FE2E: From Editor to Dense Geometry Estimator

Supplementary Material

In this appendix, we provide more implementation details, experiments, analysis, and discussions for a comprehensive evaluation and understanding of FE2E. Detailed contents are listed as follows:

A. Experiment Settings

A.1. Auxiliary Dispersion Loss

Following Diffuse-and-Disperse [47], we apply this loss to the output of the 9th block:

$$\mathcal{L}_{\text{disp}} = \log \mathbb{E}_{i,j} \left[\exp(-\|\eta_i - \eta_j\|_2^2 / \tau) \right], \tag{1}$$

where $\eta_{i,j}$ are the output features for the i-th and j-th samples in a batch, respectively, and temperature $\tau = 1$. Finally, the training loss is defined as: $\mathcal{L}_{train} = \mathcal{L}_{fm} + \lambda \mathcal{L}_{\text{disp}}, \lambda = 0.5$. The choices of λ, τ, and block all follow the optimal hyperparameters identified in the experiments from Diffuse-and-Disperse.

Table 1. Ablation study on Disperse Loss. The baseline is the ID4 model in the main paper, Table 4.

Method	KITTI		ETH3D	
	AbsRel↓	$\delta 1$ ↑	AbsRel↓	$\delta 1$ ↑
Baseline (CV + FS)	8.6	94.0	4.8	97.3
+ Disperse Loss (DL)	8.4	94.4	4.5	97.6

For integrity, we also conducted ablation studies on this dispersed loss. The performance gains observed in Table 1 confirm that this loss is also effective for dense geometric estimation tasks.

A.2. Evaluation Datasets

We evaluate our model on two tasks: **Zero-shot Affine-Invariant Depth Estimation.** We evaluate on five standard benchmarks: NYUv2 [43], ScanNet [8], KITTI [9], ETH3D [39], and DIODE [45]. Following standard practice, we report the Absolute Relative error (AbsRel) and δ_1 accuracy. **Surface Normal Prediction.** We evaluate on NYUv2, ScanNet, iBims-1 [20], and Sintel [2] benchmarks. The evaluation metrics are the mean angular error (MeanErr) and the percentage of pixels with an angular error below $11.25°$.

A.3. Evaluation Metrics

For zero-shot depth estimation, similar to [18], we employ the following evaluation metrics:

- AbsRel: $\frac{1}{|M_{vl}|} \sum_{d \in M_{vl}} |d - d_{gt}| / d_{gt}$;
- a_1: percentage of d such that $\max(\frac{d}{d_{gt}}, \frac{d_{gt}}{d}) < 1.25$;

where d_{gt} and d denote the GT and estimated pixel depth, M_{vl} is the valid mask (mask rules are consistent with [11]).

For zero-shot normal estimation, we use the following evaluation metrics:

- MeanErr $= \frac{1}{|M_{vl}|} \sum_{\mathbf{n} \in M_{vl}} \frac{180}{\pi} \arccos(\text{clamp}(\mathbf{n} \cdot \mathbf{n}_{gt}, -1, 1))$;
- $11.25°$: The percentage of $\mathbf{n}$ where the angular error is less than $11.25°$;

where $\mathbf{n}_{gt}$ and $\mathbf{n}$ are GT and estimated normal vector.

B. Training Details of Finetune Analysis

B.1. Improved Experiment Setup

For clarity, we term the direct adaptation of the original editing/generative formulation as *"DirectAdapt"* (Sec 3.2), and Table 4 shows that *DirectAdapt* fails to achieve satisfactory performance. To address this, we introduce two key improvements on training objective (Sec 3.3) and GT quantization(Sec 3.4). They can benefit both editing and generative models, and these *improved* models are better for analyzed our core motivation (Sec 3.1), as they isolate the error from training data and the denoising process. We finally introduce joint training on the editing-based model to obtain **FE2E**.

