ParaUni: Enhance Generation in Unified Multimodal Model with Reinforcement-driven Hierarchical Parallel Information Interaction

Jiangtong Tan[1,2], Lin Liu[1,2], Jie Huang[1], Xiaopeng Zhang[2], Qi Tian[2]*, Feng Zhao[1]*
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
[2]Huawei Inc.
{jttan, ll0825, hj0117}@mail.ustc.edu.cn, {zhangxiaopeng12, tian.qi1}@huawei.com
fzhao956@ustc.edu.cn

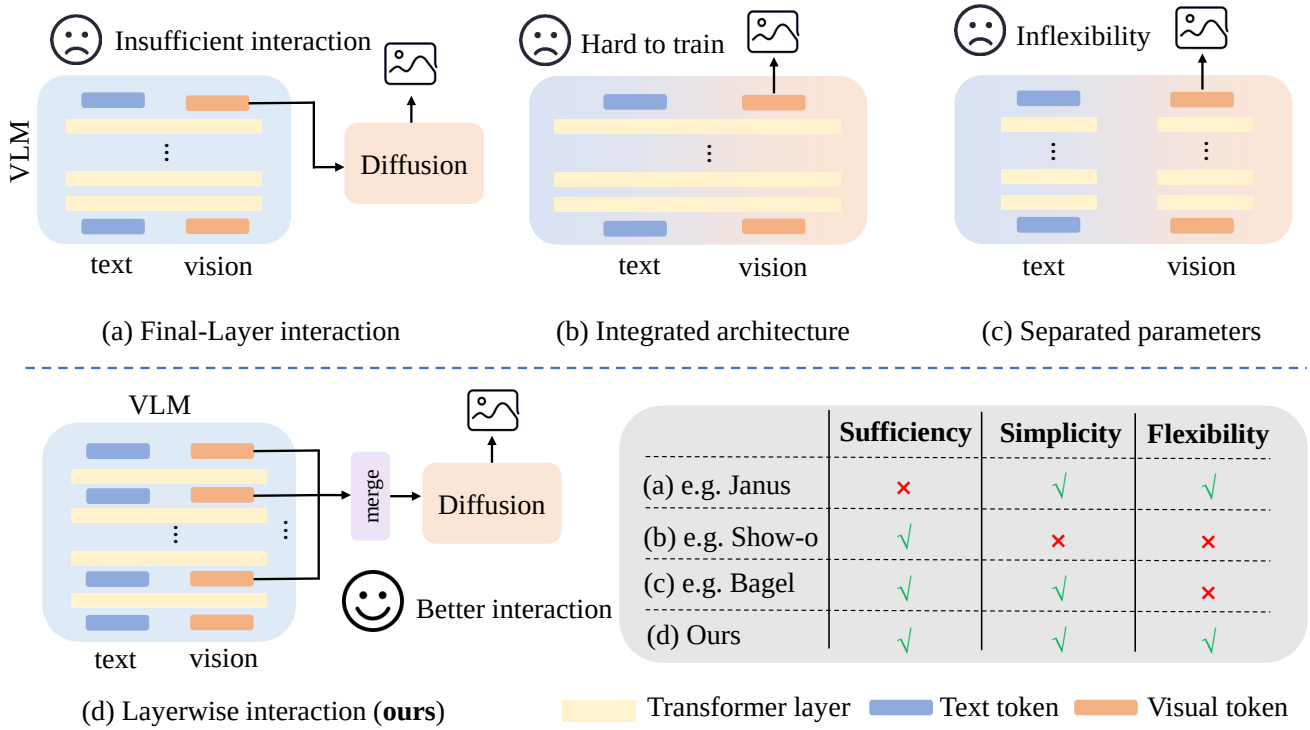


Figure 1. Illustration of different VLM and diffusion interaction methods in unified understanding and generation. (a) Final-Layer interaction, which only uses the features from the last layer of the VLM as a condition, is insufficient in capturing detailed image contents and only provides a semantic abstraction. (b) Integrated architecture, which integrates VLM and diffusion into one transformer, faces increased training difficulty due to different training objectives. (c) Separated parameters, which use two sets of parameters for understanding and generating information within one transformer, although beneficial for multimodal information interaction, faces challenges in flexibility and scalability. (d) Our layerwise interaction, which merges different levels of semantic abstraction and fine-grained details from the VLM, ensures sufficiet interaction and retains flexible separation architecture.

