UNIGP: TAMING DIFFUSION TRANSFORMER FOR PRIOR PRESERVED UNIFIED GENERATION AND PERCEPTION

Anonymous authors
Paper under double-blind review

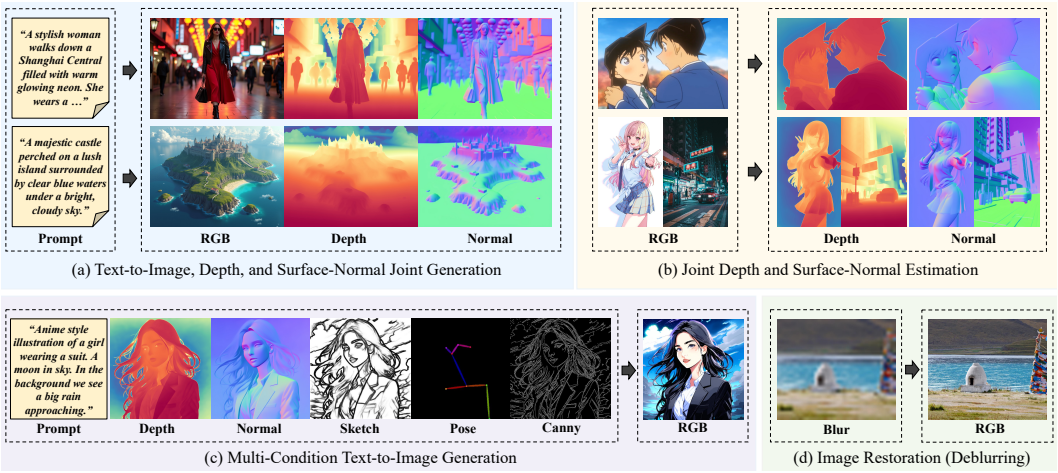


Figure 1: We present UNIGP, a Diffusion Transformer-based framework that simultaneously models RGB and dense distributions within a single framework, supporting: **(a)** Text to Image, Depth, and Surface-Normal joint generation; **(b)** Joint depth and surface-normal estimation; **(c)** Multi-condition text-to-image generation; and **(d)** Image restoration.

ABSTRACT

Recent advances in diffusion models have shown impressive performance in controllable image generation and dense prediction tasks (e.g., depth and normal estimation). However, existing approaches typically treat diffusion-based controllable generation and dense prediction as separate tasks, overlooking the potential benefits of jointly modeling the different distributions. In this work, we introduce UNIGP, a framework built upon MMDiT, which unifies controllable generation and dense prediction through simple joint training, without the need for complex task-specific designs or losses, while preserving the backbone's versatile priors. By learning controllable generation and prediction under different conditions, our model effectively captures the joint distribution of image-geometry pairs. UNIGP is capable of versatile controllable generation (as ControlNet), dense prediction (as Marigold) and joint generation (as JointNet). Specifically, the proposed UNIGP consists of DUGP and a unified dataset training strategy. The former, following the principle of Occam's razor, uses only a copied image branch of MMDiT to model dense distributions beyond RGB, while the latter integrates different types of datasets into a unified training framework to jointly model generation and perception tasks. Extensive experiments demonstrate that our unified model surpasses prior unified approaches and comparable with specialized methods. Furthermore, we show that through multi-task joint training, the performance of controllable generation and dense prediction can mutually enhance each other.

1 INTRODUCTION

In recent years, large-scale Text-to-Image (T2I) Diffusion models (Rombach et al., 2022; Esser et al., 2024; Midjourney, 2024; Huang et al., 2025) have attracted significant attention for their

exceptional performance in high-quality image generation. With advancements in generation quality, the community has increasingly focused on exploring more downstream tasks based on pretrained T2I models. These downstream tasks can be divided into two major categories: **1) Diffusion-based controllable generation:** ControlNet (Zhang et al., 2023) introduced the use of external conditions (e.g., depth, normal, and canny) to control the T2I model's generation. **2) Diffusion-based dense prediction:** Marigold (Ke et al., 2024) adapted T2I models to dense prediction tasks, such as monocular depth estimation, with later works extending this paradigm to normal estimation (Ye et al., 2024) or joint depth-normal estimation tasks (Garcia et al., 2024; Fu et al., 2025).

Although these methods have achieved remarkable results independently, they only model the transition within single distribution, limiting them to simple, single-task scenarios. Previous work, JointNet (Zhang et al., 2024), aimed to model multiple distributions simultaneously through intricate architectural design. While this approach resulted in a significant increase in parameter count, it achieved results that were not as optimal as intended. Recent works including UniCon (Li et al., 2024), OneDiffusion (Le et al., 2025) and JoDi (Xu et al., 2025) explored how to model the joint distribution but ignored the potential connection between perception and generation tasks, resulting in suboptimal performance compared to diffusion-based expert models.

In this paper, we propose UNIGP, a unified diffusion transformer (DiT) model that learns the global joint distribution of different modalities of an image with a flexible architecture. **First,** we adopt the Multi Modal DiT (MMDiT) framework (Esser et al., 2024) and introduce an effective weight initialization strategy. Following the principle of Occam's razor, we introduce **d**isentangled **u**nified **g**eneration and **p**erception branch (DUGP), which only copies the image branch from MMDiT to model distributions beyond RGB. **Second,** we propose a unified dataset and training strategy that combines datasets for both generation and perception tasks. By employing a binary loss weighting strategy, the model learns generation and perception tasks simultaneously without requiring task-specific model designs or losses. **Third,** we demonstrate through ablation studies that joint training of perception and generation tasks within a single framework leads to mutual improvement. Compared to representative controllable generation methods, our approach attains higher-quality controllability with less data. Unlike mainstream dense prediction methods, it preserves the backbone's generative capacity, mitigating catastrophic forgetting and enabling more accurate detail perception. As shown in Fig. 1, UNIGP supports various tasks within a single model: **1)** text-to-image, depth and surface-normal joint generation; **2)** joint depth and surface-normal estimation; **3)** multi-conditon to image generation; and **4)** image restoration.