B.2. Implementation Details of Generative-based Models

Step1X-Edit is fine-tuned from the generative model FLUX, and both share an almost identical DiT architecture. To further reduce confounding factors, we follow the Step1X-Edit protocol and replace the original FLUX input with a horizontally concatenated noise and RGB image. All hyperparameters, including LoRA settings, optimizer, and training data, are kept exactly the same as those used for FE2E in-depth estimation.

FLUX consists of 38 block layers, each producing outputs of consistent dimensions. After rearrangement, the feature map has the shape $B \times 192 \times H/8 \times W/8$, where B is the batch size, H and W are the height and width of the input image. Typically, the output from the final block is projected to 16 channels and passed to the VAE for reconstruction to $B \times 3 \times H \times W$. For visualization, we chose 1, 20, and 35 blocks, operate on the $B \times 192 \times H/8 \times W/8$ feature map, normalize it to $B \times 1 \times H/8 \times W/8$ using the L2 norm, upsample it to $B \times 1 \times H \times W$, and finally visualize it using the Rainbow colormap. The visualization of depth and normals follows the approach of Lotus.

Since our experimental comparisons are conducted using the *improved* model, only one single "denoising" step is performed during inference. Consequently, the output from the VAE decoder directly represents the depth map (the $B \times 3 \times H \times W$ output mentioned before was averaged to obtain a 1-channel depth map), which makes it easier to visualize meaningful features.

C. Quantization Error Calculation Details

The following calculations are based on the effective depth range of 0-80m from the Virtual KITTI dataset. The normalization scheme consistently maps an input domain X to the VAE's mandatory input range of [-1, 1] using the standard min-max scaling formula:

$$V = 2 \times \frac{X - X_{min}}{X_{max} - X_{min}} - 1$$

. While other mapping schemes from [0m, 80m] to [-1, 1] may exist, they are not explored in this work. All calculations use the worst-case precision of BF16 over the [-1, 1] interval, which corresponds to a single quantization step of $\Delta V \approx 1/256$.

C.1. Uniform Quantization

In this scheme, the depth value D is linearly mapped to the [-1, 1] interval. The depth range is $[D_{min}, D_{max}] = [0\text{m}, 80\text{m}]$. The mapping function is $V = 2 \times \frac{D-0}{80-0} - 1 = \frac{D}{40} - 1$. A quantization step of $\Delta V = 1/256$ in the normalized space corresponds to an error ΔD in the real-world depth space. This error is constant across the entire depth range:

$$\Delta D = 40 \times \Delta V = 40 \times \frac{1}{256} \approx 0.15625$$

At 80m: Error $\approx 16cm$. AbsRel = $\frac{0.16\text{m}}{80\text{m}} = 0.002$.
At 0.1m: Error $\approx 16cm$. AbsRel = $\frac{0.16\text{m}}{0.1\text{m}} = 1.600$.
This method yields an unacceptably large relative error at close distances.

C.2. Inverse Quantization

This scheme quantizes the reciprocal of depth, i.e., disparity $P = 1/D$. We consider an effective depth range of $[0.1\text{m}, 80\text{m}]$ to avoid division by zero. The corresponding disparity range is $[P_{min}, P_{max}] = [1/80, 1/0.1] = [0.0125, 10]$. The disparity P is linearly mapped to [-1, 1]. The quantization step in disparity, ΔP, is constant:

$$\Delta P = (P_{max} - P_{min}) \times \frac{\Delta V}{2} \approx 0.0195.$$

The relationship between depth error ΔD and disparity error ΔP is given by $\Delta D \approx |\frac{d(1/P)}{dP}|\Delta P = \frac{1}{P^2}\Delta P = D^2\Delta P$.

At 80m: Error = $(80\text{m})^2 \times 0.0195 = 6400 \times 0.0195 \approx 124.8\text{m} \approx 125m$. AbsRel = $\frac{125\text{m}}{80\text{m}} \approx 1.563$.

At 0.1m: Error = $(0.1\text{m})^2 \times 0.0195 = 0.01 \times 0.0195 = 0.000195\text{m} \approx 0.2mm$. AbsRel = $\frac{0.0002\text{m}}{0.1\text{m}} = 0.002$.