Abstract

Unified multimodal models significantly improve visual generation by combining vision-language models (VLMs) with diffusion models. However, existing methods struggle to fully balance sufficient interaction and flexible implementation due to vast representation difference. Considering abundant and hierarchical information in VLM's layers from low-level details to high-level semantics, we

*Corresponding author

*propose **ParaUni**. It extracts features from variants VLM's layers in a **Para**llel way for comprehensive information interaction and retains a flexible separation architecture to enhance generation in **Uni**fied multimodal model. Concretely, visual features from all VLM's layers are fed in parallel into a Layer Integration Module (LIM), which efficiently integrates fine-grained details and semantic abstractions and provides the fused representation as a condition to the diffusion model. To further enhance performance, we reveal that these hierarchical layers respond unequally to different rewards in Reinforcement Learning (RL). Crucially, we design a Layer-wise Dynamic Adjustment Mechanism (LDAM) to facilitate multiple reward improvements that aligns the hierarchical properties of these layers using RL. Extensive experiments show ParaUni leverages complementary multi-layer features to substantially improve generation quality and shows strong potential for multiple reward advances during RL stages. Code is available at https://github.com/JosephTiTan/ParaUni.*

1. Introduction

As unified multimodal understanding and generation has long been a research focus, extensive work has demonstrated strong performance of unified architectures [8, 12, 29, 53, 57, 59, 61], improving tasks such as image generation and editing. However, due to the architectural gap between autoregressive (AR)-based vision-language models (VLMs) for image understanding and diffusion-based models for image generation, unifying them in a single framework is highly challenging.

Several prior works [6, 39, 66] attempt to bridge this gap by supplying the final layer representations of a VLM as conditioning to a diffusion decoder, as shown in Fig. 1(a). However, reliance on insufficient information interaction from only a single layer restricts the fidelity and granularity of information transferred from the understanding module to the generation module, thus constraining potential gains in generation quality, which has been shown in [31]. Alternative approaches[61, 67, 68] seek to integrate the diffusion denoising process with the AR process within a single transformer to facilitate interaction, as shown in Fig. 1(b). Since the two procedures are governed by disparate optimization objectives, this integration substantially increases training complexity and precludes straightforward reuse of off-the-shelf pretrained models. Some works [51] use two separate transformer experts to handle understanding and generation, while tokens from different modalities interact via shared multimodal self-attention inside each transformer block, as shown in Fig. 1(c). This design enables richer cross-modal information interaction and reduce the difficulty of multi-task learning [12]. However, it is limited in flexibility and scalability and incurs high inference latency due to the tight coupling of the two type parameters [7].

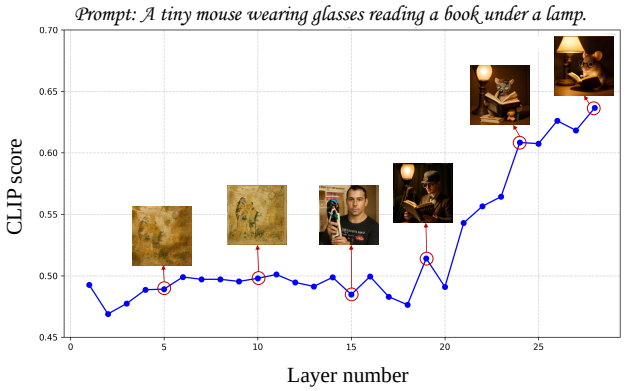


Figure 2. Illustration of CLIP score changes with 28 layer features from shallow to deep in VLM. We conduct experiments on 500 prompts and find that as the number of VLM layers increased, the images generated by the diffusion model gradually shifted from focusing on detailed textures to enhancing semantic information. The CLIP score, which measures the alignment between images and text, also increased with the number of layers. It indicates that different depths of the VLM's transformer layers encode information ranging from low-level details to high-level semantics.

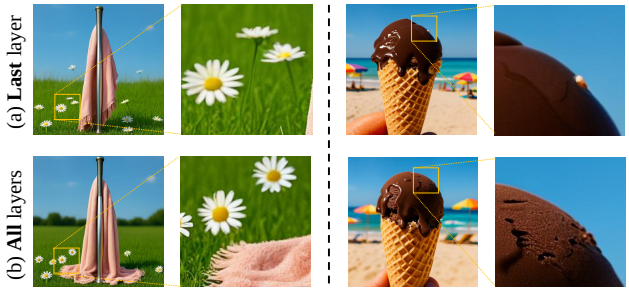
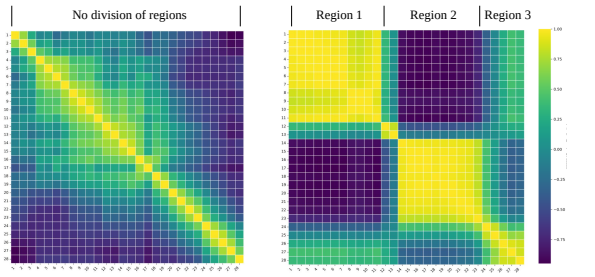


Figure 3. Comparison between (a) using only the last layer of the VLM and (b) using all layers. After using all layers, the generated images have more details, indicating that the diffusion model integrates more detailed information from the VLM, which confirms the rationality of using all layers as conditions.

