UniGP: Taming Diffusion Transformer for Prior-Preserved Unified Generation and Perception

Qin Guo[1], Hao Luo[2,3,4], Dongxu Yue[5], Weixuan Jin[5]
, Xiao Fu[6], Fan Wang[2], and Dan Xu[1,7†]

[1]The Hong Kong University of Science and Technology
[2]DAMO Academy, Alibaba Group, Zhejiang, China [3]Hupan Lab, Zhejiang Province
[4]Zhejiang University, Zhejiang, China [5]Tsinghua University
[6]The Chinese University of Hong Kong [7]Zeekr Automobile R&D Co., Ltd.
[†]Corresponding author

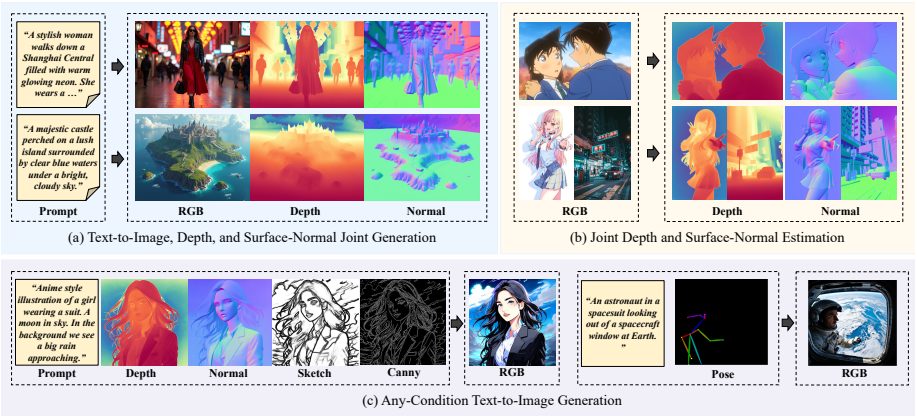

Fig. 1: We present UniGP, a Diffusion Transformer-based framework that simultaneously models RGB and dense distributions within a single framework, supporting: **(a)** Text to image, depth, and surface-normal joint generation; **(b)** Joint depth and surface-normal estimation; and **(c)** Any-condition text-to-image generation.

Abstract. Recent advances in diffusion models have shown impressive performance in controllable image generation and dense prediction tasks. However, existing approaches typically treat diffusion-based controllable generation and dense prediction as separate tasks, overlooking the potential benefits of jointly modeling the heterogeneous 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, dense prediction, and joint generation. 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 heterogeneous 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 performs on par with specialized methods. Furthermore, we demonstrate that multi-task joint training provides complementary benefits: generative priors enrich perceptual details, while perceptual learning improves structural alignment in generation. Project page: guoqincode.github.io/UniGP.

Keywords: Unified Diffusion Transformers · Diffusion-based Image Generation · Diffusion-based Perception

1 Introduction

Large-scale Text-to-Image (T2I) Diffusion models [4, 16, 22] have recently demonstrated exceptional high-quality image generation. This success has spurred exploration into downstream tasks based on pretrained T2I models, falling into two major categories: 1) Controllable generation, where models like ControlNet [35] use external conditions (*e.g.* depth, normal) to guide the T2I process. 2) Dense prediction, where models are repurposed for tasks like monocular depth estimation [11], later extended to normal estimation [32] and joint depth-normal estimation [5, 6].

Although these methods have achieved remarkable results independently, they only model the transition within a single distribution, limiting them to simple, single-task scenarios. One might ask: since regression-based feed-forward models already achieve highly accurate dense predictions, why should we pursue perception within a generative framework? The answer lies in the fundamental limitation of regression models: they act as passive, unidirectional observers. While they predict geometry accurately, they cannot internalize this geometric consciousness to actively guide complex creation processes. A truly unified model should embed perception into the generative latent space, allowing geometric understanding to act as a core generative prior.

Previous work, JointNet [34], aimed to model multiple distributions simultaneously through intricate architectural design. While this approach resulted in a significant increase in parameter count, it achieved suboptimal results. Recent works including UniCon [14], OneDiffusion [13], and JoDi [29] explored how to model the joint distribution but largely ignored the complementary mechanism between perception and generation tasks, resulting in performance gaps compared to diffusion-based expert models.

In this paper, we propose UniGP, a unified diffusion transformer (DiT) model that learns the joint distribution of different modalities of an image with a flexible architecture. **First,** we adopt the Multi-Modal DiT (MMDiT) framework [4]. Following the principle of Occam's razor, we introduce the **d**isentangled **u**nified **g**eneration and **p**erception branch (DUGP), which only copies the image branch

from MMDiT to model distributions beyond RGB, gracefully avoiding the massive computational overhead of replicating the entire backbone. **Second,** we propose a unified dataset and training strategy to bridge these heterogeneous tasks. By employing a simple binary loss weighting, our model jointly learns generation and perception, obviating the need for task-specific designs. **Third,** we demonstrate the complementary benefits between tasks. Our ablation studies show that jointly training perception and generation within a single framework allows perceptual learning to enforce strict spatial alignment for generation, while generative priors imbue perception with finer details.