As mentioned in the main text, the disparities for 39m and 78m are $1/39 \approx 0.0256$ and $1/78 \approx 0.0128$, respectively. Their difference is ≈ 0.0128, which is smaller than the disparity quantization step $\Delta P \approx 0.0195$, making them indistinguishable after quantization. This scheme fails completely at large distances.

C.3. Logarithmic Quantization

This scheme quantizes the logarithmic depth, $D_{log} = \ln(D)$. We again consider the depth range $[0.1\text{m}, 80\text{m}]$. The corresponding log-depth range is $[\ln(0.1), \ln(80)] \approx [-2.30, 4.38]$. The log-depth D_{log} is linearly mapped to [-1, 1]. The quantization step in log-depth, ΔD_{log}, is constant:

$$\Delta D_{log} = (\ln(80) - \ln(0.1)) \times \frac{\Delta V}{2} \approx 0.013.$$

The relationship between depth error ΔD and log-depth error ΔD_{log} is given by $\Delta D \approx |\frac{d(e^{D_{log}})}{dD_{log}}|\Delta D_{log} = e^{D_{log}}\Delta D_{log} = D \cdot \Delta D_{log}$. This implies that the absolute relative error, AbsRel = $\Delta D/D$, is approximately constant and equal to $\Delta D_{log} \approx 0.013$.

At 80m: AbsRel ≈ 0.013. Error = $80\text{m} \times 0.013 = 1.04\text{m}$.

At 0.1m: AbsRel ≈ 0.013. Error = $0.1\text{m} \times 0.013 = 0.0013\text{m} = 1.3mm$.

This method maintains a reasonable and nearly constant relative error across both near and far ranges, making it a well-balanced and effective solution. The percentile-based normalization used in the main text is a more robust implementation of this fundamental principle.

D. Preliminaries of Flow Matching

Flow Matching [24] is a highly effective framework for training Continuous Normalizing Flows (CNFs). The core idea is to smoothly transform a simple prior distribution p_0 (e.g., the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$) into a complex target data distribution p_1 over a continuous time variable $t \in [0, 1]$.

This transformation process can be described by an Ordinary Differential Equation (ODE), where the velocity at any time t and point $\mathbf{z}$ is defined by a vector field $v_t(\mathbf{z})$. However, estimating this marginal vector field $v_t(\mathbf{z})$ directly from data samples is challenging. The Flow Matching framework elegantly bypasses this issue by regressing a much simpler and easier-to-compute conditional vector field $u_t(\mathbf{z}|\mathbf{z}_0, \mathbf{z}_1)$ instead.

Specifically, we first sample a pair of points, $(\mathbf{z}_0, \mathbf{z}_1)$, from the prior distribution p_0 and the target distribution p_1, respectively. We then define a simple path $\mathbf{z}_t$ from $\mathbf{z}_0$ to $\mathbf{z}_1$ and its corresponding conditional vector field $u_t = \frac{d\mathbf{z}_t}{dt}$. It has been proven that if a neural network $f_\theta(\mathbf{z}, t)$ is trained to approximate this simple conditional vector field u_t, then in expectation over all sample pairs $(\mathbf{z}_0, \mathbf{z}_1)$ and time t, the network f_θ will converge to the complex marginal vector field v_t that we truly wish to learn.

Rectified Flow [26] presents a particularly simple and powerful instance of Flow Matching. It defines the path between $\mathbf{z}_0$ and $\mathbf{z}_1$ as a straight line:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0, \quad t \in [0, 1].$$

The derivative of this path is trivial, yielding a constant velocity vector that is independent of both time and space:

$$\mathbf{v} = \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0.$$

Consequently, the training objective (loss function) becomes exceedingly simple: aligning the neural network's prediction with this constant velocity vector $\mathbf{v}$:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_1,\mathbf{z}_0} \|(\mathbf{z}_1 - \mathbf{z}_0) - f_\theta(\mathbf{z}_t, t)\|^2.$$