Prior works[14, 15, 20, 24] have established that VLM's transformer layers at different depths encode varying levels of information from low-level details to high-level semantics. We argue that aggregating these features as conditions for generation with a diffusion model can effectively improve comprehensive interaction between them. To prove it, we extract the visual tokens separately from **each** layer of the VLM and use them as conditioning inputs to the diffusion model. We evaluate CLIP scores of generated images to reflect their semantic alignment[25, 44]. As shown in Fig. 2, generated images conditioned on shallow layers tend to exhibit texture-like characteristics. As layer depth increases, the generated images progressively exhibit stronger semantic content and their content becomes more consis-



(a) Layer similarity of **last** layer interaction (b) Layer similarity of **all** layers interaction

Figure 4. Layer similarity of the last layer or all layer interaction. For last layer interaction, the similarity between layers is relatively low, while for all layer interaction, cluster-like phenomena occur between layers. From [31], the phenomena is not limited to our base model with universality. We analyze the properties of each clustered region in Fig. 5.

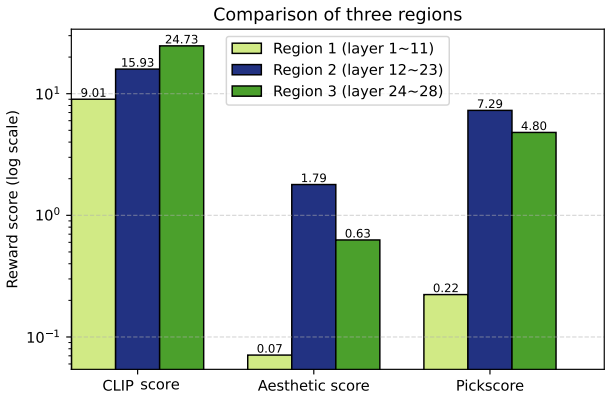


Figure 5. Comparison of the response degrees of reward scores to different regions. CLIP score is quite sensitive to changes in all three regions, but is most significantly affected by the deep layer. Aesthetic score and Pickscore are most sensitive to changes in the middle layers, but have little impact on the shallow layer. Therefore, we can influence the corresponding reward score by perturbing specific layers.

tent with the text prompt. From this analysis we conclude that VLM layers exhibit different properties from shallow to deep, so integrating all layers naturally provide more comprehensive information for the generation process. Therefore, we aggregate the visual tokens from **all** layers and feed the combined condition into the diffusion model during training. As illustrated in Fig. 3, incorporating multi-layer information into the conditioning yields images with richer, more realistic details and further improves performance.

With all layers used as conditions, we then dive into the relations and properties of individual VLM layers. We find adjacent layers exhibit higher similarity, and several regions with distinct intrinsic properties emerge, which aligns with different reward scores. **Relations**: we first extract the visual tokens from each layer after the above training and analyze their relationships using cosine similarity, as shown in Fig. 4. Similar adjacent phenomena have been observed in prior work[31] but without analysis of the properties of each layer for reward scores. **Properties**: we therefore empirically remove layers corresponding to specific regions and measure the impact on three reward scores, as shown in Fig. 5. We find that reward scores align the hierarchical properties of these layer regions: middle-layer regions align aesthetics (Aesthetic score[21]) and human preferences (Pickscore[23]) metrics, while deep-layer regions align semantic (CLIP score[44]) metrics. Consequently, within a multi-layer conditioning framework, it is possible to design an optimization strategy to facilitate layer-wise reward alignment, thereby enhancing generative performance during the Reinforcement Learning (RL) stage.

In this paper, we propose a unified framework, dubbed **ParaUni**, that processes features from every layer of a VLM in parallel to bridge and enhance generation tasks with layer-wise guided RL. Concretely, as shown in Fig. 6, our method employs a learnable-query–based unified architecture, learnable queries extract visual tokens from each layer, which are then passed into the cross-attention in diffusion for the generation task. Our framework consists of two key components: (1) **Layer Integration Module (LIM)**. We use a shared Transformer module to encode features from each layer in parallel, and then apply layer normalization to further align the different layer features. The processed features are finally fed into the diffusion model for generation. (2) **Layer-wise Dynamic Adjustment Mechanism (LDAM)**. During the RL stage when a specific reward is optimized, we perturb a specific layer's weights according to changes in corresponding reward in Fig. 5 and gradient norm to promote their improvement and stabilize training. Extensive experiments show that ParaUni effectively improves generation, enhancing image detail and quality. Moreover, during the RL phase, it significantly boosts various reward metrics, demonstrating strong performance.

