One Diffusion to Generate Them All

Duong H. Le[1,*] Tuan Pham[2,*] Sangho Lee[1] Christopher Clark[1]
Aniruddha Kembhavi[1] Stephan Mandt[2] Ranjay Krishna[1,3] Jiasen Lu[1]
[1]Allen Institute for AI [2] University of California, Irvine [3] University of Washington * Equal contribution

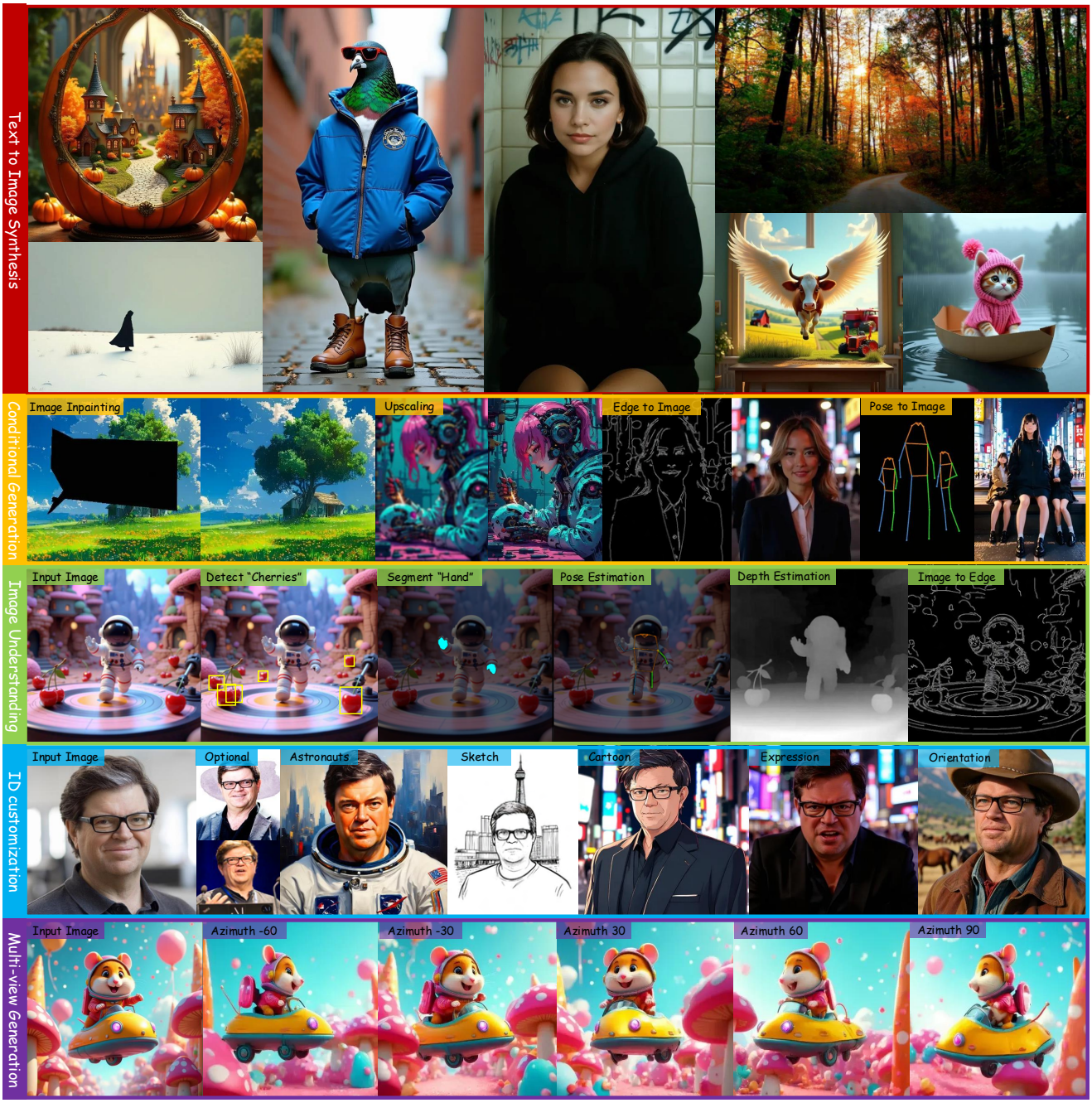

Figure 1. `OneDiffusion` is a unified diffusion model designed for both image synthesis and understanding across diverse tasks. It supports text-to-image generation (red box), conditional image generation from input images (orange box) and it's reverse task Image understanding (green box). It can also perform ID customization (blue box), and multi-view generation (purple box) with arbitrary number of input and output images.