Application in *DirectAdapt* In this paper, we adapt this framework for a conditional image editing task. Our goal is not to learn an unconditional generative model, but rather a flow from noise $\mathbf{z}_0^y$ to the target geometry latent $\mathbf{z}_1^y$, guided by the input image $\mathbf{x}$ (encoded as $\mathbf{z}^x$). Therefore, our velocity prediction model f_θ must take $\mathbf{z}^x$ as an additional condition. As shown in Eq. 2 in the main text, our loss function is:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_1^y,\mathbf{z}_0^y} \|(\mathbf{z}_1^y - \mathbf{z}_0^y) - f_\theta(t, \mathbf{z}^x)\|^2.$$

During inference, we generate the target latent $\hat{\mathbf{z}}_1^y$ by solving the following ODE, with $\mathbf{z}^x$ serving as the guiding condition:

$$\frac{d\hat{\mathbf{z}}_t^y}{dt} = f_\theta(t, \mathbf{z}^x), \quad \text{with initial value } \hat{\mathbf{z}}_0^y \sim \mathcal{N}(0, \mathbf{I}).$$

By integrating from $t = 0$ to $t = 1$ using a numerical ODE solver (e.g., Euler method), we can obtain the final prediction $\hat{\mathbf{z}}_1^y$.

E. Reviews of Related Generative and Editing Models

The fields of image generation and image editing have always been complementary, and they have undergone several paradigm shifts. The first major breakthrough was the Generative Adversarial Network (GAN) [10], which introduced a novel adversarial training process. Then, key advancements in this era include architectural refinements like DCGAN [34], the development of conditional and text-to-image GANs such as the StackGAN series [62, 63], AttnGAN [50], and Cross-Modal Contrastive Learning based models [64]. The StyleGAN series [15–17] marked a high point for GANs, achieving unprecedented photorealistic high-resolution image synthesis and offering fine-grained control over visual attributes through a disentangled latent space, which became a cornerstone for many subsequent editing techniques.

More recently, the field has transitioned to Denoising Diffusion Models [12], which have become the state-of-the-art for their superior image quality and textual coherence. A series of influential diffusion-based methods were introduced, including GLIDE [28], DALL·E [35] and its successor DALL·E 2 [36], Imagen [38], and PIXART-α [3]. The open-source Stable Diffusion (SD) [37] model, trained on the large-scale LAION-5B dataset [40], further democratized high-quality image generation and quickly became a community standard. A growing body of evidence suggests that Diffusion Transformers [6, 7, 21, 30] outperform U-Nets, motivating the shift toward training modern diffusion models with Transformer architectures.

Building on these powerful generative foundations, the domain of image editing [22] (generalized editing) has also advanced rapidly. Early diffusion-based methods like SDEdit [27] demonstrated that real images could be edited by adding noise and then denoising with a new text prompt. A significant leap was made with instruction-guided editing, pioneered by InstructPix2Pix [1], which enabled edits

Table 2. Quantitative comparison on zero-shot affine-invariant depth estimation between FE2E and the concurrent unified model.

Method	NYUv2 (Indoor)		KITTI (Outdoor)		ETH3D (Various)		ScanNet (Indoor)		DIODE (Various)		Avg
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	Rank↓
Qwen-Image	5.5	96.7	7.8	95.1	6.6	96.2	4.7	97.4	19.7	83.2	2.6
DINOv3	4.3	98.0	7.3	96.7	5.4	97.5	4.4	98.1	25.6	82.2	1.8
FE2E	**4.1**	**97.7**	**6.6**	**96.0**	**3.8**	**98.7**	**4.4**	**97.5**	**22.8**	81.2	**1.6**

based on natural language commands. The field has since diversified with numerous innovative approaches. For instance, DragGAN [29] introduced a novel point-based interaction, allowing users to "drag" pixels to precisely deform object shapes. OmniControl [44] further enhances controllability by creating a unified framework that accepts diverse spatial guidance signals for both synthesis and editing. This trend towards more powerful and versatile models is also reflected in large-scale systems like UniWorld [23], which uses a unified transformer for multi-modal understanding and generation, Step1X-Edit [25], fine-tuned from the FLUX architecture for superior instruction following, and multi-modal editors like Qwen-Image [48], which leverage Large Language Models (LLMs) to build more comprehensive visual editing frameworks.