Compared to representative controllable generation methods, UniGP achieves superior controllability with less data. Unlike mainstream dense prediction methods that quickly suffer from catastrophic forgetting, our approach preserves generative capacity, which enhances perceptual generalization. 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; and **3)** any-condition text-to-image generation.

To summarize, our main contributions are three-fold:

- We propose UniGP, an MMDiT framework that unifies generation and perception via a DUGP branch. Following Occam's razor, this branch models non-RGB data by reusing only the backbone's image branch parameters.
- We propose a unified dataset and a simple training strategy with a binary loss, enabling joint learning of heterogeneous tasks without task-specific architectures.
- Extensive experiments show UniGP significantly surpasses prior joint models and rivals task-specific experts, demonstrating clear complementary benefits between jointly trained perception and generation tasks.

2 Related Work

Controllable Diffusion Models. Controllable diffusion models aim to use external conditions to guide generation. ControlNet [35] and its subsequent models [18, 24, 36] 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 strictly to generation tasks and require massive datasets for precise control. In contrast, our method jointly trains controllable generation and dense prediction tasks, enabling faster convergence and superior geometric consistency with significantly less data. We propose a novel controlling module design and training strategy for MMDiT, enhancing capabilities within the current SOTA diffusion framework.

Diffusion Models in Perception Tasks. Currently, a notable trend involves adapting pretrained T2I diffusion models for dense prediction tasks [26], such as depth and surface normal estimation. Marigold [11] fine-tuned Stable Diffusion (SD) [22] to generate precise depth maps, leveraging SD's strong geometric priors. Follow-up works [5,6,8,9,28,33] improve its performance and efficiency. StableNormal [32] extends this paradigm to surface normal estimation. GeoWizard [5]

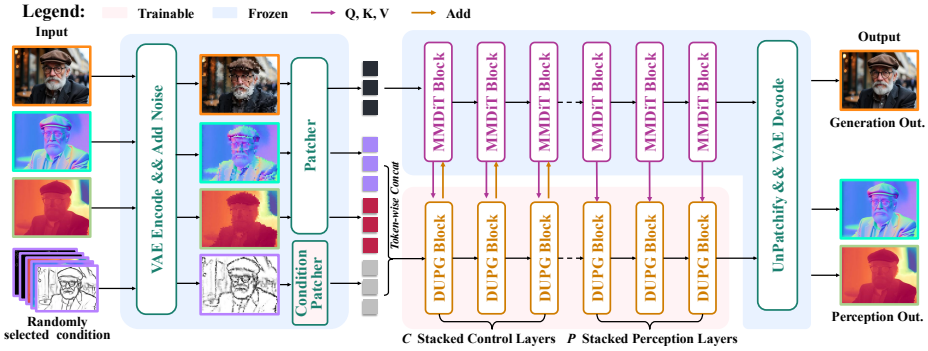

Fig. 2: Framework of UniGP. 1) Our inputs include: **a)** RGB images; **b)** Depth and Normal images; and **c)** Randomly selected condition as described in Sec. 3.1. **d)** Prompts (omitted for brevity). **2)** After VAE encoding and adding noise, the noisy RGB, depth and normal latents are fed into 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 concatenated token-wise and pass through the **Stacked Control Layers** and **Stacked Perception Layers**. **3)** Finally, the backbone generates RGB images, while DUGP generates depth and normal maps.

jointly predicts depth and normal using parallel branches. The above models entirely fine-tune diffusion for perception tasks, quickly losing versatile generative priors and narrowing their applicability. Unlike prior works that compromise generation when adapted to perception, our method explicitly bridges generative and perception distributions, quarantining generative priors while learning highly accurate perception.

Unified Diffusion Models. While less common, some works have pursued unified diffusion to enable a single model to handle multiple modalities. LDM3D [23] jointly generates images and depth within an RGBD space. JointNet [34] adopts a symmetric UNet structure using an inpainting approach to support both generation and perception. UniCon [14] improves upon JointNet with fewer parameters. Recent works like OneDiffusion [13] and JoDi [29] coarsely model different distributions using Transformer attention before fine-tuning the entire model. However, previous methods treat generation and perception merely as conditional image generation, largely ignoring the underlying complementary mechanisms between the two tasks, resulting in suboptimal results. We demonstrate that under a tailored model design and dataset strategy, jointly learning generation and perception yields mutual benefits, achieving state-of-the-art results among unified frameworks.

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.