Abstract

We introduce `OneDiffusion`*, a versatile, large-scale diffusion model that seamlessly supports bidirectional image synthesis and understanding across diverse tasks. It enables conditional generation from inputs such as text, depth, pose, layout, and semantic maps, while also handling tasks like image deblurring, upscaling, and reverse processes such as depth estimation and segmentation. Additionally,* `OneDiffusion` *allows for multi-view generation, camera pose estimation, and instant personalization using sequential image inputs. Our model takes a straightforward yet effective approach by treating all tasks as frame sequences with varying noise scales during training, allowing any frame to act as a conditioning image at inference time. Our unified training framework removes the need for specialized architectures, supports scalable multi-task training, and adapts smoothly to any resolution, enhancing both generalization and scalability. Experimental results demonstrate competitive performance across tasks in both generation and prediction such as text-to-image, multiview generation, ID preservation, depth estimation and camera pose estimation despite a relatively small training dataset. Our code and checkpoint are freely available at* `https://github.com/lehduong/OneDiffusion`.

1. Introduction

Diffusion models, particularly in text-to-image (T2I) generation, have recently achieved remarkable results. Models such as DALL-E [46], Imagen [46], and Stable Diffusion [15, 44, 50] have established new benchmarks for generating high-quality, photorealistic images from text prompts. Additionally, recent studies have demonstrated the effectiveness of diffusion models in various other computer vision tasks, such as depth estimation [23] or optical flow estimation [38, 53], *etc*. However, despite these advancements, diffusion models are typically trained individually for either T2I generation or specific tasks.

In contrast, large language models (LLMs) (*e.g*. GPT-4 [1]) have demonstrated their ability to function as universal models. They can perform a wide range of tasks across different domains without the need for task-specific modules, and can effectively handle tasks they have not been explicitly trained in a *zero-shot* manner. This universality has been immensely valuable; it has dramatically simplified using training and scaling these models, and ultimately led to better performance. This incentivizes us to ask whether diffusion models can become universal in a similar way.

Designing a unified architecture for diverse image synthesis tasks presents significant challenges. Current methods often depend on external add-ons to handle new tasks. For example, ControlNet [73] or T2I-Adapter [40] require specialized modules to encode the conditional inputs, and personalization models typically require encoding the identity through a pretrained facial recognition network and adding auxiliary losses to preserve identity [21, 63, 68]. Additionally, tasks vary widely in their input requirements. For instance, multi-view generation alone requires handling arbitrary input-output view combinations, posed or unposed images, and camera pose conditioning [18, 25, 35, 54, 62], while image understanding tasks require diverse outputs such as depth, pose, or segmentation. Finally, existing training recipes are often tightly tuned to particular tasks and therefore cannot be relied on to generalize between tasks.

In this work, we present `OneDiffusion` – a unified diffusion model that seamlessly supports bidirectional image synthesis and understanding across diverse tasks. Our approach enables a single model to perform multiple tasks without the need for external losses and add-ons. Inspired by recent advances in diffusion models for sequential data [7, 51, 74], we model all conditions and target images as a sequence of "views" with varying noise levels during training. At inference, any of the views can be used as a conditional input, or set to noise and then used to generate an output image. Conditioning text can also be changed to define the task, and specify additional conditioning details (*e.g*. camera pose). The simple, but flexible, framework allows our model to support many kinds image generation and image understanding tasks with a unified architecture and training objective.

To demonstrate how general purpose our training algorithm is, we train `OneDiffusion` completely from scratch. First, we train on text-to-image task to equip the model with general image synthesis abilities, then on our One-Gen dataset to learn the full set of tasks. Our final model has 2.8 billion parameters and is equipped with a diverse set of skills, shown in Figure 1. The model also adapts naturally to various resolutions, enabling zero-shot high-resolution generation even when such resolutions were not encountered during training.

We evaluate `OneDiffusion` on a diverse set of both generative and predictive tasks. On T2I, `OneDiffusion` efficiently generates high-quality images while utilizing fewer number of parameters. In the multiview generation task, `OneDiffusion` demonstrates performance comparable to state-of-the-art methods that are specifically designed and exclusively trained for this purpose. We also show that `OneDiffusion` supports novel conditioning setups, such as text-to-multi-view and image-to-multi-view. For high-variability tasks like face identification from a single image, the model is capable of generating multiple consistent images featuring diverse expressions and poses, demonstrating strong generalization to unseen domains.