We also note several recent studies from our broader collaborators on adjacent topics, including visual geometry and 4D reconstruction [4, 5, 13], 3D generation, editing, and multimodal understanding [51–53], controllable generation, virtual try-on, and adaptive inference for image editing [14, 33, 46, 49, 57, 58], controllable diffusion and language-guided depth estimation [59–61], preference alignment or reinforcement learning for diffusion and flow-based models [31, 32, 41, 42], and a broader set of downstream applications such as compositional vision-language adaptation, autonomous driving, action recognition, and risk analysis [19, 54–56, 65].

F. Addition Experiments Results

Comparison with Concurrent Unified Model The field of dense geometry estimation is advancing rapidly, with the task of depth estimation particularly fast. Recently, several concurrent works have been explored to unify the visual tasks, which also include depth estimation benchmarks. As shown in Table 2, our method consistently achieves the top average ranking, even though they are trained with extremely huge data compared to FE2E (e.g., Qwen Image utilizes billions of samples, and DINO v3 is trained on 1.7 billion images).

Additional Qualitative Comparison Fig. 1 presents a qualitative comparison between FE2E and other methods. The results demonstrate that our approach produces more refined and accurate depth predictions, particularly in structurally complex regions that may not be fully captured by quantitative metrics. Furthermore, as illustrated in Fig. 2, FE2E consistently delivers precise surface normal predictions, effectively handling intricate geometries and diverse environments. These results highlight the robustness of our method in fine-grained prediction tasks.

Table 3. Performance comparison of different models.

Methods	Marigold	Lotus-D	Qwen Image	DINO v3	**FE2E**
MACs	133T	2.65T	2.13P	14.5T	**28.9T**
RunTime	9.67s	212ms	63.4s	632ms	**1.78s**
AbsRel	6.5	6.1	6.6	5.4	**3.8**

G. Limitations and Future Work

Large computational load We present the inference latency and computational complexity of the FE2E model in Table 3, alongside comparisons with previous SD-based and unified methods. Although incorporating DiT does lead to a notable increase in computational complexity relative to other self-supervised approaches, FE2E strikes a trade-off between performance and computational efficiency.

Diversifying foundation models The field of image editing is evolving rapidly, and our approach is designed to be model-agnostic. In future work, we plan to incorporate a broader range of editing models to further substantiate the motivation and conclusions presented in this paper.

Scaling up the training data While a key contribution of this work is demonstrating strong generalization performance with a limited amount of training data, we still anticipate that scaling up the training dataset could further improve the model's capabilities. This direction is meaningful for domains that are not sensitive to computational complexity but require extremely high prediction accuracy. We leave the exploration for future research.