In summary, our contributions are as follows:

- We reveal that VLM layers encode properties from low-level detail to high-level semantics, and that integrating visual tokens from these layers as conditions for diffusion substantially improves generation quality.
- To make the best use of this characteristic, we design the LIM for parallel processing of VLM layers and introduce the LDAM with RL to further enhance performance.
- Extensive experiments demonstrate that our method outperforms baselines and achieves state-of-the-art results. Our method enhances the details and quality of the generated images, and at the same time promotes the improvement of multiple rewards in the RL stage.

2. Related Works

2.1. Unified Image Understanding and Generation

Unified multimodal models [8, 12, 29, 53, 57, 59, 61] aim to integrate visual understanding and generation within a single framework. Previous work uses autoregressive VLMs [9–11, 26, 34, 35, 64, 69] for image understanding and diffusion models [1, 3, 5, 43, 45–47, 71] for image generation, creating an architectural gap that hinders unification. Some studies[39, 40, 55, 57] unify VLMs and diffusion models by feeding VLM's last hidden-layer features as conditioning to diffusion decoders. While leveraging both understanding and generation capabilities, simple concatenation limits information transfer. Other studies[61, 67, 68] integrate diffusion's noise prediction into the autoregressive process, avoiding concatenation complexity and enabling information exchange. However, this causes mutual interference between tasks, reducing flexibility. Alternative paradigms[12, 32, 51] use separate parameters for each task to reduce conflicts, but require substantial training resources. Some work using multi-layer VLM features[2, 31] also fail to leverage prior knowledge effectively, achieving suboptimal results.

2.2. RL algorithms for Image Generation

Optimizing image generation models using reinforcement learning (RL) to align with human preferences has always been a key focus in generative tasks. [62] is the first work that optimizes diffusion models by using reward model scores. Diffusion-DPO [54, 65] introduces DPO into diffusion models, further improving the quality of generated images. With significant breakthroughs achieved by group relative policy optimization (GRPO) in large language models, RL for generation has gradually made rapid progress. [37] and [63] introduced randomness by reformulating the original ODE as an SDE, thereby incorporating GRPO into flow matching models. In addition, due to the instability of GRPO, some works[30, 56] have proposed improvements, which further enhances the effectiveness of reinforcement learning.

3. Method

3.1. Preliminaries

Diffusion Models. Diffusion models are a class of generative models that capture data distributions by inverting a forward corruption process, which is defined as:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \quad t \sim \mathcal{U}(0,1),\ \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where α_t and σ_t are monotonically decreasing and increasing functions of t, respectively. To reverse this process, a denoising network $\mathcal{F}_\theta$ is trained to predict the target y at each time step t, conditioned on c such as a class label or text prompt. In our setting, the condition c is obtained from a VLM. This can be formulated as:

$$\mathcal{L} = \mathbb{E}_{x_0,c,\epsilon,t}[\|y - \mathcal{F}_\theta(x_t, c, t)\|_2^2], \tag{2}$$

where the training target y can be either the Gaussian noise ϵ in DDPM or the vector field $\epsilon - x_0$ in flow model.

RL for diffusion model. Following the Flow-GRPO[38], the flow model generates a set of G candidate images $\{x_0^i\}_{i=1}^G$ along with trajectories $\{x_T^i, x_{T-1}^i, \ldots, x_0^i\}_{i=1}^G$ by reformulating trajectory sampling as a stochastic differential equation (SDE), injecting randomness into the process:

$$dx_t = \left(v_t + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_t)\right)dt + \sigma_t dw_t, \tag{3}$$

which solves the problem of original deterministic ODE formulation's inability to generate diverse samples.