2. Related work

Diffusion models for generative tasks Recent advancements in diffusion models have greatly improved image generation capabilities, with models like Stable Diffusion [3, 4, 15, 44, 50, 77] setting new standards in text-to-image synthesis. Beyond general image generation, controllable diffusion models such as ControlNet [73] and T2I-Adapter [40] allow fine-grained control via auxiliary inputs like edge or depth maps. Similar structured conditioning has been applied to inverse problems [42, 43, 57], enabling applications such as super-resolution or inpainting. Meanwhile, InstructPix2Pix [5] introduces natural language-guided image editing, making these tools more user-friendly. For personalized applications, identity-focused models, including IP-Adapter [68], InstantID [63], PhotoMaker [28], and PuLiD [21], personalize generation by conditioning on reference images. Moreover, in multi-view generation, recent methods [18, 35, 54, 62], employ camera ray embeddings or 3D geometry to achieve consistent viewpoints. Together, these innovations showcase the versatility of diffusion models in delivering controllable, personalized, and multi-perspective image synthesis.

Diffusion models for predictive tasks Beyond image generation and manipulation, diffusion models have also proven effective for predictive tasks within computer vision. Marigold [23] fine-tunes the Stable Diffusion model [50] to perform monocular depth estimation, demonstrating the adaptability of diffusion models for prediction-based applications. Furthermore, diffusion models have been utilized for optical flow estimation, as shown in the works of Saxena et al. [53] and Luo et al. [38], where the models predict pixel-level motion between consecutive frames. Additionally, Li et al. [27] trained a diffusion model for open-vocabulary semantic segmentation, showcasing the potential of these models for more complex vision tasks. Prior works have attempt to unify diffusion model for predictive tasks [17, 19]. These studies show that diffusion models are not only useful for generating images but also highly effective for various predictive tasks in computer vision.

Unified diffusion models Several attempts have been made to unify diffusion model for different type of controls [45, 65, 75]. However, they are limited to utilization of multiple image conditions. These models usually requires to design complicated adapters for different conditions. [36, 37, 60, 76] propose unified models for language and images. Concurrently, [64] propose finetuning multimodal large language model with diffusion objective on diverse tasks like text-to-image, editing, and subject-driven generation etc. In contrast, our model distinguishes itself by leveraging bidirectional capabilities of diffusion models and addressing a wide range of diverse tasks.

3. Methodology

3.1. Flow matching for generative modeling

Flow matching [2, 31, 34] is a framework for training continuous-time generative models by learning a time-dependent vector field that transports between two probability distributions. More specifically, a time-dependent vector field $u_t : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ governs the transformation from a base distribution p_0 to the target distribution $p_1 \approx q$ through an ODE $dx = u_t(x)dt$.

The solution of this ODE is a flow $\phi_t : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ with initial condition $\phi_0(x) = x$, and this flow characterizes a push-forward operation $p_t = [\phi_t]_{\#} p_0$, in which p_t is the density of samples $x \sim p_0$ transported by u from time 0 to time t. The goal is approximate this ODE using a learned time-dependent vector field parameterized as a neural network $v_\theta(t, x)$. Due to the intractable nature of u_t, [31] proposed to learn $v_\theta(t, x)$ using the conditional flow matching (CFM) objective:

$$\mathcal{L}_{\text{CFM}}(\theta) := \mathbb{E}_{t,q(z),p_t(x|z)} \left\| v_\theta(t, x) - u_t(x|z) \right\|^2 \quad (1)$$

This objective is equivalent to the original flow matching objective, and only requires the samples from the target distribution and a suitable conditional probability path.

3.2. Proposed Approach

Objective We cast the problem of image generation with multimodal conditions as sequential modeling. Inspired by previous work on diffusion model for sequential data [7, 51, 74], we jointly model all conditions and target images as a sequence of "*views*". Note that the number of views N is determined by tasks. Particularly, $N = 1$ for text-to-image tasks, $N = 2$ for image-to-image translation such depth/pose/image editing, etc, $N > 2$ for multiview generation or ID customization.

Mathematically, let N views $\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{H \times W \times D}$ be sampled from a training dataset $q(\mathbf{x}_1, ..., \mathbf{x}_N)$. Given time variables t_i, our goal is to learn a function $v_\theta(t_1, ..., t_N, \mathbf{x}_1, ..., \mathbf{x}_N) : [0, 1]^N \times \mathbb{R}^{N \times H \times W \times D} \to \mathbb{R}^{H \times W \times N \times D}$. Intuitively, v_θ serves as a generalized time-dependent vector field where each input $\mathbf{x}_i$ paired with its respective time variable t_i.

Learning v_θ enables arbitrary conditional generation, where any subset of views can be selected as conditions to generate the remaining views, as explained below. This setup allows us to dynamically configure the generation process, supporting flexible applications across a range of generative tasks.