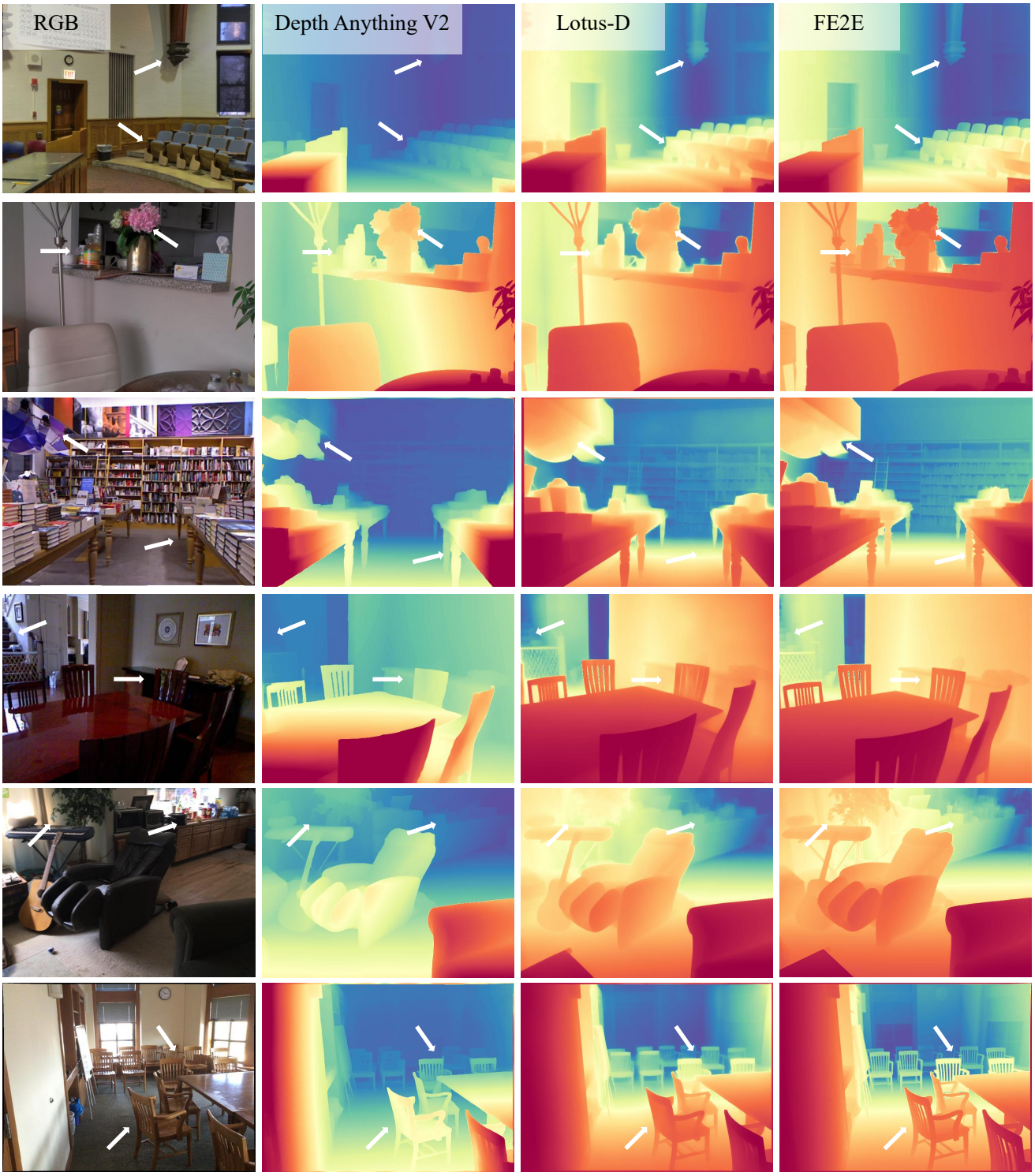

Figure 1. **Additional qualitative comparison on zero-shot affine-invariant depth estimation.** FE2E achieves more accurate depth predictions, particularly in structurally complex regions. White arrows highlight these improvements.