Algorithm 1 Layer-wise Dynamic Adjustment Mechanism

Require: *Generation model* $\mathcal{G}$, *Reward model* $\mathcal{R}$.
Initial: $g_{n-1}, r_{n-1}, n_{cool}, r_s \leftarrow inf, 0, 0, 0$
for $n \leftarrow 1$ **to** *epochs* **do**
  $g_n \leftarrow$ *Generation model* $\mathcal{G}$
  $r_n \leftarrow$ *Reward model* $\mathcal{R}$
  if $g_n \geq g_{n-1} * 10^2$ **then**:
    $g_s \leftarrow True$ ▷ Gradient norm guidance.
  end if
  if $r_n \leq r_{n-1}$ **then**:
    $r_s \leftarrow r_s + 1$ ▷ Reward guidance.
  end if
  if $n_{cold} = 0$ **then**:
    if g_s and $r_s \geq 5$ **then**:
      $\epsilon \sim \mathcal{N}(0, I)$
      $i \leftarrow$Selected layer number ▷ In Sec. 4
      $\gamma \leftarrow \gamma(g, r)$ ▷ In supplementary material
      $c_i \leftarrow c_i(1 + \gamma\epsilon)$ ▷ Core operation.
      $r_s \leftarrow 0$ ▷ Restore to the original state.
      $g_s \leftarrow False$
      $n_{cool} \leftarrow n$ ▷ Cooling-off period begin.
    end if
  else
    $n_{cool} \leftarrow n_{cool} - 1$
  end if
  $g_{n-1} \leftarrow g_n$ ▷ Prepare for the next round.
  $r_{n-1} \leftarrow r_n$
end for

3.2. ParaUni

In this section we introduce ParaUni, a unified architecture that simultaneously processes every layer of a VLM and feeds the integrated features into a diffusion model, improving generation quality for unified multimodal model.

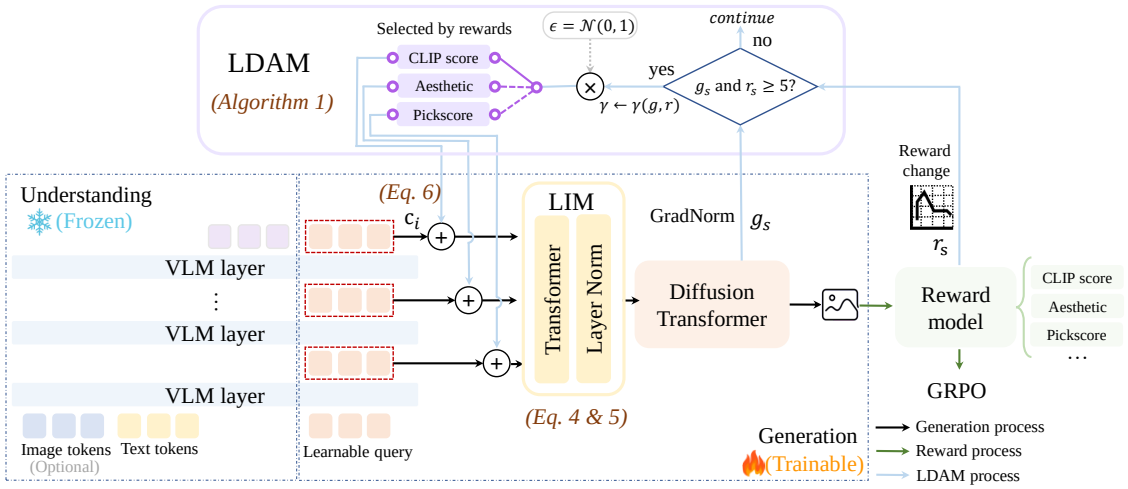


Figure 6. Overview of ParaUni. For understanding, after inputting image tokens and text tokens, the VLM performs understanding through an autoregressive process. For generation , the context information is compressed through a learnable query, and then the learnable queries from all layers are fed into a Layer Integration Module (LIM), which consists of a Transformer module and a Layer Norm layer. The integrated features are then sent into the cross attention in diffusion for generation. In the RL phase, we designed a Layer-wise Dynamic Adjustment Mechanism (LDAM), which adds Gaussian noise perturbation to specific layers through reward and GradNorm guidance. And different layers enable perturbations under different rewards, as detailed in Algorithm 1.

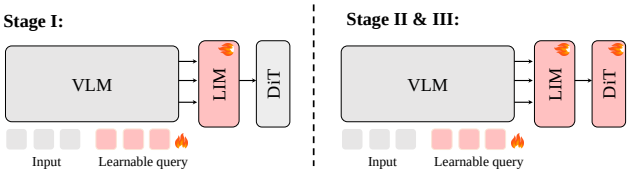


Figure 7. Illustration of the three training stages. Stage I is aligning VLM and diffusion, Stage II is fine-tuning with high-quality data, and Stage III uses RL to further enhance performance.