To summarize, our main contributions are three-fold:

- We present UNIGP, an MMDiT-based framework that unifies generation and perception by jointly modeling multiple distributions. Following Occam's razor, UNIGP reuses only the image modeling components and parameters from the backbone for initialization, achieving stronger performance with fewer parameters.
- We propose a unified dataset and training strategy that seamlessly integrates the generation and perception datasets into a single training process, enabling UNIGP to efficiently learn both tasks.
- Extensive experiments demonstrate the superiority of UNIGP, surpassing existing unified models and performing on par with task-specific expert models. Moreover, we have shown that joint training of perception and generation yields mutual performance gains.

2 RELATED WORK

Controllable Diffusion Models. Controllable diffusion models are an important research direction aimed at using external conditions to control diffusion model generation. Representative works (Zhang et al., 2023; Mou et al., 2024; Mo et al., 2024) propose general frameworks for processing various spatial conditions. ControlNet (Zhang et al., 2023) and its subsequent models (Qin et al., 2024; Zhao et al., 2024; Sun et al., 2024b) extend T2I generation by encoding condition signals into latent representations using a trainable UNet encoder, injected into the backbone via zero convolution. However, these methods are limited to generation tasks and require large datasets for precise control. In contrast, our method jointly trains controllable generation and dense prediction tasks, enabling faster convergence with significantly less data. We propose a novel control design for MMDiT,

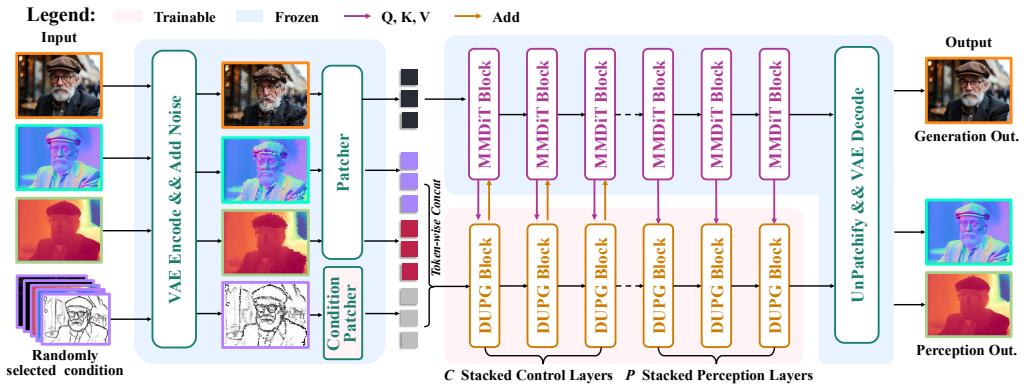


Figure 2: **Framework of UNIGP. 1)** Our inputs include: **a)** RGB images; **b)** Depth and Normal images; and **c)** Randomly selected condition as described in Sec. 3.1. **d)** Prompts (omitted for brevity). **2)** After VAE encoding and adding noise, the noisy RGB, depth and normal latents are fed into the backbone's patcher, while the clean condition latents are passed to the **Condition Patcher** of DUGP. Then, the tokens of noisy depth/normal and condition are concated in token-wise and passing through the **Stacked Control Layers** and **Stacked Perception Layers**. **3)** Finally, the backbone generates RGB images, while DUGP generates depth and normal maps.

enhancing control capabilities within the current SOTA diffusion framework and offering new insights to the community.

Diffusion Models in Perception Tasks. Currently, a notable trend involves adopting pretrained T2I diffusion models into dense prediction tasks Vandenhende et al. (2021); Xu et al. (2018), such as monocular depth estimation and surface normal estimation. Marigold (Ke et al., 2024) fine-tuned Stable Diffusion (SD) (Rombach et al., 2022) to generate precise depth maps conditioned on images, leveraging SD's strong geometric and semantic priors. Its success stems from training on high-quality synthetic datasets with perfect ground truth and smoothly transitioning from text-conditioned image generation to image-conditioned depth generation, preserving SD's generalization. Follow-up works (Gui et al., 2024; Garcia et al., 2024; He et al., 2024; Ye & Xu, 2024; Fu et al., 2025; Xu et al., 2024) on Marigold improve its performance and efficiency. StableNormal (Ye et al., 2024) extends Marigold's paradigm to surface normal estimation through a two-stage training. GeoWizard (Fu et al., 2025) jointly predicts depth and normal using two parallel UNet branches and cross-domain attention. The above models fine-tune diffusion for perception tasks, quickly losing versatile generative priors and reducing perception accuracy. Unlike prior works that compromise generation when adapted to perception, our method bridges generative and perception distributions, preserving generative priors while learning accurate perception.

Unified Diffusion Models. While less common than controllability or estimation, some works have pursued unified diffusion to enable a single model to model multiple modalities. LDM3D (Stan et al., 2023) jointly generates images and corresponding depth within an RGBD space. JointNet (Zhang et al., 2024) adopts a symmetric Unet structure to generate both images and depth, using an inpainting approach to support both generation and perception tasks. UniCon (Li et al., 2024) improves upon JointNet by using fewer additional parameters and an optimized training strategy, providing more versatile capabilities across different scenarios. Moreover, recent works such as OneDiffusion (Le et al., 2025) and JoDi (Xu et al., 2025) coarsely model different distributions using attention mechanism in Transformer and then fine-tune the entire model. However, previous methods treat generation and perception tasks merely as naive conditional image generation, overlooking the potential synergy between the two tasks and their differing training needs , resulting in suboptimal results. We demonstrate that under a tailored training strategy, jointly learning generation and perception distribution yields mutual gains, achieving optimal results in both tasks.