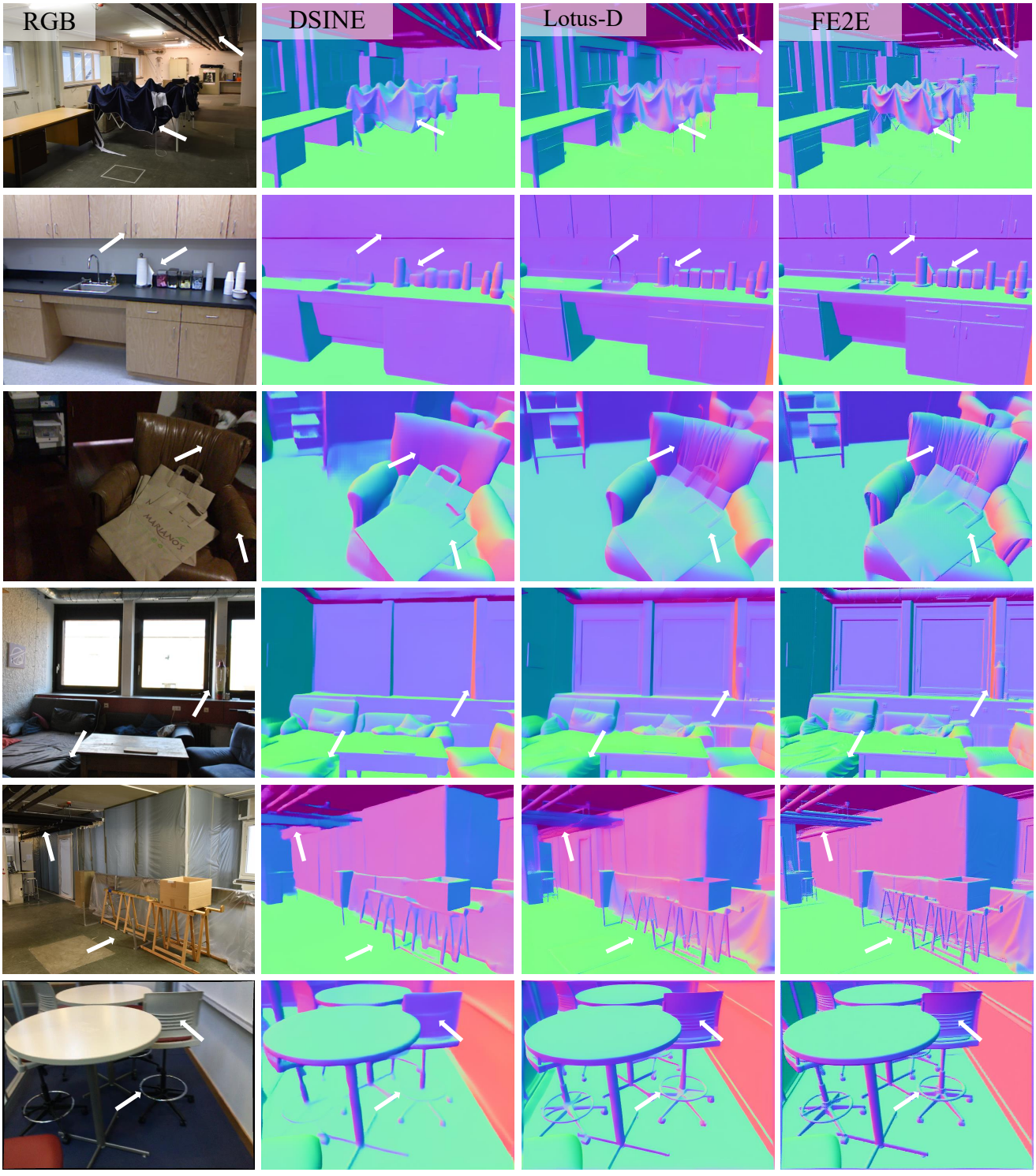

Figure 2. **Additional qualitative comparison on zero-shot surface normal estimation.** FE2E offers improved accuracy, particularly in detailed and complex regions.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012.

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[4] Chong Cheng, Yu Hu, Sicheng Yu, Beizhen Zhao, Zijian Wang, and Hao Wang. Reggs: Unposed sparse views gaussian splatting with 3dgs registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8100–8109, 2025.

[5] Chong Cheng, Xianda Chen, Tao Xie, Wei Yin, Weiqiang Ren, Qian Zhang, Xiaoyuang Guo, and Hao Wang. Longstream: Long-sequence streaming autoregressive visual geometry. *arXiv preprint arXiv:2602.13172*, 2026.

[6] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[7] Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. In *ICCV*, 2025.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[11] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[13] Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. *arXiv preprint arXiv:2511.19971*, 2025.