Training Our training pipeline is visualized on the left side of Figure 2. At each training step, we independently sample $t_i \sim \text{LogNorm}(0, 1)$ [15] and Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$. This results in different noise levels for each

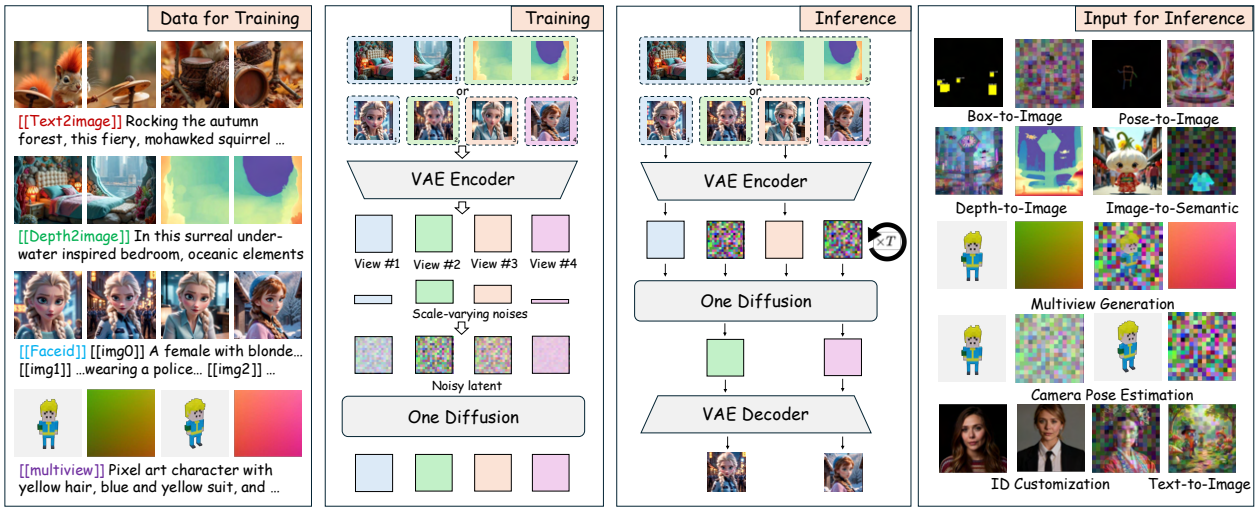

Figure 2. Illustration of training and inference pipeline for `OneDiffusion`. We encode the desired task for each sample via a special task token. During training we independently sample different diffusion timesteps for each view and add noise to them accordingly. In inference, we replace input image(s) with Gaussian noises while setting timesteps of conditions to 0.

views. We apply an interpolation-based forward process:

$$\mathbf{x}_i^{t_i} = \alpha_{t_i}\mathbf{x}_i + \beta_{t_i}\epsilon_i \tag{2}$$

where α_t and β_t satisfy the boundary conditions $\alpha_0 = 0, \alpha_1 = 1$ and $\beta_0 = 1, \beta_1 = 0$. Similar to [77], we adopt the linear interpolation schedule:

$$\mathbf{x}_i^{t_i} = t_i\mathbf{x}_i + (1 - t_i)\epsilon_i \tag{3}$$

the corresponding velocity field u_i for each view $\mathbf{x}_i$ is:

$$u_i(t_i, \mathbf{x}_i) = \mathbf{x}_i - \epsilon_i \tag{4}$$

with the aggregated target as $u = (\mathbf{x}_1 - \epsilon_1, ..., \mathbf{x}_N - \epsilon_N) \in \mathbb{R}^{N\times H\times W\times D}$, our training loss is the joint flow-matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|v_\theta(t_1, ..., t_N, \mathbf{x}_1, ..., \mathbf{x}_N) - u\|^2\right] \tag{5}$$

This flow matching objective [2, 31, 34] guides the model to learn the optimal velocity field v_θ by minimizing the difference from the target velocity field u.

Inference Our framework allows for both joint sampling and conditional sampling across any chosen set of views. In details, we define the target views we want to sample as $\mathbf{x}_K = \{\mathbf{x}_i\}_{i\in K}$, and the set of conditional views as $\mathbf{x}_{\setminus K} = \{\mathbf{x}_i\}_{i\notin K}$. To perform conditional sampling, we start by initializing the target views $\mathbf{x}_K$ as Gaussian noise. At each timestep t, we compute the corresponding time-dependent vector field $v_\theta^K(t, \mathbf{x}|\bar{\mathbf{x}}_{\setminus K})$ by fixing the conditional views to their known values $\mathbf{x}_{\setminus K} = \bar{\mathbf{x}}$ and setting their time variables to zero $t_{\setminus K} = 0$; while keeping the time variables of the target views as $t_K = t$:

$$v_\theta^K(t, \mathbf{x}|\mathbf{x}_{\setminus K} = \bar{\mathbf{x}}) = v_\theta(t_K = t, t_{\setminus K} = 0, \mathbf{x}_K = \mathbf{x}, \mathbf{x}_{\setminus K} = \bar{\mathbf{x}}) \tag{6}$$

Note that unlike v_θ, v_θ^K is a valid time dependent vector field, as all the views in K now has the same t. Thus, by integrating this vector field using an ordinary differential equation (ODE) solver, we can generate the conditional samples we are interested in. We illustration the inference on the right side of Figure 2.

