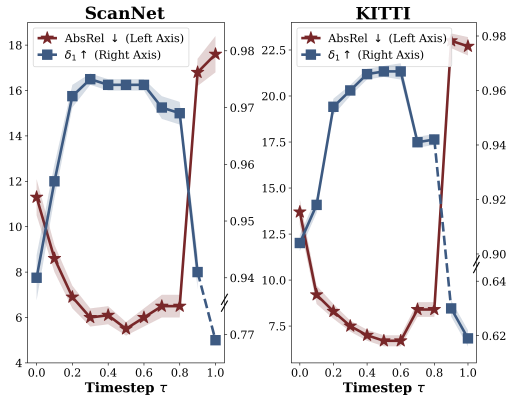

Figure 13. **Ablation of timestep τ.** The structural anchor τ dictates a fidelity-stability trade-off. While outdoor scenes are more sensitive, both datasets achieve a balance at $\tau = 0.5$.

Cross-Backbone Generalization. To verify that our deterministic adaptation paradigm is universally applicable rather than specific to a single architecture, we further deploy `DVD` on CogVideoX-5B (Hong et al., 2022). As shown in Table 6, ablating the structural anchor τ on KITTI perfectly corroborates our findings from the Wan2.1-1.3B backbone: extreme timesteps degrade geometry, whereas the mid-range value ($\tau = 0.5$) provides the optimal conditioning (AbsRel 7.4, δ_1 0.938). While the CogVideoX-5B + `DVD` variant yields expectedly lower metrics than our default Wan2.1 backbone (due to differing foundation capacities and pre-training distributions), its fundamental capability for structural extraction remains profoundly robust. As visualized in Fig. 10, even with a different generative backbone, `DVD` effortlessly recovers fine-grained, high-frequency geometries that are severely over-smoothed by the state-of-the-art discriminative baseline, VDA. This provides initial evidence that the empirical findings can transfer across different video foundation models.

Analysis of Pre-trained Structural Anchors. Table 7 details the exact numerical breakdown of the fidelity-stability trade-off discussed in Fig. 13. As previously established, entirely removing the structural anchor (represented as w/o τ, which is equivalent to the $\tau = 0.0$ boundary state) significantly degrades global metric accuracy. To further investigate whether this conditioning acts merely as a standard trainable parameter or an irreplaceable pre-trained key, we introduce an additional extreme ablation: **learning** τ. In this setting, we replace the fixed sinusoidal frequency basis with a randomly initialized, fully learnable embedding of identical dimensions. As shown in Table 7, learning a new anchor from scratch triggers a catastrophic performance collapse, with AbsRel skyrocketing to 16.3 and 23.7 on ScanNet and KITTI, respectively. We attribute this to the fact that the profound structural priors within the pre-trained video DiT backbone are fundamentally entangled with its original sinusoidal frequency encodings. A newly initialized embedding fails to activate these pre-trained pathways, effectively rendering the zero-shot generative priors inaccessible. This confirms that our fixed τ anchor natively unlocks the foundation model's geometric capacity, and cannot be replaced by naive fine-tuning.

Table 7. Analysis of structural anchor τ, which is evaluated on ScanNet (***Left***) and KITTI (***Right***).

Timestep	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
w/o τ	11.3	0.940	13.7	0.904
$\tau = 0.1$	8.6	0.957	9.2	0.918
$\tau = 0.2$	6.9	0.972	8.3	0.954
$\tau = 0.3$	6.0	0.975	7.5	0.960
$\tau = 0.4$	6.1	0.974	7.0	0.966
$\tau = 0.5$	5.5	0.974	6.7	0.967
$\tau = 0.6$	6.0	0.974	6.7	0.967
$\tau = 0.7$	6.5	0.970	8.4	0.941
$\tau = 0.8$	6.5	0.969	8.4	0.942
$\tau = 0.9$	16.8	0.941	23.0	0.630
$\tau = 1.0$	17.6	0.769	22.7	0.619
learning τ	16.3	0.811	23.7	0.699

Analysis of Different Regularization Strategies. To isolate Latent Manifold Rectification (LMR)'s efficacy against mean collapse, we compare it with widely adopted regularizers in Table 8. The vanilla $\mathcal{L}_2$ baseline yields sub-optimal accuracy (AbsRel 8.5) and structural fidelity (B-F1 0.210). Adding RGB reconstruction distracts the network from decoding textures rather than geometry. Interestingly, existing geometric regularizers present a strict trade-off: edge-aware smoothness improves global metrics (AbsRel 7.5, δ_1 0.978) but severely over-smooths high-frequency details (B-F1 drops to 0.193). Conversely, multi-scale gradient matching sharpens boundaries (B-F1 0.257) but offers marginal global scale correction (AbsRel 8.2). In stark contrast, LMR breaks this dilemma. By enforcing latent differential constraints, it simultaneously minimizes AbsRel (7.3) and maximizes boundary precision (B-F1 0.259), suggesting that LMR provides a better balance between global accuracy and boundary preservation under our experimental setting.