Overview. Our framework follows the MetaQuery[42] design and comprises N learnable queries, a VLM, and a diffusion model. The VLM's visual understanding capabilities remain unchanged, since its parameters are frozen. During image generation, the learnable queries extract textual information—or optionally visual information—during the VLM's forward pass, and then pass that information to the diffusion model. In ParaUni, we feed the learnable queries from all layers as conditioning into a Layer Integration Module (LIM) to consolidate information across layers; the integrated representation is finally supplied to the diffusion model. During the RL stage, we propose a Layer-wise Dynamic Adjustment Mechanism (LDAM) that applies different optimization strategies to different layers according to different rewards, enabling improvement on multiple reward signals. In the sections below, we describe in detail how the LIM enhances generation quality and how the LDAM is designed for the RL stage.

Layer Integration Module. Our early exploratory experiments reveal that the VLM's layers encode rich, hierarchical representations. Therefore, instead of following the common practice of using only the final-layer features as conditioning, we introduce information from all layers into the diffusion model. Suppose we extract the learnable query q_i from each layer i of the VLM:

$$c_i = LN(f_\theta(q_i)), i \in [0, L], \tag{4}$$

$$c = \frac{1}{n}(\sum_{i=1}^{n}(c_i)), \tag{5}$$

where L denotes the max layer number, f_θ denotes a Transformer module and LN denotes a LayerNorm module. We use c as the condition passed to DiT's cross-attention, which produces final high-quality image after denoising iterations.

Layer-wise Dynamic Adjustment Mechanism. Because different VLM layers respond differently to different rewards, we apply dynamic adjustments to specific layers during optimization of each reward. However, large adjustments can make training unstable. Inspired by prior work on inference-time scaling[49], which injects perturbations into text conditioning to improve generation quality and diversity, we perturb the layers that exhibit strong responses for a given reward when optimizing that reward. Concretely, we monitor each reward's training signal in real time. If a reward shows a sustained decline, we use it as a guidance and inject noise sampled from a Gaussian distribution into the corresponding layers to encourage the model to explore a broader set of potential optima, expressed as follows:

$$c_i = c_i(1 + \gamma\epsilon), \epsilon \sim \mathcal{N}(0, I), \tag{6}$$

Table 1. Quantitative Results from the GenEval benchmark for text-to-image generation.

Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	LlamaGen [52]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [48]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [48]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α [4]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [48]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [46]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [55]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [43]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [1]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [16]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Unified	Chameleon [53]	-	-	-	-	-	-	0.39
	SEED-X [17]	0.97	0.58	0.26	0.80	0.19	0.14	0.51
	LMFusion [51]	-	-	-	-	-	-	0.63
	Show-o [61]	0.95	0.52	0.49	0.82	0.11	0.28	0.68
	TokenFlow-XL [36]	0.95	0.60	0.41	0.81	0.16	0.24	0.63
	Janus [57]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	MetaQuery [42]	-	-	-	-	-	-	0.74
	BLIP3-o-4B [6]	-	-	-	-	-	-	0.81
	BAGEL [12]	0.99	0.94	**0.81**	0.88	0.64	0.63	0.82
	OpenUni[58]	0.99	0.92	0.76	0.91	0.82	**0.77**	0.86
	ParaUni	**0.99**	**0.94**	0.78	**0.91**	**0.83**	0.76	**0.87**

Figure 8. Samples generated by ParaUni.

where γ is a scale factor to control the degree of perturbations, defined in supplementary material. Additionally, considering GradNorm as an indicator of training stability, we use it as a guidance for whether to apply the perturbation strategy. When GradNorm exhibits a large spike, we apply perturbations to the corresponding layer to encourage the model to escape the current unstable state. In order to ensure stable training, we have set a cooling-off period, which gradually increases with the number of training iterations. During training, we first perform RL for one reward using the method described above. We then use the resulting checkpoint as the starting point for RL on the next reward, while preserving the corresponding layer perturbations. We repeat this process progressively to perform RL across multiple rewards. Finally, the detailed description of Layer-wise Dynamic Adjustment Mechanism is expressed as Algorithm 1.

3.3. Training Recipe

Our ParaUni training consists of three stages. First, we align the VLM and diffusion models, then fine-tune using high-quality data, and finally conduct RL, as shown in Fig. 7.

Stage I. Aligning the VLM and diffusion models. We follow the setup of OpenUni, in which the VLM is based on Intern-VL3-2B[70] and the diffusion part is based on SANA-1.5-1.6B1024px[60]. The training data includes text-to-image-2M[19], LAION-Aesthetic-6M[50], Megalith-10M[41], RedCaps-5M[13]. In this stage, only the LIM module and learnable query are trained, while the other modules are frozen.

Stage II. We fine-tune the model using the high-quality data from BLIP3-o-60k[6]. In this stage, we train the parameters of the learnable query, the LIM module and the diffusion module. After this training stage, the model's performance has surpassed that of models trained only on the last layer with the same configuration.