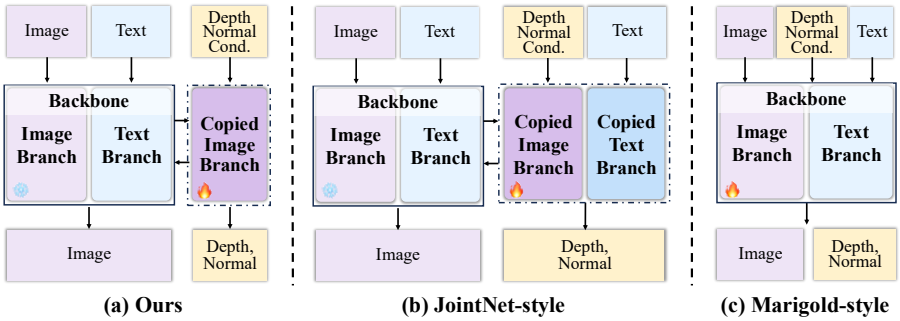

Fig. 3: Demonstration of representative design paradigms. UniGP copies only the image branch from MMDiT to model additional visual distributions while explicitly preserving the backbone's versatile priors. JointNet-style duplicates the entire backbone, incurring heavy computation; Marigold-style fine-tunes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

3.1 Preliminaries and Problem Setting

Diffusion Transformer [17] replaces the commonly used U-Net backbone in diffusion models with Transformer [27]. DiT first converts spatial inputs into a sequence of tokens and then performs denoising through a series of transformer blocks. Among these, MMDiT is a powerful variant widely adopted in recent SOTA visual generation models [4, 12]. 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 $\mathbf{x}$, depth map $\mathbf{d}$, and 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* or *Sketch*. The model synthesizes $\mathbf{x}$ while simultaneously inferring its geometric properties ($\mathbf{d}$, $\mathbf{n}$). ***ii) Perception.*** The condition $\mathbf{c}$ is an RGB image. Here, 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 the joint generation of the image and its geometry purely from $\mathbf{c}_t$. More formally, our objective is to learn a unified diffusion model $\mathcal{F}(\mathbf{c}_t, \mathbf{c})$.

3.2 UniGP: Unified Generation and Perception

UniGP consists of two parts, with the overall framework illustrated in Fig. 2. In Sec. 3.2, we introduce DUGP, which enables our framework to model joint distributions of RGB and its geometry pairs flexibly. In Sec. 3.2, we present the unified dataset and training strategy.

Disentangled Unified Generation and Perception To achieve unified generation and perception on the MMDiT architecture, one straightforward

option is to duplicate the entire backbone, as in JointNet [34]. However, this doubles the parameters and incurs substantial computational overhead. Another option, following Marigold [11], is to directly fine-tune the backbone itself, but this quickly erases the versatile generative priors.

We revisit the MMDiT architecture, as 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 $\mathcal{F}_i$. Although the text branch $\mathcal{F}_t$ accounts for half the parameters, it cannot model the visual distribution. Following the principle of Occam's razor, the image branch $\mathcal{F}_i$ is the only necessary entirety to model additional visual distributions. Thus, **we only copy a trainable image branch** $\mathcal{F}_i'$ based on $\mathcal{F}_i$ as an additional branch for perception-related visual distributions. As shown in Fig. 4 *right*, this new branch is referred to as DUGP. We divide DUGP into a condition patcher, stacked control layers, and stacked perception layers, where the control and perception layers consist of m and n blocks respectively.

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 of the original model, $\mathcal{F}_i$. This new layer is zero-initialized to prevent the control signal from abruptly affecting the model during early training. Concurrently, the original pre-trained patcher processes the noisy depth and normal latents. Finally, the output tokens of noisy depth/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, DUGP employs the attention mechanism and shares the backbone's text branch $\mathcal{F}_t$. The attention calculation between DUGP and the backbone is as follows, where $\sigma(\cdot)$ denotes the Softmax function:

$$[\mathbf{A}_d, \mathbf{A}_n] = \sigma\left(\frac{[\mathbf{Q}_d, \mathbf{Q}_n][\mathbf{K}_t, \mathbf{K}_d, \mathbf{K}_n]^T}{\sqrt{d_k}}\right)[\mathbf{V}_t, \mathbf{V}_d, \mathbf{V}_n], \tag{1}$$

$$\mathbf{A}_c = \sigma\left(\frac{\mathbf{Q}_c[\mathbf{K}_t, \mathbf{K}_i, \mathbf{K}_c]^T}{\sqrt{d_k}}\right)[\mathbf{V}_t, \mathbf{V}_i, \mathbf{V}_c]. \tag{2}$$

In Eqs. (1) and (2), $[\cdot, \cdot]$ denotes sequence concatenation. At the end of each block in the stacked control layers, we obtain the output $\mathbf{I}$ of the backbone's image branch $\mathcal{F}_i$ and the condition's outputs of $\mathcal{F}_i'$, denoted as $\mathbf{C}$. We then add $\mathbf{C}$ to $\mathbf{I}$ through a zero-initialized linear layer $\mathcal{L}_{\text{zero}}$:

$$\mathbf{I}' = \mathbf{I} + \mathcal{L}_{\text{zero}}(\mathbf{C}). \tag{3}$$

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. We modify the computation here as follows:

$$[\mathbf{A}_d, \mathbf{A}_n] = \sigma\left(\frac{[\mathbf{Q}_d, \mathbf{Q}_n][\mathbf{K}_d, \mathbf{K}_n, \mathbf{K}_t, \mathbf{K}_i, \mathbf{K}_c]^T}{\sqrt{d_k}}\right)[\mathbf{V}_d, \mathbf{V}_n, \mathbf{V}_t, \mathbf{V}_i, \mathbf{V}_c]. \tag{4}$$

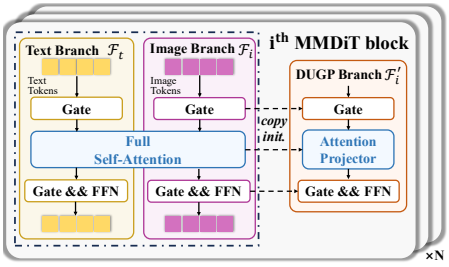

Fig. 4: The MMDiT block and initialization of DUGP. The *left dashed box* represents the MMDiT, which consists of separate text and image branches. The *right dashed box* represents DUGP , initialized by copying the image branch of MMDiT.