3.3. Implementation Details

Model architecture We adopt the Next-DiT architecture [77] in our model. By leveraging a full transformer-based architecture, our model can work with different numbers of views N. We independently encode each frame *i.e.* images and conditions as latent $\mathbf{z} \in \mathbb{R}^{N\times H\times W\times C}$ with a VAE tokenizer [15] and concatenate them in N dimension. With flexible N, our approach establishes a universal framework which supports diverse input-modality with variable length. Following [77], we also apply 3D RoPE [58] for positional encoding, enabling generalization to different resolutions and aspect ratios.

Text-to-Image (1 views) With only a single "view", training and inference follow the same process as standard text-to-image diffusion models. We prepend the task label `[[text2image]]` to the caption to specify the task.

Image-to-Image (2 views) We set the first view as the target image and the second as the conditioning input. During inference, we can use one or both views for generation,

Figure 3. High-resolution samples **from text** of our `OneDiffusion` model, showcasing its capabilities in precise prompt adherence, attention to fine details, and high image quality across a wide variety of styles.

and the model is trained to produce the target image. For tasks like bounding box or semantic map generation, we add the hexadecimal color code and class label to the prompt. For instance, to segment a mouse with a yellow mask, the prompt is: `[[semantic2image]] <#FFFF00 yellow mask: mouse> photo of a ...` Further details are provided in the appendix.

ID Customization (2-4 views) We sample images of the same individual across views, concatenating captions for each input image and using a token `[[imgX]]` to denote each image. We also prepend the task label `[[faceid]]` to the captions. At inference, we can condition on an arbitrary number of images and generate multiple outputs, leading to more consistent results.

Multiview Generation (4-12 views) Inspired by [18], we use Plücker ray embeddings to represent camera poses. For each image patch, we calculate Plücker coordinates as $\boldsymbol{r} = (\boldsymbol{o} \times \boldsymbol{d}, \boldsymbol{d})$ using its ray origin $\boldsymbol{o}$ and direction $\boldsymbol{d}$. The result embedding has dimensions $H/8 \times W/8 \times 6$, matching the spatial size of the latent, and is replicated across channels to form a 16 channel embedding. Unlike [18], we treat ray embedding as a independent "view" following image latents as a unified sequence rather than concatenating by channels. This design allows flexible denoising, enabling multi-view image generation conditioned on camera poses or sampling ray embeddings to predict poses from image conditions, similar to RayDiffusion [72]. We scale the ray embeddings to have unit variance, as in [50].

As with other tasks, we prepend the task label `[[multiview]]` to the caption. During inference, we can substitute images or Plücker ray embeddings with Gaussian noise for multi-view generation and camera pose estimation, respectively.

Training details Our model is trained from *scratch* using a flow-matching objective. Similar to prior works [8, 15], we use a three stage training recipe. In the **first stage**, we pretrained the text-to-image model with resolution of 256^2 (500K steps) and 512^2 (500K steps). In the **second stage**, we continue training on a mixed of tasks, using 512^2 for T2I and 256^2 for other tasks, for a total of 1M steps. Finally, in the **last stage**, we finetune the model at a high resolution of (1024) for T2I. For ID customization fine-tuning, we use 2-5 views. For fewer views (2-3), we apply a resolution of 512^2, while for more views, we use 256^2 resolution.

During training, we use an in-batch sampling strategy at each stage, sampling tasks (T2I, Image-to-Image, ID customization, and multiview generation) with equal probability. The noise scheduler's shift value is set to 3, as suggested in [15]. We use AdamW optimizer with learning rate $\eta = 0.0005$. Training is performed on a TPU v3-256 pod with a global batch size of 256 in the first two phases, and the final fine-tuning stage is completed on 64 H100 GPUs using the same configuration.