3 METHOD

To achieve unified generation and perception, we present UNIGP. We outline the preliminaries and problem setting in Sec. 3.1, and then introduce UNIGP in Sec. 3.2. Built upon the MMDiT framework, UNIGP jointly models generation and perception while enhancing both.

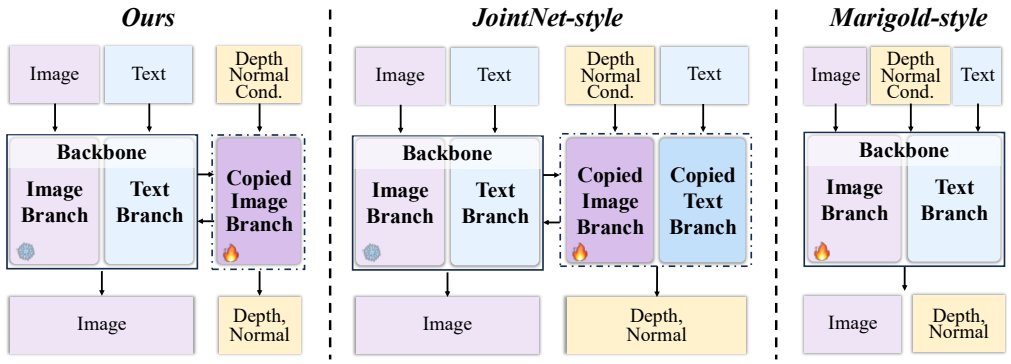


Figure 3: **Demonstration of representative design paradigms.** UNIGP copies only the image branch from MMDiT to model additional visual distributions while explicitly preserving the backbone's versatile priors. JointNet-style duplicates the entire backbone, incurring heavy computation; Marigold-style fine-tunes the backbone itself, quickly forgetting generative priors.

3.1 PRELIMINARIES AND PROBLEM SETTING

Diffusion Transformer (Peebles & Xie, 2023) replaces the commonly used U-Net backbone in diffusion models with Transformer (Vaswani et al., 2017). DiT first converts spatial inputs into a sequence of tokens and then performs denoising through a series of transformer blocks. DiT has achieved remarkable results in visual synthesis (Esser et al., 2024; Labs, 2024) . Among these, MMDiT is a powerful variant of DiT that has been widely adopted in recent SOTA DiT-based visual generation models (Esser et al., 2024; Labs, 2024). Specifically, as shown in the dashed box of Fig. 4, MMDiT models text and image features in two separate branches, applying full attention only once in each transformer block. We denote the text and image branches as $\mathcal{F}_t$ and $\mathcal{F}_i$, respectively.

Problem Setting. Our model is designed to jointly produce three outputs: an RGB image $\mathbf{x}$ and corresponding depth map $\mathbf{d}$, surface normal map $\mathbf{n}$ based on the input text prompt $\mathbf{c}_t$ and condition $\mathbf{c}$. We investigate three primary settings for this condition: ***i) Controllable Generation.*** The condition $\mathbf{c}$ is a spatial map, such as *Depth, Normal* and *Sketch*. The model synthesizes $\mathbf{x}$ while inferring its geometric properties ($\mathbf{d}$, $\mathbf{n}$); ***ii) Perception.*** The condition $\mathbf{c}$ is an RGB image. In this setting, the model's goal is to predict its geometric pairs ($\mathbf{d}$, $\mathbf{n}$); ***iii) Joint Generation.*** The condition consists solely of a text prompt, allowing for joint generation of the image and its geometry from $\mathbf{c}_t$. More formally, our objective is to learn a diffusion model $\mathcal{F}(\mathbf{c}_t, \mathbf{c})$ conditioned on the text prompt $\mathbf{c}_t$ and condition $\mathbf{c}$ to generate the corresponding image $\mathbf{x}$, depth $\mathbf{d}$, and surface-normal $\mathbf{n}$.

3.2 UNIGP: UNIFIED GENERATION AND PERCEPTION

UNIGP consists of two parts, with the overall framework illustrated in Fig. 2. In Sec. 3.2.1, we introduce DUGP, which enables our framework to model joint distributions of RGB and its geometry pairs in a plug-and-play manner. In Sec. 3.2.2, we present the unified dataset and training strategy designed to facilitate joint training for generation and perception tasks.

3.2.1 DUGP: DISENTANGLED UNIFIED GENERATION AND PERCEPTION BRANCH

To achieve unified generation and perception tasks on the MMDiT architecture, one straightforward option is to duplicate the entire backbone, as in JointNet (Zhang et al., 2024), to model additional distributions. However, this doubles the parameters and incurs substantial computational overhead. Another option, following Marigold (Ke et al., 2024), is to directly fine-tune the backbone itself, but this quickly erases the versatile generative priors.

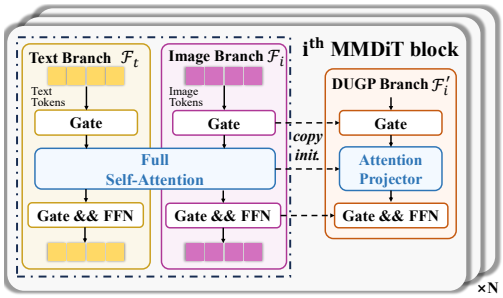


Figure 4: **The MMDiT block and initialization of DUGP.** ***Left dashed box***: The MMDiT consists of separate text and image branches . ***Right***: DUGP is initialized by copying the image branch of MMDiT.