[14] Dongyang Jin, Ryan Xu, Jianhao Zeng, Rui Lan, Yancheng Bai, Lei Sun, and Xiangxiang Chu. Semantic context matters: Improving conditioning for autoregressive models. *arXiv preprint arXiv:2511.14063*, 2025.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[18] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[19] Zong Ke, Yuqing Cao, Zhenrui Chen, Yuchen Yin, Shouchao He, and Yu Cheng. Early warning of cryptocurrency reversal risks via multi-source data. *Finance Research Letters*, page 107890, 2025.

[20] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[22] Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing, 2025.

[23] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[24] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022.

[25] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[26] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022.

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022.

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[29] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[31] Bowen Ping, Chengyou Jia, Minnan Luo, Changliang Xia, Xin Shen, Zhuohang Dang, and Hangwei Qian. Paco-rl: Advancing reinforcement learning for consistent image generation with pairwise reward modeling. *arXiv preprint arXiv:2512.04784*, 2025.

[32] Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026.

[33] Xiangyan Qu, Zhenlong Yuan, Jing Tang, Rui Chen, Datao Tang, Meng Yu, Lei Sun, Yancheng Bai, Xiangxiang Chu, Gaopeng Gou, Gang Xiong, and Yujun Cai. From scale to speed: Adaptive test-time scaling for image editing, 2026.

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[39] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[41] Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization. *arXiv preprint arXiv:2505.10250*, 2025.

[42] Wenhao Shen, Hao Wang, Wanqi Yin, Fayao Liu, Xulei Yang, Chao Liang, Zhongang Cai, and Guosheng Lin. Vlm-guided group preference alignment for diffusion-based human mesh recovery. *arXiv preprint arXiv:2602.19180*, 2026.

[43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[44] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025.

[45] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[46] Honglie Wang, Yan-Ming Zhang, Wangzi Yao, Fei Yin, and Cheng-Lin Liu. Learning to generate stylized handwritten text via a unified representation of style, content, and noise. In *The Fourteenth International Conference on Learning Representations*, 2026.

[47] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization, 2025.

[48] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[49] Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. *arXiv preprint arXiv:2507.19946*, 2025.

[50] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[51] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference, 2024.

[52] Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025.

[53] Junliang Ye, Shenghao Xie, Ruowen Zhao, Zhengyi Wang, Hongyu Yan, Wenqiang Zu, Lei Ma, and Jun Zhu. Nano3d: A training-free approach for efficient 3d editing without masks, 2025.

[54] Qian Yu, Zong Ke, Guofu Xiong, Yu Cheng, and Xiaojun Guo. Identifying money laundering risks in digital asset transactions based on ai algorithms. In *2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC)*, pages 1081–1085. IEEE, 2024.

[55] Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. AutoDrive-R^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.

[56] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al. Video-star: Reinforcing open-vocabulary action recognition with tools. *arXiv preprint arXiv:2510.08480*, 2025.

[57] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8372–8382, 2024.

[58] Jianhao Zeng, Yancheng Bai, Ruidong Chen, Xuanpu Zhang, Lei Sun, Dongyang Jin, Ryan Xu, Nannan Zhang, Dan Song, and Xiangxiang Chu. Eevee: Towards close-up high-resolution video-based virtual try-on. *arXiv preprint arXiv:2511.18957*, 2025.

[59] Ziyao Zeng, Jingcheng Ni, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, and Alex Wong. Iris: Integrating language into diffusion-based monocular depth estimation. *arXiv preprint arXiv:2411.16750*, 2024.

[60] Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. Rsa: Resolving scale ambiguities in monocular depth estimators through language descriptions. *Advances in neural information processing systems*, 37: 112684–112705, 2024.

[61] Ziyao Zeng, Jingcheng Ni, Ruyi Liu, and Alex Wong. Coffee: Controllable diffusion fine-tuning. *arXiv preprint arXiv:2511.14113*, 2025.

[62] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[63] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[64] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.

[65] Heng Zhou, Jing Tang, Jusheng zhang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, and Jiawei Yao. Comem: Compositional concept-graph memory for vision–language adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.