4. One-Gen Datasets

Text-to-Image We leverage both public and internal (synthetic) datasets. The public datasets including: PixelProse [56], Unsplash, Coyo [6], JourneyDB [41]. Additionally, we use a 10M internal synthetic dataset consisting of images re-captioned with LLaVA-NeXT [32] and Molmo [11]. The

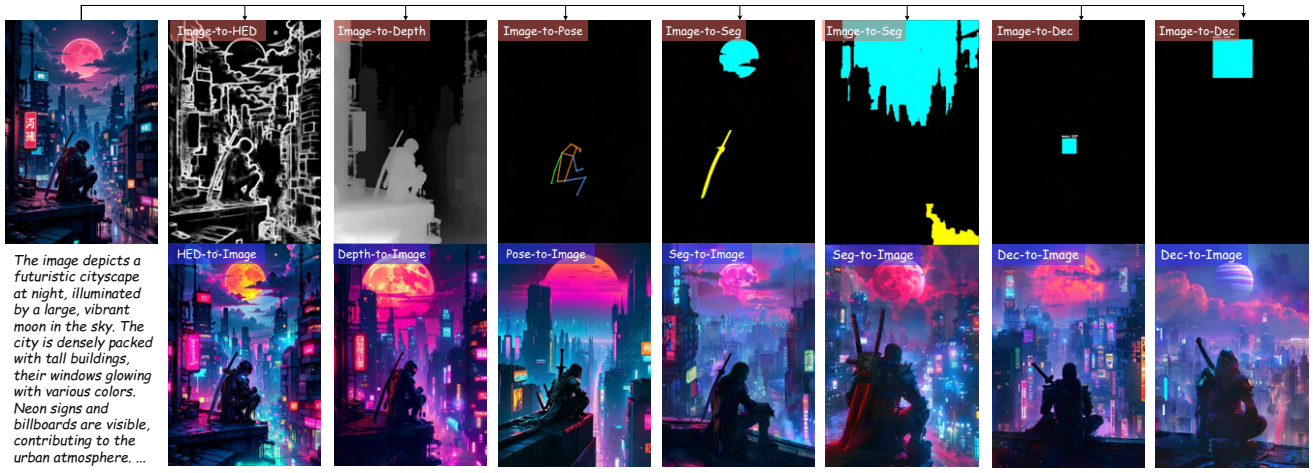

Figure 4. Illustration of our model capability to generate HED, depth, human pose, semantic mask, and bounding box from input image. For semantic segmentation, we segment the sword (highlighted in yellow) and the moon (highlighted in cyan) the first example, while segmenting road (yellow), sky (cyan) in the second. For object detection, We localize the head and moon (both highlighted in cyan). Leveraging these conditions, we can reverse the process to recreate a variant of the input image based on the same caption. Additionally, we can edit the image by modifying specific elements, such as replacing the moon with Saturn (last example).

length of the text description for each image varies from 100 to 150 words. When an original prompt is available, we use both the LLM-generated caption and the original caption.

Image-to-Image For simpler tasks *e.g.* deblurring, inpainting, image generation from canny edge, or upscaling, we use a $1M$-sample subset of our synthetic data and apply the corresponding pre-processor for each image to create an input condition. For more complex tasks, we create a synthetic dataset from outputs generated by Midjourney, Stable Diffusion, and Flux-dev following the below process:

- **Semantic Map and Detection** For each image, we use the LLaVA-NeXT [32] model to identify entities or subjects (e.g., person, shirt, dog, building), with a maximum of 10 entities per image. Based on these subject names from LLaVA-Next, we perform semantic segmentation using SAM [24] and extract bounding boxes. Each class is assigned a random color from a predefined list. This dataset contains 350K triplets consisting of a semantic map, bounding box, and the original image.
- **Depth Map** We generate the depth dataset by applying DepthAnything-v2 [67] to 500K images sampled from various datasets, including both real and synthetic images. Additionally, we caption 40K images from Hypersim dataset [49] with LLaVA-NeXT and incorporate these into the training set.
- **Human Poses** We collect a different subset with 50K images, primarily featuring human for pose conditioning. We use YOLOv5 to detect the bounding boxes for region of interests and apply ViTPose [66] for pose estimation.

ID Customization We collect a dataset of celebrities and characters from games and movies by from publicly available images. After filtering to ensure each subject has at least four images and removing NSFW content, the dataset includes approximately 60K subjects and a total of 1.3M images. We caption these images using the LLaVA-NeXT.

Multiview Generation We use the DL3DV-10K dataset [30], Objaverse [10], CO3D [48]. For Objaverse dataset, we utilize the 80K filtered split from LGM [59] and caption provided by Cap3D [39]. In the DL3DV dataset, we sample an image from each scene and caption it using LLaVA-Next. For CO3D, we exclude captions and include only the task token in the text input.

5. Experiments

We evaluate our `OneDiffusion` model on broad range of image generation and understanding tasks. We do not perform task-specific finetuning in any results. Details about additional qualitive examples are in the Appendix.

5.1. Text-to-Image

Qualitative results of `OneDiffusion` for text-to-image task is illustrated in Figure 3. Thanks to the diversity of our One-Gen dataset, the model can handle various art styles, spanning both artistic and photorealistic designs.