Table 8. Analysis of regularization strategies. All variants use the same backbone, training, and inference settings.

Regularizer	AbsRel↓	δ_1 ↑	B-F1↑
$\mathcal{L}_2$ only	8.5	0.966	0.210
+ RGB reconstruction	10.5	0.951	0.174
+ Edge-aware smoothness	<u>7.5</u>	**0.978**	0.193
+ Multi-scale gradient matching	8.2	0.969	<u>0.257</u>
+ LMR (Ours)	**7.3**	<u>0.977</u>	**0.259**

Analysis of Overlap Size. To determine the optimal balance between temporal consistency and computational efficiency, we further ablate the sliding window overlap size (O) on KITTI (Geiger et al., 2012) here. As detailed in Table 9, an extremely small overlap ($O = 3$) yields sub-optimal accuracy (AbsRel 7.9), as the limited pixel population makes the affine estimation susceptible to local dynamic outliers. Expanding the overlap region enriches the statistical basis for our Global Affine Coherence, effectively reducing inter-window discrepancies (*e.g.*, AbsRel drops to 7.3 at $O = 9$). However, further enlarging the overlap ($O \geq 14$) yields

Table 9. Analysis of overlap size on KITTI. Increasing the number of overlapping frames (O) improves geometric accuracy but incurs diminishing returns and computational overhead (Rel. Time).

Overlap Size	AbsRel↓	δ_1 ↑	Rel. Time↓
$O = 3$	7.9	0.937	1.00×
$O = 6$	7.7	0.941	1.04×
$O = 9$	7.3	0.945	1.17×
$O = 14$	7.2	0.948	1.34×
$O = 19$	7.1	0.947	1.55×