We revisit the MMDiT architecture, as described in Sec. 3.1 and illustrated in the *left dashed box* of Fig. 4. MMDiT (denoted as $\mathcal{F}$) consists of a text branch $\mathcal{F}_t$ and an image branch $\mathcal{F}_i$, both symmetric

in design. Although the text branch $\mathcal{F}_t$ accounts for half the parameters of $\mathcal{F}$, it cannot model the visual distribution. Therefore, when modeling additional visual distributions, only the image branch $\mathcal{F}_i$ is the necessary entirety in the principle of Occam's razor. Thus, **we only copy a trainable image branch** $\mathcal{F}_i'$ based on $\mathcal{F}_i$ as an additional branch to model the visual distribution related to perception. As shown in Fig. 4 *right*, the new branch initialized by copying the image branch is referred to as DUGP. As illustrated in Fig. 2, We divide DUGP into **condition patcher**, **stacked control layers**, and **stacked perception layers**, where the control and perception layers consist of m and n blocks respectively, which sum to the backbone's l total layers.

Condition Patcher. To seamlessly integrate the condition $\mathbf{c}$, we introduce a parallel processing path at the DUGP's input section. First, we copy the patcher (i.e., the patchify layer) of the original model, $\mathcal{F}_i$. This new layer is then zero-initialized and serves as a dedicated module to process the control signal $\mathbf{c}$. This design is intended to prevent the control signal from affecting the model during the early stages of training and learn how to use condition gradually. Concurrently, the original pre-trained patcher is used to process the noisy depth and normal latents. Finally, the output tokens of noisy depth and normal latents and condition $\mathbf{c}$ are concatenated along the sequence dimension and fed into $\mathcal{F}_i'$.

Stacked Control Layers. To control the backbone, we enable DUGP to modulate it. First, DUGP employs the attention mechanism with backbone and shares the backbone's text branch $\mathcal{F}_t$. The following describes the attention calculation between DUGP and backbone:

$$[\mathrm{A_d}, \mathrm{A_n}] = \mathrm{Softmax}\left(\frac{[Q_d, Q_n] \cdot [K_t, K_d, K_n]^T}{\sqrt{d_k}}\right) \cdot [V_t, V_d, V_n], \tag{1}$$

$$\mathrm{A_c} = \mathrm{Softmax}\left(\frac{Q_c \cdot [K_t, K_i, K_c]^T}{\sqrt{d_k}}\right) \cdot [V_t, V_i, V_c], \tag{2}$$

In Eqs. (1) and (2), $[\cdot, \cdot]$ denotes sequence concatenation. $(\mathrm{A_m}, Q_m, K_m, V_m)$ denote the attention output, query, key, and value projections for modality m respectively.

At the end of each block in the stacked control layers, we obtain the output $\mathbf{I}$ of the backbone's image branch $\mathcal{F}_i$. Additionally, we obtain the condition's outputs of $\mathcal{F}_i'$, denoted as $\mathbf{C}$. Following ControlNet (Zhang et al., 2023), we add $\mathbf{C}$ to $\mathbf{I}$ through a zero-initialized linear layer:

$$\mathbf{I} = \mathbf{I} + \mathrm{Zero\_Linear}(\mathbf{C}) \tag{3}$$

Stacked Perception Layers. In the stacked perception layers, our goal is to extract features from the input condition $\mathbf{c}$ and the backbone to output the corresponding depth and normal. In addition to utilizing the information from $\mathbf{c}$ provided by the condition patcher in $\mathcal{F}_i'$, we also query features from the backbone. Specifically, we modify the computation in the stacked perception layers as follows:

$$[A_d, A_n] = \mathrm{Softmax}\left(\frac{[Q_d, Q_n] \cdot [K_d, K_n, K_t, K_i, K_c]^T}{\sqrt{d_k}}\right) \cdot [V_d, V_n, V_t, V_i, V_c], \tag{4}$$

Notably, the feature addition from Eq. (3) is bypassed in the perception layers, although the conditional attention for $\mathrm{A_c}$ (Eq. (2)) is retained.

Through the process described above, the parameters and computations of the original backbone are not altered. All modifications required for unified generation and perception are achieved within $\mathcal{F}_i'$. We rely merely on the powerful modeling capability inherent in the attention mechanism to achieve joint modeling of multiple distributions.

3.2.2 Unified Dataset Training Strategy

Unified Dataset. Diffusion-based generation and perception tasks have traditionally relied on entirely different datasets. Generation tasks typically employ datasets such as LAION (Schuhmann et al., 2022), JourneyDB (Sun et al., 2024a), or proprietary high-quality collections (Chen et al., 2024; Esser et al., 2024), which emphasize large-scale image quantity and aesthetic quality. In contrast, perception tasks often use synthetic datasets such as Hypersim (Roberts et al., 2021) and Virtual KITTI (Cabon et al., 2020), which provide precise geometric annotations but suffer from lower visual quality and limited realism compared to natural images. To unify generation and perception,