Following previous works [15], we evaluated the text-to-image capabilities of our model on GenEval benchmark [20]. For each prompt, we generate 4 images using Euler solver with 100 steps and guidance scale of 5. The results for `OneDiffusion`, along with those of baseline models, are presented in Table 1. Our model demonstrates strong performance compared to similarly sized baselines, excelling in multitasking capabilities despite being trained on a relatively smaller dataset.This performance is largely

Figure 5. Illustration of the multiview generation with single input image. We equally slice the azimuth in range of $[-45, 60]$ and elevation in range of $[-15, 45]$ for the left scenes. For the right scene, the azimuth range is set to $[0; 360]$ and elevation range is set to $[-15; 15]$.

Methods	Params (B)	# Data (M)	GenEval ↑
LUMINA-Next [77]	2.0	14	0.46
PixArt-Σ [9]	0.6	33	0.54
SDXL [44]	2.6	–	0.55
PlayGroundv2.5 [26]	2.6	–	0.56
IF-XL	5.5	1200	0.61
SD3-medium [15]	2.0	1000	0.62
Hunyuan-DiT [29]	1.5	–	0.63
DALLE3	–	–	0.67
FLUX-dev	12.0	–	0.67
FLUX-schnell	12.0	–	**0.71**
`OneDiffusion`	2.8	75	*0.65*

Table 1. Comparison of text-to-image performance on the GenEval benchmark at a resolution of 1024 × 1024.

Model	Condition	PSNR ↑
Zero123 [33]	1-view	18.51
Zero123-XL [12]	1-view	18.93
EscherNet [25]	1-view	20.24
	2-view	22.91
	3-view	24.09
`OneDiffusion`	1-view	19.01
	2-view (unknown poses)	19.83
	2-view (known poses)	20.22
	3-view (unknown poses)	20.64
	3-view (known poses)	21.79

Table 2. Comparison of NVS metrics across different number of condition view settings. Increasing the number of condition views improves the reconstruction quality.

attributed to the diversity of the dataset and the comprehensive captions provided for each sample.

5.2. Controllable Image generation

We show the experiment with image-to-image translation using various source domains, including HED, depth map, human pose, semantic map, bounding boxes. We report the qualitative results in Figure 4 and 19 in appendix. Generated images of `OneDiffusion` consistently conform various types of conditions by purely utilizing attention mechanisms and supplementary information from captions.

5.3. Multiview Generation

We assess our method's multiview generation capabilities using the Google Scanned Object dataset. Table 2 compares our approach (`OneDiffusion`) with state-of-the-art methods like Zero123, Zero123-XL, and EscherNet, which are tailored for multiview tasks. Unlike these, `OneDiffusion` supports variable conditional inputs and can handle additional views with unknown camera poses due to its flexible denoising framework.

In Table 2, `OneDiffusion` outperforms Zero123 and Zero123-XL in the 1-view condition and maintains strong results with unknown poses, e.g., a PSNR of 19.83 (2-view, unknown) vs. 20.22 (known), and 20.64 (3-view, unknown) vs. 21.79 (known). Figure 5 shows consistent multiview outputs from a single front-view image, with more examples in Appendix Figures 10 and 11. Our framework also enables text-to-multiview generation using only camera poses, as shown in Figure 12.

5.4. ID Customization

We further evaluate `OneDiffusion` on ID customization tasks, which involve using one or multiple ID images as inputs for personalized generation. To assess performance, we compare with STOA methods, including InstantID [63], PuLID [21], and PhotoMaker [28], using both qualitative and quantitative analyses. Our evaluation extends beyond the standard benchmark (Unsplash-50 [16]) to test generalization on ID customization tasks, such as varying expressions, viewpoints, and even non-human images.

Figure 6 illustrates examples of altering facial expressions and gaze directions (first row), changing viewpoints (second row), and customizing non-human IDs (third row). Our method achieves success in these tasks, where all other methods fail. Unlike previous approaches that rely on face embeddings and primarily "replicate" the original face, `OneDiffusion` employs attention mechanisms between images and text conditions. This enables flexible end-to-end training and generates more expressive outputs, making our method suitable for a wider range of applications. Intuitively, the mechanism that ensures consistent multiview generation also proves effective for manipulating camera angles in ID customization, highlighting its adaptability across related applications. Additional visualizations are provided in Figure 13 and 14.

We also present the quantitative results on the Unsplash-50 [16] benchmark in Table 3. This benchmark focuses

Figure 6. Illustration of ID customization using reference images. Unlike prior methods that rely on face embeddings and often fail to generalize, our model demonstrates superior generalization. It effectively adjusts facial expressions and gaze directions (first row), changes viewpoints (second row), and even customizes non-human IDs (third row). All results in the third row are generated from a single reference image, while InstantID fails as its face detector cannot detect faces in the input.