Stage III. We conduct reinforcement learning train-

Table 2. Quantitative Results from the DPG-Bench for text-to-image generation.

Type	Method	Global	Entity	Attribute	Relation	Other	Overall↑
Gen. Only	SDv1.5 [48]	74.63	74.23	75.39	73.49	67.81	63.18
	PixArt-α [4]	74.97	79.32	78.60	82.57	76.96	71.11
	Lumina-Next [71]	82.82	88.65	86.44	80.53	81.82	74.63
	SDXL [43]	83.27	82.43	80.91	86.76	80.41	74.65
	Playground v2.5 [27]	83.06	82.59	81.20	84.08	83.50	75.47
	Hunyuan-DiT [33]	84.59	80.59	88.01	74.36	86.41	78.87
	PixArt-Σ [5]	86.89	82.89	88.94	86.59	87.68	80.54
	Emu3-Gen [55]	85.21	86.68	86.84	90.22	83.15	80.60
Unified	Show-o [61]	-	-	-	-	-	67.27
	Janus [57]	82.33	87.38	87.70	85.46	86.41	79.68
	Janus-Pro-1B [8]	87.58	88.63	88.17	88.98	88.30	82.63
	MetaQuery [42]	-	-	-	-	-	80.04
	BLIP3-o-4B [6]	-	-	-	-	-	79.36
	OpenUni [58]	87.01	**90.02**	**89.63**	90.28	**88.62**	83.08
	ParaUni	**90.01**	89.31	89.18	**91.85**	85.61	**83.45**

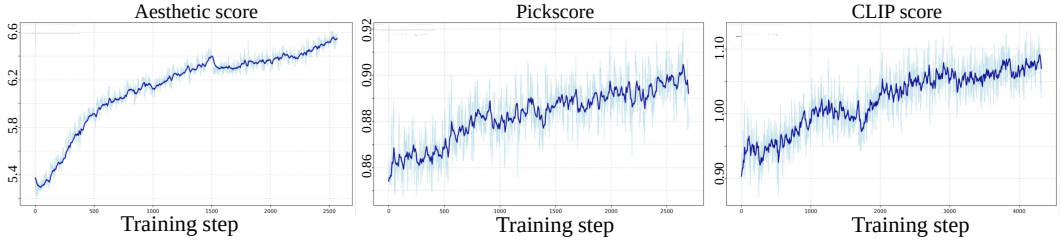


Figure 9. Reward change curves during the training process. As the training progresses, all reward scores gradually increase, demonstrating the effectiveness of our method.

ing using multiple rewards following the FlowGRPO[38] framework. The rewards include Aesthetic score, Pickscore and CLIP score. We first train using the Aesthetic score and Pickscore sequentially, then retain the weights of the corresponding layers, and finally train using the CLIP score. In this stage, the trainable parameters of the model is the same as Stage II. More configuration will be detailed in the supplementary materials.

Novelty Distinctions. Considering prior work that also conditions diffusion models on all VLM layers, we now detail how ParaUni differs from those approaches. [28] and [2] perform one-to-one interactions between each VLM layer and each diffusion layer. Although this design leverages information from all VLM layers, it requires tight coupling between the VLM and the diffusion model, which reduces flexibility. [31] is similar to our method in that it integrates multi-layer VLM features, but it does not analyze the distinct properties of those layers nor exploit those properties to accelerate or guide the RL process. In contrast, ParaUni explicitly studies layer-wise characteristics and uses that analysis to drive a Layer-wise Dynamic Adjustment Mechanism , allowing targeted perturbations and optimization strategies for different layers during RL while maintaining modularity and flexibility between VLM and diffusion model.

4. Experiments

4.1. Implementation Details

We train all three stages using the flow matching, which implements iterative denoising by predicting the velocity field. All models are optimized with the AdamW optimizer at a learning rate of 1e-4. The batch size is 512 and weight decay was set to 0.05. For the diffusion model we adopt a cosine schedule. The VLM employs 256 learnable queries and 28 transformer layers. All experiments are conducted on NVIDIA A800 GPUs. In Algorithm 1, for CLIP score, $i \in [24, 28]$; for Pickscore and Aesthetic score, $i \in [12, 23]$, which align with Fig. 5.

Table 3. Ablation from the GenEval for text-to-image generation.