The feature addition from Eq. (3) is bypassed in the perception layers, forming the core mechanism for our "Prior Preserved" objective via strict gradient isolation. The fundamental dilemma in unifying generation and dense prediction lies in their conflicting optimization objectives: perception requires deterministic, discriminative pixel-level mapping, which often collapses the highly diverse distributions learned by generative models. By severing the additive $\mathcal{L}_{\text{zero}}$ path for perception layers, gradients originating from the dense prediction loss flow exclusively through the DUGP branch $\mathcal{F}'_i$. This prevents them from altering the pre-trained weights of the backbone $\mathcal{F}_i$. The backbone acts purely as a robust feature donor, forcing the perception layers to learn as non-invasive "readers" and ensuring that multi-modal generative priors are perfectly quarantined and preserved.

Handling Multiple Input Conditions. For scenarios requiring multiple simultaneous conditions, we duplicate the DUGP condition patcher for each condition. Crucially, the interaction with the backbone is performed via sequence concatenation along the same dimension. This means all condition latents are processed within a *single forward pass* of the backbone, effectively preventing the inference time from scaling linearly with the number of conditions and maintaining high efficiency.

Unified Dataset Training Strategy Diffusion generation and perception tasks traditionally use separate datasets. To unify these domains, our dataset integrates both, combining filtered generation data [18] with filtered synthetic perception data [2, 21]. We address the heterogeneous nature of this data by randomly constructing batches on the fly and adjusting the loss weight adaptively based on the dataset origin.

Specifically, we learn a network v_θ that predicts the velocity field [4, 15] u_t given the image condition $\mathbf{c}$ and text prompt $\mathbf{c}_t$. We minimize the training objective as follows:

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

Here, x_1, d_1, and n_1 represent the latents for RGB, depth, and normal, respectively. The routing weights $\lambda_g, \lambda_p \in \{0, 1\}$ act as binary task selectors. Rather than manually balancing complex task-specific loss scales, we dynamically assign these weights based on the data source of each sample within the mixed batch. This straightforward yet highly effective strategy allows us to jointly optimize tasks on

Algorithm 1 Joint Training Strategy of UniGP

Require: Pre-trained MMDiT backbone $\mathcal{F}$, initialized DUGP branch $\mathcal{F}'$, Generation Dataset $\mathcal{D}_g$, Perception Dataset $\mathcal{D}_p$
Ensure: Optimized DUGP parameters θ

while not converged **do**
  Sample a mixed batch $B = B_g \cup B_p$ where $B_g \sim \mathcal{D}_g$ and $B_p \sim \mathcal{D}_p$
  for each sample $s_i \in B$ **do**
    if $s_i \in B_g$ **then**
      Set loss weights: $\lambda_g^{(i)} \leftarrow 1$, $\lambda_p^{(i)} \leftarrow 0$
      Set condition $\mathbf{c} \leftarrow$ random spatial condition or NONE
    else
      Set loss weights: $\lambda_g^{(i)} \leftarrow 0$, $\lambda_p^{(i)} \leftarrow 1$
      Set condition $\mathbf{c} \leftarrow$ Original RGB image
    end if
    Sample timestep $t \sim \mathcal{U}(0, 1)$
    Add noise to latents (x_1, d_1, n_1) to get (x_t, d_t, n_t)
    Forward pass backbone $\mathcal{F}$ and DUGP $\mathcal{F}'$ with $\mathbf{c}_t$ and $\mathbf{c}$
    Compute flow matching loss $\mathcal{L}^{(i)}$ using Eq. (5)
  end for
  Compute batch gradient $\nabla_\theta \frac{1}{|B|} \sum_i \mathcal{L}^{(i)}$
  Update parameters θ of DUGP using Adam optimizer
end while

a strictly heterogeneous dataset. The complete joint training pipeline, detailing this adaptive sample-level routing, is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Implementation Details. We implement UniGP on SD3-medium and optimize using the Adam optimizer with a learning rate of 1×10^{-4}. The backbone is frozen, meaning we only train the DUGP branch. All experiments are conducted on 16 NVIDIA A800 GPUs with a total batch size of 64, for 40K steps.
Training Datasets. Our unified dataset integrates two heterogeneous data sources. **1) Perception Component:** We use synthetic datasets covering both indoor and outdoor scenes to ensure perfect geometric ground truth. Specifically, we use 39K complete samples filtered from Hypersim (indoor) resized to 576×768, and 20K samples from Virtual KITTI 2 (outdoor) with the far plane set at 80m. **2) Generation Component:** We filter and randomly select 1M samples from MultiGen-20M, ensuring a minimum edge length of 768. We utilize its existing annotations (*normal, canny, sketch, human pose*). Importantly, we re-annotate the *depth* using GeoWizard because original discriminative annotations like MiDaS or DepthAnything output inverse normalized depth, which is structurally incompatible with diffusion-based generation. During training, perception samples use the original RGB image as the condition, while generation samples randomly select a spatial condition or NONE (for joint-generation) with equal probability.