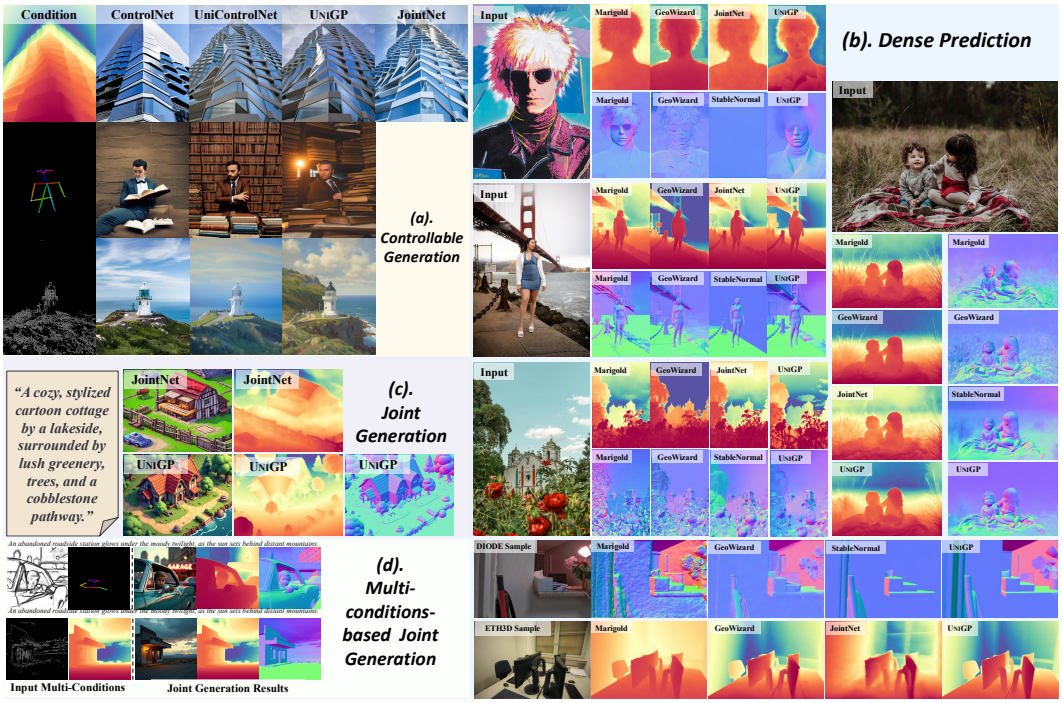


Figure 5: **Qualitative comparison and results** on **(a).** controllable generation, **(b).** dense prediction, **(c).** joint-generation and **(d).** multi-conditions-based joint generation tasks between UniGP and representative diffusion-based methods. UniGP outperforms previous diffusion-based experts and unified models across all tasks.

we integrate these two categories of data. Specifically, our dataset combines the filtered MultiGen-20M (Qin et al., 2024; Schuhmann et al., 2022), primarily supporting generation tasks, with filtered synthetic datasets (Roberts et al., 2021; Cabon et al., 2020), which are tailored for perception tasks.

Training Strategy and Objective. We mix generation and perception datasets together for training and randomly construct batches on the fly. Then, we adjust the loss weight adaptively based on the dataset from which each sample in the batch originates. Specifically, we learn a network v_θ that predicts the velocity field (Esser et al., 2024; Lipman et al., 2022) u_t given the image condition $\mathbf{c}$ and text prompt $\mathbf{c}_t$. We minimize the training objective as follows:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{t,x,d,n,\mathbf{c}_t,\mathbf{c}} \Big[& \lambda_g \left\| v_\theta(x_t, t, \mathbf{c}_t, \mathbf{c}) - u_t\left(x_t, t \mid x_1\right) \right\|^2 \\ & + \lambda_p \left\| v_\theta(d_t, t, \mathbf{c}_t, \mathbf{c}) - u_t\left(d_t, t \mid d_1\right) \right\|^2 \\ & + \lambda_p \left\| v_\theta(n_t, t, \mathbf{c}_t, \mathbf{c}) - u_t\left(n_t, t \mid n_1\right) \right\|^2 \Big]\end{aligned} \tag{5}$$

Specifically, x_1, d_1, and n_1 represent the RGB's latent and the corresponding latent for depth and normal, respectively. x_t, d_t, and n_t are the versions after adding noise at timestep t. $\lambda_g, \lambda_p \in \mathbb{R}^{\texttt{batch\_size}}$, For the i-th data sample, if it comes from the generation dataset, we set $\lambda_g[i] = 1$ and $\lambda_p[i] = 0$; if it comes from the perception dataset, we set $\lambda_g[i] = 0$ and $\lambda_p[i] = 1$. This strategy allows us to jointly optimize the generation and perception tasks.

4 Experiments

4.1 Experimental Settings

Implementation Details. We implement UniGP on SD3-medium (Esser et al., 2024), a commonly used MMDiT model, and optimize using Adam optimizer with a learning rate of 1×10^{-4}. The backbone is frozen, training only DUGP. All experiments are conducted on 16 NVIDIA A800 GPUs with a batch size of 64, over 40,000 steps.

Methods	Training Data	Text to Image				Depth to Image			Normal to Image			Canny to Image	
		FID ↓	CLIP ↑	GD ↓	GN ↓	FID ↓	CLIP ↑	RMSE ↓	FID ↓	CLIP ↑	RMSE ↓	FID ↓	CLIP ↑
Base T2I Models													
SD1.5 (Rombach et al., 2022)	-	23.60	30.89	-	-	-	-	-	-	-	-	-	-
SD3-medium (Esser et al., 2024)	-	20.09	31.77	-	-	-	-	-	-	-	-	-	-
Single-Control Methods													
ControlNet (Zhang et al., 2023)	-	-	-	-	-	19.80	25.30	13.86	22.18	25.10	18.15	16.16	25.34
T2I-Adapter (Mou et al., 2024)	-	-	-	-	-	20.08	25.67	15.62	-	-	-	18.76	25.25
SD3-ControlNet (Team, 2024)	-	-	-	-	-	18.00	27.09	12.50	-	-	-	17.92	26.88
Multi-Control Methods													
UniControlNet (Zhao et al., 2024)	10M	-	-	-	-	20.09	25.25	15.93	-	-	-	17.79	25.39
UniControl (Qin et al., 2024)	2.8M	-	-	-	-	20.67	25.51	14.07	20.04	25.78	17.81	16.69	25.15
ControlNetPlus (Github, 2024)	10M	-	-	-	-	19.27	27.99	13.20	20.11	26.80	16.00	16.12	27.90
Joint-Generation Methods													
LDM3D (Stan et al., 2023)	-	30.19	26.02	18.13	-	-	-	-	-	-	-	-	-
JointNet (Zhang et al., 2024)	2.56M	28.02	27.00	16.80	-	25.66	25.09	14.88	-	-	-	-	-
UniCon (Li et al., 2024)	16K	-	-	-	-	19.21	25.27	12.91	-	-	-	-	-
UniGP (Ours)	1M	21.05	31.50	6.05	6.41	17.95	28.41	6.89	18.60	28.45	7.12	15.91	30.35

Table 1: **Comparison of UniGP with task-specific baselines across four representative tasks.** The best and second best performances are highlighted. The symbol '-' indicates that the results for the models are not reported, or that the models do not support the corresponding tasks.