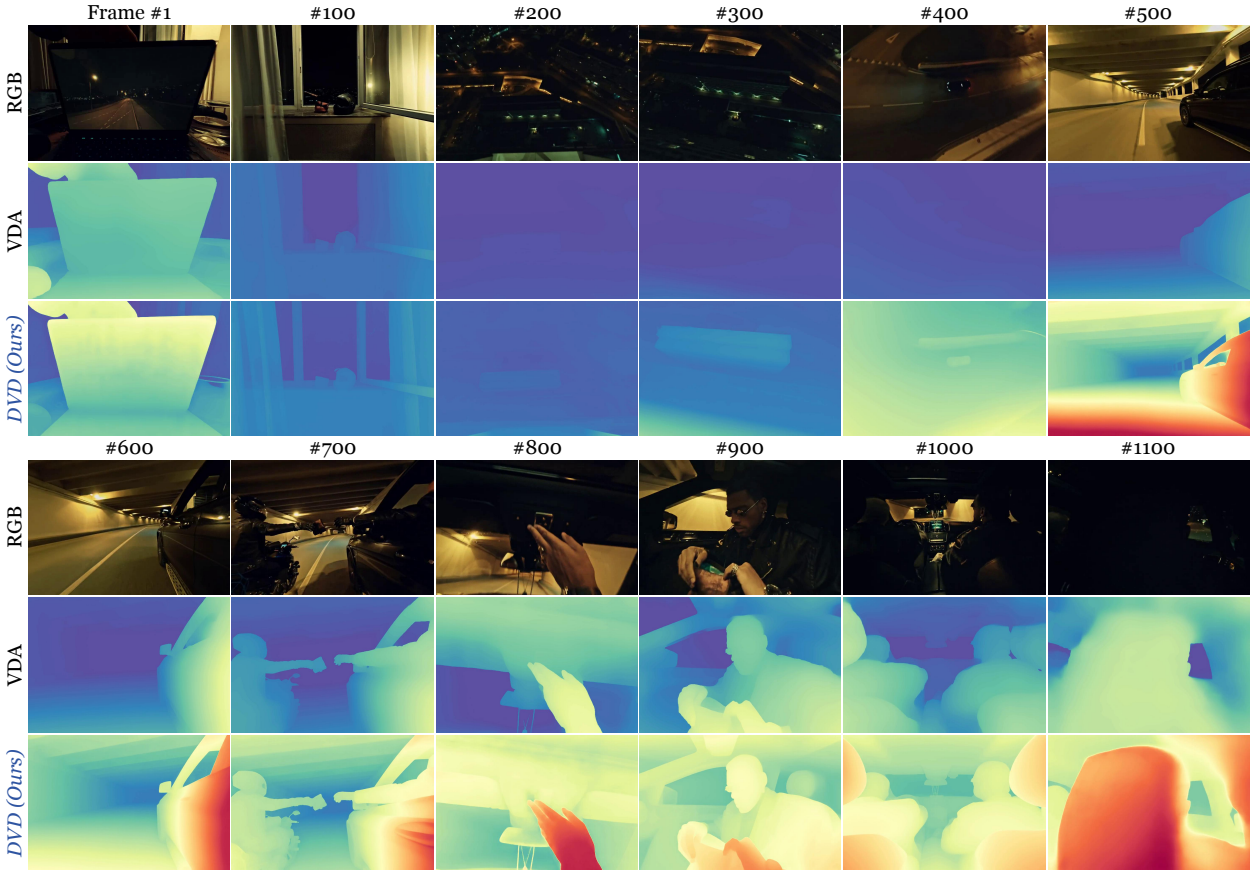

Figure 14. Failure case analysis on 1100-frame sequence with massive scene transitions. While both DVD and VDA (Chen et al., 2025b) inevitably suffer from global scale drift across disjointed scenes (*e.g.*, contrasting the absolute depth representations between Frame #1 and #500), DVD consistently maintains significantly sharper local structural fidelity (*e.g.*, hands in Frames #800–#900).

diminishing geometric returns, and even saturates structural fidelity (δ_1 slightly drops at $O = 19$), while incurring severe computational overhead (reaching $1.55\times$ relative latency). Consequently, a moderate overlap configuration provides a highly robust, jitter-free geometric transition without unnecessarily sacrificing inference efficiency.

Analysis of LoRA Rank. Unlike standard LoRA applications that employ low ranks for superficial style transfer, adapting a video diffusion backbone into a dense geometric regressor requires modeling highly complex mappings. Table 10 ablates the LoRA capacity on ScanNet. We observe that a moderate rank of 256 yields sub-optimal accuracy, while expanding to rank 512 significantly improves structural fidelity (AbsRel drops to 7.3). Further scaling to rank 1024 provides only marginal gains. Empirically, we found that extremely low ranks struggle to capture high-frequency details, whereas full parameter fine-tuning tends to overfit the limited training data and degrade the model's pre-trained zero-shot priors. Consequently, rank 512 offers an optimal balance between geometric capacity and prior preservation.

Table 10. Analysis of LoRA ranks.

LoRA Rank	AbsRel↓	δ_1 ↑
256	7.7	0.974
512	7.3	0.977
1024	7.3	0.979

F. Failure Case Analysis

As discussed in Section §A, the empirical affine coherence of DVD relies on geometric overlap between adjacent temporal windows. Consequently, the model may experience scale inconsistencies in unconstrained long-video scenarios characterized by massive ego-motion or abrupt scene transitions.

Fig. 14 illustrates a highly challenging 1100-frame sequence featuring drastic environmental shifts (*e.g.*, transitioning

from an indoor desk to an outdoor tunnel). When comparing distant frames with zero visual overlap (*e.g.*, Frame #1 *vs.* Frame #500), `DVD` inevitably exhibits global scale drift, struggling to anchor a unified absolute depth range across completely disjointed scenes.

This case highlights a boundary condition of our affine alignment strategy. When adjacent windows share little or no visual overlap due to abrupt scene transitions or large ego-motion, the estimated affine transformation lacks a reliable geometric anchor, and global scale consistency may degrade. We observe that this issue is not unique to DVD: the strong discriminative baseline VDA (Chen et al., 2025b) also exhibits temporal scale degradation under the same extreme dynamic conditions. Nevertheless, DVD still preserves sharper local structures in many regions, such as the laptop screen in Frame #1 and hand geometries in Frames #800–#900, whereas VDA tends to produce smoother predictions. These results suggest that, although infinite-length metric anchoring remains challenging under disjoint scene content, deterministic adaptation of generative video priors can still provide strong local geometric fidelity.