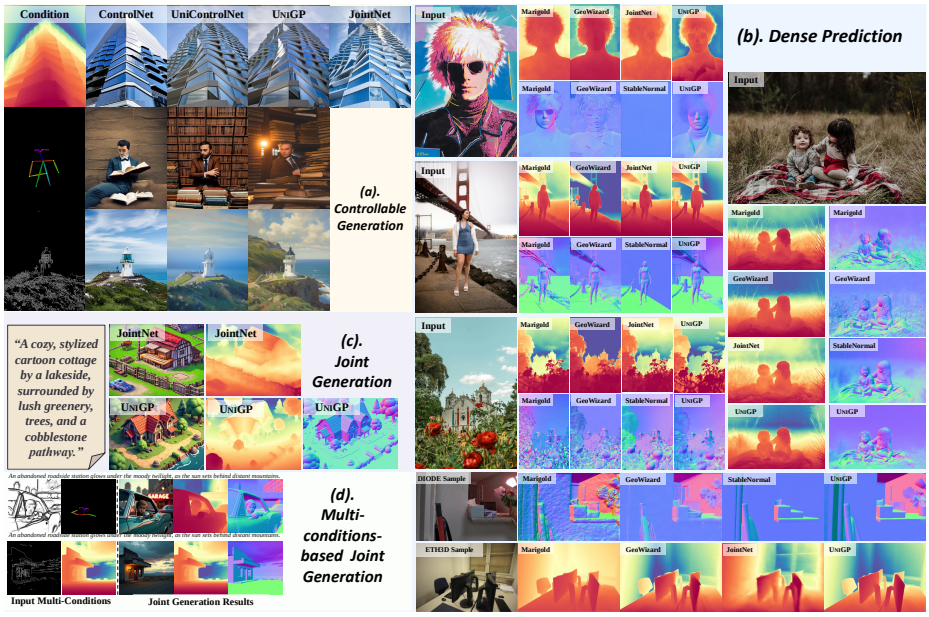

Fig. 5: Qualitative comparison and results on **(a)** controllable generation, **(b)** dense prediction, **(c)** joint-generation and **(d)** multi-condition-based joint generation tasks between UniGP and representative diffusion-based methods. UniGP outperforms previous diffusion-based experts and unified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

Evaluation Datasets and Metrics. 1) Perception Tasks: For zero-shot affine-invariant depth estimation, we evaluate on NYUv2, ScanNet, KITTI, and ETH3D using absolute mean relative error (AbsRel) and inlier metrics (δ_1, δ_2). For surface normal estimation, we test on NYUv2, ScanNet, iBims-1, and Sintel, reporting mean angular error (m.) and accuracy ($11.25°, 30°$). **2) Generation Tasks:** We evaluate on 5K samples from the MSCOCO validation set using Fréchet Inception Distance (FID) and CLIP Score. **3) Cross-modal Consistency Metrics:** Existing metrics like FID or CLIP score primarily evaluate the visual quality or text-alignment of generated images, completely ignoring whether the jointly generated RGB and geometries are structurally aligned. To explicitly measure this geometric consistency in joint-generation tasks, we propose **GD** (**G**enerated **D**epth RMSE) and **GN** (**G**enerated **N**ormal RMSE). These are computed by passing the generated RGB through an expert model (GeoWizard) to recalculate the pseudo-ground-truth depth/normal, and then calculating the RMSE against our model's directly generated depth/normal. Lower GD/GN strictly indicates higher structural alignment and geometric consistency. For controllable generation, RMSE is calculated directly between the generated geometry and the input condition.

Comparison Methods. We benchmark UniGP against state-of-the-art methods across multiple domains. **1) Controllable Generation:** We compare with single-control models including ControlNet [35] and SD3-ControlNet [25] (*i.e.* SD3-CNet), as well as multi-control models including UniControlNet [36] and

ControlNetPlus [7] (*i.e.* CNetPlus). **2) Joint Generation:** We evaluate against recent unified frameworks, including LDM3D [23], JointNet [34], and UniCon [14]. **3) Dense Prediction:** For depth and surface normal estimation, we comprehensively compare against standard regression experts (MiDaS [20], Omnidata V1/V2 [3, 10], DPT [19], DepthAnything V1/V2 [30, 31], and DSINE [1]) as well as recent generative perception methods (Marigold [11], GeoWizard [5], StableNormal [32], OneDiffusion [13], and JoDi [29]).