Training Datasets. 1) For perception tasks, we use Hypersim (Roberts et al., 2021) and Virtual KITTI (Cabon et al., 2020) following previous works, covering both indoor and outdoor scenes. Specifically, we use 59K complete samples filtered from these datasets. **2)** For generation tasks, we use 1M samples from MultiGen-20M (Qin et al., 2024), annotated with *depth, normal, canny, sketch, human pose, and blur*. For comprehensive details on the training datasets, please refer to Sec. A.1.1.

Evaluation Datasets and Metrics. For perception, generation, and image restoration tasks, we use commonly adopted benchmarks and metrics. Additionally, we propose the following metrics to better evaluate our model: **1)** GD and GN are computed by passing the generated RGB through GeoWizard (Fu et al., 2025) to recalculate depth/normal and then computing RMSE against the generated depth/normal. **2)** For the depth/normal to image task, following Anycontrol (Sun et al., 2024b), RMSE is calculated between the generated and input depth/normal. For more details on evaluation, please refer to Sec. A.1.2.

4.2 Main Results

4.2.1 Qualitative Evaluation

In Fig. 5, we show main results generated by UniGP on different tasks and compare them with representative methods. *Note that our results are generated by a single model.* **1)** First, in terms of controllable generation, our method supports more conditions than ControlNet (Zhang et al., 2023), which requires switching models for different conditions. Compared to UniControlNet (Zhao et al., 2024), our approach achieves better generation quality and improved image-text consistency. Additionally, we outperform both ControlNet (Zhang et al., 2023) and UniControlNet (Zhao et al., 2024) in control accuracy. **2)** Second, in perception tasks, UniGP outperforms previous generative-based dense prediction methods on both real-world samples and benchmark samples, providing fine-grained details and accurate depth and surface normal estimations while effectively handling complex geometries and diverse environments. Compared to the unified model JointNet (Zhang et al., 2024), UniGP not only estimates depth and normal simultaneously but also achieves significantly higher depth estimation accuracy. **3)** Third, in terms of joint-generation, JointNet can only generate RGB and depth simultaneously, whereas UniGP is capable of generating RGB, depth, and normal simultaneously. Moreover, UniGP produces better consistency and richer details compared to JointNet. We demonstrate more results and more perception tasks in Sec. A.3.2.

4.2.2 Quantitative Evaluation

Generation Results. We show quantitative generation results in Tab. 1, from our experiments, we observe the following: **1)** In terms of controllable generation, UniGP outperforms most previous methods. Thanks to the joint training of generation and perception tasks, our model achieves the lowest RMSE between the generated images and the given conditions in *depth/normal to image*, Thanks to the joint training of generation and perception tasks, our model achieves the lowest RMSE between generated images and the given conditions in *depth/normal-to-image*, demonstrating strong geometric input-output consistency. **2)** In terms of joint-generation (Text to Image column),

Method	Training Data	NYUv2 (Indoor) AbsRel↓	$\delta1$↑	$\delta2$↑	KITTI (Outdoor) AbsRel↓	$\delta1$↑	$\delta2$↑	ETH3D (Various) AbsRel↓	$\delta1$↑	$\delta2$↑	ScanNet (Indoor) AbsRel↓	$\delta1$↑	$\delta2$↑	Avg. Rank
Discriminative Methods														
DiverseDepth (Yin et al., 2021a)	320K	11.7	87.5	-	19.0	70.4	-	22.8	69.4	-	10.9	88.2	-	12.8
MiDaS (Ranftl et al., 2020)	2M	11.1	88.5	-	23.6	63.0	-	18.4	75.2	-	12.1	84.6	-	12.9
LeRes (Yin et al., 2021b)	354K	9.0	91.6	-	14.9	78.4	-	17.1	77.7	-	9.1	91.7	-	9.3
Omnidata (Eftekhar et al., 2021)	12.2M	7.4	94.5	-	14.9	83.5	-	16.6	77.8	-	7.5	93.6	-	7.4
DPT (Ranftl et al., 2021)	1.4M	9.8	90.3	-	10.0	90.1	-	7.8	94.6	-	8.2	93.4	-	7.4
HDN (Zhang et al., 2022)	300K	6.9	94.8	-	11.5	86.7	-	12.1	83.3	-	8.0	93.9	-	6.1
DepthAnything (Yang et al., 2024a)	62.6M	4.3	98.1	99.6	7.6	94.7	99.2	12.7	88.2	-	4.3	98.1	99.6	2.1
DepthAnything V2 (Yang et al., 2024b)	62.6M	4.5	97.9	99.3	7.4	94.6	98.6	13.1	86.5	-	4.2	97.8	99.3	2.5
Generative Methods														
GeoWizard (Fu et al., 2025)	280K	5.6	96.3	99.1	14.4	82.0	96.6	6.6	95.8	98.4	6.4	95.0	98.4	4.9
Marigold (Ke et al., 2024)	74K	5.5	96.4	99.1	9.9	91.6	98.7	6.5	95.9	99.0	6.4	95.2	98.8	4.2
JointNet (Zhang et al., 2024)	2.56M	13.6	84.1	86.0	29.9	59.6	62.3	19.2	78.7	80.2	11.9	84.8	86.7	14.8
UniCon (Li et al., 2024)	16K	7.9	93.9	-	-	-	-	-	-	-	9.2	91.9	-	8.3
OneDiffusion (Le et al., 2025)	75M	8.9	92.0	98.2	-	-	-	-	-	-	9.7	90.7	98.0	9.0
JoDi (Xu et al., 2025)	290K	8.3	92.0	98.2	-	-	-	-	-	-	9.7	90.7	98.0	8.7
UniGP (Ours)	59K	5.2	96.6	99.4	8.3	93.3	98.9	6.0	96.3	99.1	5.5	97.9	99.3	2.8

Table 2: **Quantitative comparison on zero-shot affine-invariant depth estimation** between UniGP and SOTA methods. **UniGP outperforms all other generative methods on average**, however, it lags behind the DepthAnything series on most metrics, which is trained on 62.6M images while UniGP is only trained on 0.059M images for perception capacity.