Method	ID ↑	CLIP-T ↑
PhotoMaker [28]	0.193	27.38
InstantID [63]	0.648	26.41
PuLID [21]	**0.654**	**31.23**
Ours	0.283	26.80

Table 3. Quantitative results on Unsplash-50.

solely on style changes and re-contextualization, where PuLID [21] demonstrates strong performance by leveraging embeddings from ID encoder networks trained on human faces for discrimination tasks. While this approach effectively preserves the identity traits of input images, it faces significant limitations when handling more complex face manipulations.

5.5. Depth Estimation

For image understanding tasks, we evaluate our model's performance on monocular depth estimation using standard benchmarks: NYUv2 [55] and DIODE [61]. We report the results in Table 4. Our model achieves competitive performance compared to baselines that leverage pretrained text-to-image diffusion models, such as Marigold [23]. Notably, as illustrated in Figures 15 and 16, our model demonstrates superior robustness than diffusion-based depth estimators like Marigold. Specifically, it excels in handling open-world images, including paintings, hazy weather, and unconventional textures.

Method	NYUv2		DIODE	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
DiverseDepth [69]	11.7	87.5	37.6	63.1
MiDaS [47]	11.1	88.5	33.2	71.5
DPT [47]	9.8	90.3	**18.2**	75.8
LeReS [70]	9.0	91.6	27.1	76.6
Omnidata [14]	7.4	94.5	33.9	74.2
HDN [71]	6.9	94.8	24.6	**78.0**
Marigold [23]	6.0	95.9	31.0	77.2
DepthAnything-2 [67]	**4.6**	**97.7**	27.1	74.8
Ours	6.8	95.2	29.4	75.2

Table 4. Comparison of depth estimation methods on NYUv2 and DIODE datasets. `OneDiffusion` achieves competitive performance compared to previous depth estimation methods.

6. Conclusion

Our experiments demonstrate that `OneDiffusion` achieves impressive results across a variety of tasks, including conditional T2I generation, depth estimation, open vocabulary semantic segmentation, pose estimation, multi-view generation, ID customization and camera pose estimation. We believe this work advances the capabilities of diffusion models, providing a versatile and scalable solution comparable to the flexibility offered by large language models. This represents a significant step toward developing a general-purpose vision model that can serve as the backbone for a wide variety of applications.

7. Acknowledgements

Stephan Mandt acknowledges support from the National Science Foundation (NSF) under an NSF CAREER Award IIS-2047418 and IIS-2007719, the NSF LEAP Center, by the Department of Energy under grant DE-SC0022331, the IARPA WRIVA program, the Hasso Plattner Research Center at UCI, the Chan Zuckerberg Initiative, and gifts from Qualcomm and Disney.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 3, 4

[3] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024. 3

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 5

[7] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024. 2, 3

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 5

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 7

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2022. 6

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Christopher Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Christopher Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jennifer Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Marie Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hanna Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 5, 2

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 8

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 4, 5, 6, 7

[16] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. *arXiv preprint arXiv:2404.03620*, 2024. 7

[17] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023. 3

[18] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 3, 5

[19] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720, 2024. 3

[20] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[21] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 2, 3, 7, 8

[22] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 1

[23] 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. 2, 3, 8, 9

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[25] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 2, 7

[26] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 7

[27] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 3

[28] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 3, 7, 8, 2

[29] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 7

[30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 6

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 4

[32] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5, 6

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 7

[34] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3, 4

[35] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3

[36] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[37] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023. 3

[38] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19167–19176, 2024. 2, 3

[39] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[40] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 3

[41] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023. 5

[42] Kushagra Pandey, Ruihan Yang, and Stephan Mandt. Fast samplers for inverse problems in iterative refinement models. *Advances in Neural Information Processing Systems*, 37:26872–26914, 2024. 3

[43] Kushagra Pandey, Farrin Marouf Sofian, Felix Draxler, Theofanis Karaletsos, and Stephan Mandt. Variational control for guidance in diffusion models. *arXiv preprint arXiv:2502.03686*, 2025. 3

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 7

[45] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3, 2

[46] 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. 2

[47] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 8

[48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 6

[49] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 6

[50] 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. 2, 3, 5

[51] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024. 2, 3

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[53] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[54] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3

[55] 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. 8

[56] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. 5

[57] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023. 3

[58] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[59] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 6

[60] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3

[61] 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. 8

[62] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 3

[63] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3, 7, 8

[64] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 3

[65] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 3

[66] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 6

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 6, 8, 2, 9

[68] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3

[69] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 8

[70] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to

recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 8

[71] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022. 8

[72] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 5, 1

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[74] Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[75] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models.(may 2023), 2023. 3, 2

[76] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 3

[77] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024. 3, 4, 7