4.2 Main Results

Qualitative evaluation In Fig. 5, we show main results generated by UniGP on different tasks and compare them with representative methods. *Note that all our results are generated by a single model.* **1)** First, in controllable generation, our method supports more conditions than ControlNet and achieves better generation quality and text-image consistency than UniControlNet. **2)** Second, in perception tasks, UniGP outperforms previous generative-based dense prediction methods, providing fine-grained details and accurate depth/normal estimations for complex geometries. Compared to the unified model JointNet, UniGP simultaneously estimates both depth and normals, achieving 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 comprehensively.

Quantitative evaluation We comprehensively evaluate the quantitative performance of UniGP across generation and perception tasks.
Generation Results. We show quantitative joint generation results in Tab. 1 and a comprehensive evaluation across diverse controllable generation tasks in Tab. 2. **1)** In terms of joint-generation, UniGP outperforms the baseline models across all metrics by a large margin, demonstrating our method's superiority in preserving the backbone's generative capabilities while maintaining high geometry consistency. **2)** In controllable generation, UniGP consistently achieves the leading FID and CLIP scores across dense geometric modalities (*e.g.* depth, normal) and sparse structural modalities (*e.g.* sketch, pose). Notably, it achieves the lowest RMSE between generated images and the given conditions in depth-to-image and normal-to-image tasks, demonstrating rigorous spatial and structural alignment.

Table 1: Comparison on joint generation. Bold and underline denote the best and second-best results, respectively.

Methods	**Training Data**	**Joint Generation Results**			
		FID ↓	CLIP ↑	GD ↓	GN ↓
LDM3D	-	30.19	26.02	18.13	-
JointNet	2.56M	<u>28.02</u>	<u>27.00</u>	<u>16.80</u>	-
UniGP (Ours)	1M	**21.05**	**31.50**	**6.05**	**6.41**

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

Fig. 6: Ablation study on the relationship between generation and perception. Comparison areas are highlighted with purple boxes. Adding perception training makes the generation results strictly align with the condition boundaries. Adding generation training improves the perception results with finer details.

Perception Results. 1) We present depth estimation results in Tab. 3, where UniGP achieves the best performance among all generative baselines and is on par with the SOTA regression DepthAnything series on indoor/outdoor datasets, and outperforms them on diverse datasets (*e.g.* ETH3D). This highlights the power of generative priors. Unlike regression methods trained discriminatively from scratch on 62.6M images, UniGP trains perception on only 0.059M images. By retaining the rich world knowledge of the SD3 backbone, it achieves superior zero-shot generalization. Notably, JointNet's performance lags far behind, ranking 10th. **2)** Surface normal estimation results in Tab. 4 show that UniGP achieves comparable performance to DSINE, a recent SOTA regression model, and surpasses all generative methods.

Generalization to Diverse Tasks. Beyond the core tasks evaluated above, a fundamental advantage of embedding perception into a generative framework is the inherent versatility of the generative latent space. As detailed in the supplementary material, UniGP can be easily extended to a broader range of perception tasks, such as semantic segmentation, albedo estimation, and shading estimation. This demonstrates that our DUGP branch effectively captures a generalized representation of physical world properties.

Table 2: Comparison of UniGP with task-specific baselines on controllable generation tasks. We evaluate across five spatial conditions. * denotes single-control methods, † denotes multi-control methods, and ‡ denotes joint-generation methods.

Condition	Metric	ControlNet*	SD3-CNet*	UniControlNet†	CNetPlus†	JointNet‡	UniCon‡	UniGP (Ours)‡
Training Data		-	-	10M	10M	2.56M	16K	1M
Depth	FID ↓	19.80	<u>18.00</u>	20.09	19.27	25.66	19.21	**17.95**
	CLIP ↑	25.30	27.09	25.25	<u>27.99</u>	25.09	25.27	**28.41**
	RMSE ↓	13.86	<u>12.50</u>	15.93	13.20	14.88	12.91	**6.89**
Normal	FID ↓	22.18	-	-	<u>20.11</u>	-	21.35	**18.60**
	CLIP ↑	25.10	-	-	<u>26.80</u>	-	25.12	**28.45**
	RMSE ↓	18.15	-	-	<u>16.00</u>	-	16.55	**7.12**
Canny	FID ↓	16.16	17.92	17.79	<u>16.12</u>	-	17.84	**15.91**
	CLIP ↑	25.34	26.88	25.39	<u>27.90</u>	-	25.31	**30.35**
Sketch	FID ↓	<u>17.35</u>	19.12	18.55	17.44	-	18.62	**17.02**
	CLIP ↑	25.45	27.20	25.52	<u>28.15</u>	-	25.18	**30.01**
Pose	FID ↓	23.42	20.18	22.90	<u>18.95</u>	-	21.55	**15.40**
	CLIP ↑	25.25	27.45	25.38	<u>28.60</u>	-	25.22	**30.79**

Table 3: Quantitative comparison on zero-shot affine-invariant depth estimation. We compare UniGP with both regression and generative SOTA methods. Underline marks the overall best, and **Bold** marks the best among generative methods. Note: Generative baselines utilize different backbones.