Method	Training Data	NYUv2 (Indoor) m.↓	11.25°↑	30°↑	ScanNet (Indoor) m.↓	11.25°↑	30°↑	iBims-1 (Indoor) m.↓	11.25°↑	30°↑	Sintel (Outdoor) m.↓	11.25°↑	30°↑	Avg. Rank
Discriminative Methods														
OASIS (Chen et al., 2020)	110K	29.2	23.8	60.7	32.8	15.4	52.6	32.6	23.5	57.4	43.1	7.0	35.7	9.8
Omnidata (Eftekhar et al., 2021)	12.2M	23.1	45.8	73.6	22.9	47.4	73.2	19.0	62.1	80.1	41.5	11.4	42.0	7.2
EESNU (Bae et al., 2021)	2.5M	16.2	58.6	83.5	-	-	-	20.0	58.5	78.2	42.1	11.5	41.2	6.1
Omnidata V2 (Kar et al., 2022)	12.2M	17.2	55.5	83.0	16.2	60.2	84.7	18.2	63.9	81.1	40.5	14.7	43.5	4.5
DSINE (Bae & Davison, 2024)	160K	16.4	59.6	83.5	16.2	61.0	84.4	17.1	67.4	82.3	34.9	21.5	52.7	1.9
Generative Methods														
Marigold (Ke et al., 2024)	74K	20.9	50.5	-	21.3	45.6	-	18.5	64.7	-	-	-	-	6.7
GeoWizard (Fu et al., 2025)	280K	18.9	50.7	81.5	17.4	53.8	83.5	19.3	63.0	80.3	40.3	12.3	43.5	6.0
StableNormal (Ye et al., 2024)	250K	18.6	53.5	81.7	17.1	57.4	84.1	18.2	65.0	82.4	36.7	14.1	50.7	4.1
JoDi (Xu et al., 2025)	290K	18.6	-	-	20.3	-	-	18.2	-	-	-	-	-	4.4
UniGP (Ours)	59K	16.4	59.2	83.4	14.9	65.1	86.0	17.3	66.5	82.8	35.0	20.1	55.1	1.7

Table 3: **Quantitative comparison on zero-shot surface normal estimation** between UniGP and SOTA methods. UniGP outperform all other discriminative and generative methods.

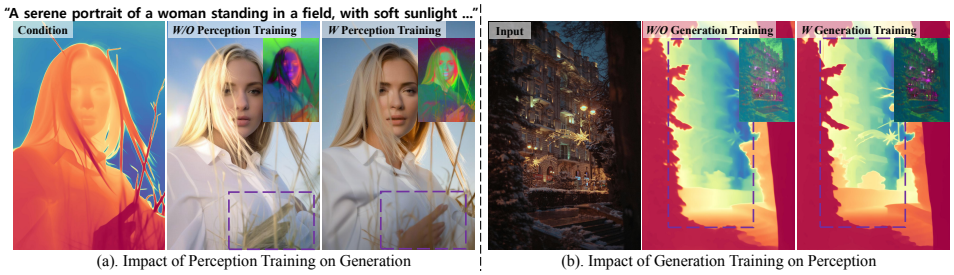
UniGP outperforms the baseline models across all metrics by a large margin, demonstrating the superiority of our method in preserving the backbone's generative capabilities as well as its joint-generation accuracy and effectiveness.

Perception Results. 1) We present depth estimation results in Tab. 2, where UniGP achieves the best performance among all generative baselines. However, it lags behind the SOTA discriminative DepthAnything series (Yang et al., 2024a;b) on indoor and outdoor datasets, but outperforms the DepthAnything series on diverse datasets (e.g., ETH3D). This is understandable, as UniGP trains its depth estimation using only 0.059M images from a single indoor and outdoor dataset—less than one-thousandth of the 62.6M images used by DepthAnything. However, generative models inherently possess extensive world knowledge. Unlike the DepthAnything series, which are trained discriminatively from scratch, UniGP retains the generative priors of the backbone, offering better generalization across diverse datasets. Notably, JointNet's depth estimation performance lags far behind ours, ranking 10th overall. **2)** We present surface normal estimation results in Tab. 3, where UniGP achieves comparable performance to DSINE (Bae & Davison, 2024), a recent SOTA discriminative model, and surpasses all other generative and discriminative methods in zero-shot surface normal estimation.

4.3 Ablation Study

In this section, we conduct ablation studies to evaluate UniGP. Due to computational constraints, the ablations are performed on representative benchmarks, with generation evaluated on COCO-5K and perception on NYUv2.

Relation between Generation and Perception. In Tab. 4 (a), we analyze the impact of removing the training for either perception or generation, reducing the training data for either task, and omitting the task-specific training strategy (i.e., $\lambda_g, \lambda_d = 1$) on the model's performance. The results show



(a). Impact of Perception Training on Generation (b). Impact of Generation Training on Perception

Figure 6: **Ablation study on the relationship between generation and perception.** Comparison areas are highlighted with purple boxes. The average attention map is visualized in the top-right corner. Adding perception training makes the generation results better align with the conditions, with clearer features, adding generation training improves the perception results with finer details.