Method	Training	NYUv2 (Indoor)		KITTI (Outdoor)		ETH3D (Various)		ScanNet (Indoor)		Avg.
	Data	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	Rank
Regression Methods										
MiDaS	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	9.88
Omnidata	12.2M	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	6.50
DPT	1.4M	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	6.50
DepthAnything	62.6M	4.3	98.1	7.6	94.7	12.7	88.2	4.3	98.1	2.25
DepthAnything V2	62.6M	4.5	97.9	7.4	94.6	13.1	86.5	4.2	97.8	2.88
Generative Methods										
GeoWizard	280K	5.6	96.3	14.4	82.0	6.6	95.8	6.4	95.0	4.81
Marigold	74K	5.5	96.4	9.9	91.6	6.5	95.9	6.4	95.2	3.56
JointNet	2.56M	13.6	84.1	29.9	59.6	19.2	78.7	11.9	84.8	10.00
UniCon	16K	7.9	93.9	-	-	-	-	9.2	91.9	7.50
OneDiffusion	75M	8.9	92.0	-	-	-	-	9.7	90.7	8.88
JoDi	290K	8.3	92.0	-	-	-	-	9.9	90.3	9.13
UniGP (Ours)	59K	**5.2**	**96.6**	**8.3**	**93.3**	**6.0**	**96.3**	**5.5**	**97.9**	**2.38**

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: generation on COCO-5K and perception on NYUv2.

Relation between Generation and Perception. In Tab. 5, we analyze the impact of isolating perception from generation, reducing the training data scale, and omitting our task-specific training strategy. The results demonstrate clear and strong **complementary benefits** between the two domains.

1) Quantitative Improvements: Incorporating perception training vastly improves the structural alignment of controllable generation, evidenced by the steep drop in depth-to-image RMSE from 10.73 down to 6.89. Conversely, incorporating massive generation training data imbues the perception layers with richer generative priors, pushing the limits of zero-shot perception metrics such as AbsRel and mean angular error by capturing finer and generalized details. Furthermore, omitting our task-specific batching strategy or halving the data scale leads to noticeable degradation across both domains, validating the necessity of our unified dataset construction.

2) Deep Dive into Cross-Task Synergy: To intuitively understand these quantitative gains, we provide a qualitative analysis in Fig. 6. When the model is trained exclusively on generation tasks without perception training, the generated pixels often bleed across the conditional spatial constraints. As highlighted by the purple boxes in Fig. 6(a), the model fails to strictly distinguish the foreground hand from the surrounding grass defined in the depth condition. However, once perception training is introduced, the model develops explicit spatial consciousness, forcing the generation results to strictly align with the condition boundaries. Conversely, as shown in Fig. 6(b), pure perception training

Table 4: Quantitative comparison on zero-shot surface normal estimation. Note: Generative baselines utilize different backbones.

Method	Training Data	NYUv2 (Indoor)			ScanNet (Indoor)			iBims-1 (Indoor)			Sintel (Outdoor)			Avg.
		m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	m.↓	11.25°↑	30°↑	Rank
Regression Methods														
Omnidata	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	6.92
Omnidata V2	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	3.50
DSINE	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	1.67
Generative Methods														
Marigold	74K	20.9	50.5	-	21.3	45.6	-	18.5	64.7	-	-	-	-	7.40
GeoWizard	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	5.29
StableNormal	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	3.63
JoDi	290K	18.6	-	-	20.3	-	-	18.2	-	-	-	-	-	4.83
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**	**1.58**

Table 5: Impact of Joint Training and Dataset Strategy. Evaluating the complementary benefits of cross-task training on representative benchmarks.

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↓	δ1↑	m.↓	11.25°↑
w/o Perception Train.	21.07	31.57	-	-	17.99	28.33	10.73	18.85	28.19	11.74	-	-	-	-
50% 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 Train.	-	-	-	-	-	-	-	-	-	-	5.5	96.5	17.1	56.0
50% 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
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**

tends to produce overly smoothed depth maps. The addition of generation training injects robust and high-frequency texture priors into the perception branch, enabling the model to capture realistic micro-structures, such as the intricate architectural details of the building facade, that are otherwise completely lost in standard regression models.

Table 6: Balance between Stacked Control and Perception Layers. We evaluate different block allocation ratios (m, n) where $m + n = 24$.

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↓	δ1↑	m.↓	11.25°↑
(m, n) = *(0,24)*	-	-	-	-	-	-	-	-	-	-	5.4	96.3	17.2	58.0
(m, n) = *(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
(m, n) = *(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
(m, n) = *(24,0)*	21.78	31.37	7.56	7.95	19.63	28.18	10.65	18.91	28.21	10.49	-	-	-	-
UniGP (12,12)	**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**

Balance between Stacked Control and Perception Layers. We denote the number of stacked control/perception layers as (m, n), with the default setting m=n=12 for the 24-layer SD3 backbone. As shown in Tab. 6, the optimal joint performance is consistently achieved when the capacities are balanced. Specifically, allocating too few control layers (*e.g.*, $m = 6$) severely degrades the controllable generation quality, as the network lacks the required depth to effectively inject and process spatial conditions. Conversely, allocating too few perception lay-

ers (*e.g.*, $n = 6$) significantly harms the dense prediction accuracy, indicating that extracting fine-grained geometric representations demands sufficient feature transformations. This insight implies that an equal capacity allocation allows the modules to co-develop a robust shared representation space, thereby maximizing cross-task synergy.

Table 7: Model Design Choices. Comparing DUGP against alternative unified network paradigms, all implemented on the exact same SD3-Medium backbone.

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↓	$\delta1$↑	m.↓	11.25°↑
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
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**

Model Design Choices. As shown in Tab. 7 and conceptualized in Fig. 3, we explored alternative designs on the SD3 backbone: duplicating the entire backbone (JointNet-style) and fully fine-tuning the backbone itself (Marigold-style). Importantly, even when utilizing the same powerful SD3 backbone, both alternative architectures yield significantly inferior performance compared to DUGP. This confirms that UniGP's superior performance stems primarily from our disentangled gradient flow and architectural design, rather than merely from scaling the backbone model.

5 Conclusion and Limitations

In this work, we present UniGP, an MMDiT-based framework for unified generation and perception tasks. UniGP comprises DUGP and a unified dataset training strategy. The former uses only a copied image branch of MMDiT to model dense distributions beyond RGB, minimizing computational bloat, while the latter combines heterogeneous datasets into a unified framework to jointly optimize tasks efficiently. UniGP demonstrates outstanding performance across evaluations, surpassing existing unified models and performing on par with specialized regression expert models. Crucially, our experiments reveal that generative priors and perceptual constraints provide complementary benefits, bridging the separated fields of diffusion-based synthesis and dense geometry prediction.

Limitations. While UniGP provides a highly versatile framework, it inevitably introduces additional computational overhead. Specifically, duplicating the image branch increases the parameter count from 2B to 3B, and the 20-step inference time (FP16, NVIDIA V100) increases from 5.07s to 9.54s compared to the base model. However, considering that UniGP effectively replaces multiple specialized network architectures with a single unified model capable of diverse controllable generation and dense prediction tasks, we consider this computational trade-off a reasonable and acceptable compromise for significantly expanded capabilities.

Acknowledgment

The research is supported in part by Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321, ITF PRP/046/24FX, Science & Technology Cooperation Program of Shandong under grant No. SDST26EG01, SAIL Research Project, and HKUST-Zeekr Collaborative Research Fund. This work was also supported by Damo Academy through Damo Academy Research Intern Program.

References

1. Bae, G., Davison, A.J.: Rethinking inductive biases for surface normal estimation. In: CVPR (2024)
2. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
3. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: ICCV (2021)
4. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024)
5. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: ECCV (2024)
6. Garcia, G.M., Zeid, K.A., Schmidt, C., de Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. arXiv preprint arXiv:2409.11355 (2024)
7. Github: Controlnetplus (2024), `https://github.com/xinsir6/ControlNetPlus`
8. Gui, M., Fischer, J.S., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast monocular depth estimation with flow matching. arXiv preprint arXiv:2403.13788 (2024)
9. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Liu, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 (2024)
10. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3d common corruptions and data augmentation. In: CVPR (2022)
11. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
12. Labs, B.F.: Flux (2024), `https://blackforestlabs.ai/announcing-black-forest-labs/`
13. Le, D.H., Pham, T., Lee, S., Clark, C., Kembhavi, A., Mandt, S., Krishna, R., Lu, J.: One diffusion to generate them all. In: CVPR (2025)
14. Li, X., Herrmann, C., Chan, K.C., Li, Y., Sun, D., Yang, M.H.: A simple approach to unifying diffusion-based conditional generation. arXiv preprint arxiv:2410.11439 (2024)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
16. Midjourney: Midjourney (2024), `https://www.midjourney.com`

17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
18. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. In: NeurIPS (2024)
19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI **44**(3), 1623–1637 (2020)
21. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
23. Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Muller, M., et al.: Ldm3d: Latent diffusion model for 3d. arXiv preprint arXiv:2305.10853 (2023)
24. Sun, Y., Liu, Y., Tang, Y., Pei, W., Chen, K.: Anycontrol: Create your artwork with versatile control on text-to-image generation. arXiv preprint arXiv:2406.18958 (2024)
25. Team, I.: Sd3-medium-controlnet (2024), `https://huggingface.co/InstantX/SD3-Controlnet-Canny`
26. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. TPAMI **44**(7), 3614–3633 (2021)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
28. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: Diffusion models trained with large data are transferable visual models. arXiv preprint arXiv:2403.06090 (2024)
29. Xu, Y., He, Z., Kan, M., Shan, S., Chen, X.: Jodi: Unification of visual generation and understanding via joint modeling. arXiv preprint arXiv:2505.19084 (2025)
30. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
31. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
32. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: Stablenormal: Reducing diffusion variance for stable and sharp normal. TOG (2024)
33. Ye, H., Xu, D.: Diffusionmtl: Learning multi-task denoising diffusion model from partially annotated data. In: CVPR (2024)
34. Zhang, J., Li, S., Lu, Y., Fang, T., McKinnon, D.N., Tsin, Y., Quan, L., Yao, Y.: Jointnet: Extending text-to-image diffusion for dense distribution modeling. In: ICLR (2024)
35. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
36. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. In: NeurIPS (2024)