Methods	Text to Image				Depth to Image			Normal to Image			Depth Esti.		Normal Esti.	
	FID ↓	CLIP ↑	GD ↓	GN ↓	FID ↓	CLIP ↑	RMSE ↓	FID ↓	CLIP ↑	RMSE ↓	AbsRel↓	$\delta 1$↑	m.↓	11.25°↑
(a) Relation between Generation and Perception														
w/o Perception Training	21.07	31.57	-	-	17.99	28.33	10.73	18.85	28.19	11.74	-	-	-	-
Half Perception Data	21.33	31.46	6.91	7.98	18.10	28.17	7.86	18.93	28.25	9.37	5.9	95.4	19.6	54.3
w/o Generation Training	-	-	-	-	-	-	-	-	-	-	5.5	96.5	17.1	56.0
Half Generation Data	21.86	31.40	6.97	8.06	18.97	28.12	8.40	19.91	28.14	9.72	5.9	95.1	19.7	55.6
w/o Training Strategy	25.59	29.29	11.14	13.05	23.21	26.61	8.74	24.45	27.14	10.36	6.9	94.3	22.6	46.9
(b) Balance between Stacked Control and Perception Layers														
(C,P) = (0,24)	-	-	-	-	-	-	-	-	-	-	5.4	96.3	17.2	58.0
(C,P) = (6,18)	25.11	29.85	9.15	9.87	21.15	28.04	9.21	23.25	27.41	10.25	5.3	96.0	17.1	57.6
(C,P) = (18,6)	24.83	29.48	10.38	11.43	20.91	28.33	7.85	20.09	28.48	9.28	5.9	95.4	18.2	54.9
(C,P) = (24,0)	21.78	31.37	7.56	7.95	19.63	28.18	10.65	18.91	28.21	10.49	-	-	-	-
(c) Different Design Choices														
JointNet Style	21.71	30.53	7.82	8.24	19.65	28.78	8.99	19.94	28.74	9.37	5.4	96.3	17.5	58.2
Marigold Style	25.71	28.46	12.15	12.03	28.35	28.01	10.84	28.85	27.14	11.61	6.0	95.0	19.7	54.9
Final Version														
UNIGP (Ours)	21.05	31.50	6.05	6.41	17.95	28.41	6.89	18.60	28.45	7.12	5.2	96.6	16.4	59.2

Table 4: **Ablation study on UNIGP.** Evaluating the impact of key components and design across representative benchmarks for each task.

that: **1)** Reducing either perception or generation training process or data scale negatively impacts the performance of the other, highlighting their mutual dependency. For example, incorporating perception training improves generation performance (e.g., GD, GN, and RMSE metrics), while adding generation training leads to better perception results. This phenomenon shows that when visual generation and perception share a unified space, their optimization objectives are intrinsically aligned, both relying on the better understanding of external/internal visual features, thus creating a cross-task synergy. **2)** Our proposed training strategy improves both generative and perceptive performance. As shown in Fig. 6, **(a)** perception training helps generation better adhere to conditional constraints (i.e., better RMSE), while **(b)** generation training improves perception by capturing finer details, which is further supported by the averaged attention map (Tumanyan et al., 2023). .

Balance between Stacked Control and Perception Layers. We denote the number of stacked control/perception layers as *(C, P)* with the default setting *C=P=12* . As shown in Tab. 4 (b), best performance is achieved when *C* and *P* are balanced.

Different Design Choices. As shown in Tab. 4 (c) and Fig. 3, we explored designs similar to JointNet (copying the entire backbone) and Marigold-style (fine-tuning backbone itself) on SD3 backbone. Both methods yielded significantly inferior performance compared to the proposed design.

5 CONCLUSION

In this work, we present UNIGP, an MMDiT-based framework for unified plug-and-play generation and perception tasks. UNIGP comprises DUGP and a unified dataset training strategy. The former, following the principle of Occam's razor, uses only a copied image branch of MMDiT to model dense distributions beyond RGB, while the latter combines various datasets into a unified training framework to jointly model generation and perception tasks. UNIGP demonstrates outstanding performance in both qualitative and quantitative evaluations, surpassing existing unified models and performing on par with SOTA expert models. Furthermore, our experiments reveal that perception and generation tasks mutually enhance each other within the diffusion framework.

6 THE USE OF LARGE LANGUAGE MODELS

We clarify that Large Language Models were solely used to refine the writing of this paper. They were not employed to generate, verify, or retrieve any factual content.

REFERENCES

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9545, 2024.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13137–13146, 2021.

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*, pp. 611–625. Springer, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

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.

Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 679–688, 2020.

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*, pp. 5828–5839, 2017.

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*, pp. 10786–10796, 2021.

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. *arXiv preprint arXiv:2403.03206*, 2024.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2025.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Github. Controlnetplus, 2024. URL https://github.com/xinsir6/ControlNetPlus.

Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, and Jianzhuang Liu. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18974, 2022.

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*, 2024.

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*, pp. 0–0, 2018.

Black Forest Labs. Flux, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2671–2682, 2025.

Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. *arXiv preprint arxiv:2410.11439*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Midjourney. Midjourney, 2024. URL https://www.midjourney.com.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 2024.

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 *Association for the Advancement of Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, pp. 4195–4205, 2023.

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. In *Advances in neural information processing systems*, volume 36, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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.

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

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*, pp. 10912–10922, 2021.

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*, pp. 10684–10695, 2022.

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*, pp. 3260–3269, 2017.

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.

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*, pp. 746–760. Springer, 2012.

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36, 2024a.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: Create your artwork with versatile control on text-to-image generation. *arXiv preprint arXiv:2406.18958*, 2024b.

InstantX Team. Sd3-medium-controlnet, 2024. URL https://huggingface.co/InstantX/SD3-Controlnet-Canny.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *TPAMI*, 44(7): 3614–3633, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024.

Yifeng Xu, Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. Jodi: Unification of visual generation and understanding via joint modeling. *arXiv preprint arXiv:2505.19084*, 2025.

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*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024b.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024.

Hanrong Ye and Dan Xu. Diffusionmtl: Learning multi-task denoising diffusion model from partially annotated data. In *CVPR*, 2024.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021a.

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*, pp. 204–213, 2021b.

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.

Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David Neil McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kv5xE1p3jz.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pp. 3836–3847, 2023.

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. In *Advances in neural information processing systems*, volume 36, 2024.