Type	Single Obj.	Colors	Position	Overall↑
(1)	0.98	0.88	0.75	0.82
(2)	0.99	0.90	0.81	0.85
(3)	0.99	0.90	0.82	0.84
(4)	**1.00**	0.90	0.81	0.86
(5)	0.98	0.61	0.75	0.73
(6)	0.98	0.90	0.82	0.86
(7)	0.98	0.90	0.81	0.86
Ours	0.99	**0.91**	**0.83**	**0.87**

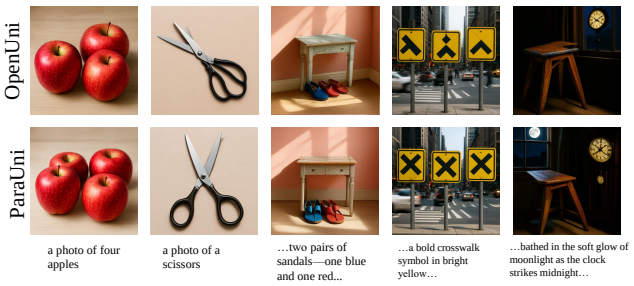


Figure 10. Qualitative comparison.

4.2. Evaluation Benchmarks

Because we do not train the VLM's parameters, ParaUni's understanding ability derives from the initial model InternVL3. The results for InternVL3's understanding performance are provided in the supplementary material. To evaluate ParaUni's generative capabilities, we primarily focus on the model's adherence to prompts and the degree of semantic alignment. Specifically, we employ **GenEval**[18], which assesses a model's ability to follow complex prompts and concentrates on generating images with correct object attributes, counts, positions, and colors. Results are reported across categories including single-object, multi-object, counting, color, and spatial position. We also use **DPG-Bench**[22] to examine the models' complex semantic alignment for text-to-image generation under verbose and densely specified prompts. Finally, we assess the effectiveness of the RL stage by statistically evaluating reward scores, which include CLIP score for semantic alignment, Aesthetic score for image quality, and Pickscore reflecting human preference. Additionally, to evaluate ParaUni on image-to-image tasks, we report reconstruction performance using PSNR and SSIM on low level details, as well as CLIP-based image similarity on high level semantics in supplementary material.

4.3. Comparison with the Baseline

For baselines, we select purely generative models, including both diffusion- and autoregressive-based methods such as LlamaGen, SDv1.5, and DALL·E 2, as well as recent state-of-the-art unified understanding-and-generation models, such as Chameleon, Show-o, BLIP3-o, and BAGEL.

Tab. 1 compares our method against these baselines. As shown, our approach achieves a GenEval score of 0.87 and a DPG-Bench score of 83.45, outperforming the unified understanding-and-generation baselines and substantially exceeding the purely generative models. This result indicates that unified understanding and generation architectures can markedly raise the performance ceiling of generative models relative to purely generative approaches. Furthermore, because we condition the generator on features from all VLM layers, we obtain an additional performance improvement over methods that use only a single layer as the conditioning signal.

Fig. 10 presents quantitative examples of outputs produced by our method. It illustrates that our approach effectively enhances both fine-grained visual detail and semantic alignment, yielding higher visual quality than models that condition on a single VLM layer such as [58].

4.4. Performance of RL

Fig. 9 the effects of our method during the RL stage. The figure shows that each reward consistently increases over the course of training under our method, demonstrating robust generalization. Additionally, we provide quantitative results before and after RL and a comparison without our method in the supplementary materials.

4.5. Ablation

We conduct extensive ablation studies on the various modules of ParaUni. The results are shown in Tab. 3.

Ablation on the Impact of Different Layer Selection Strategies on Generation. We first evaluate the effect of conditioning on fewer visual-token layers, including (1) removing shallow, (2) removing middle, or (3) removing deep subsets, as well as (4) an interleaved scheme that uses every other layer. Tab. 3 shows that omitting subsets of visual-token layers leads to degraded performance across metrics. Consequently, ParaUni conditions on **all** VLM layers.

Ablations on the LIM. We test alternative LIM configurations, including (5) removing LayerNorm module. Since removing the Transformer module would cause a significant drop in performance, it is not included in the ablation study. In each ablated variant in Tab. 3, performance declines relative to the full design, indicating that the components incorporated into our LIM each contribute positively to overall performance.

Ablations on the LDAM. We replace our proposed LDAM with alternative strategies, including (6) removing GradNorm-based guidance, (7) removing the reward degradation guidance. Removing either GradNorm guidance or the reward-degradation guidance led to worse training out-

comes. Tab. 3 shows that the LDAM design choices are important and effectively improve RL performance.

5. Conclusion

In this paper, we propose ParaUni, a unified understanding-and-generation framework that conditions diffusion-based generation in parallel on information from all layers of a vision-language model. By leveraging multi-level representations across every VLM layer, we introduce a Layer Integration Module to aggregate these signals and thereby improve overall generation quality. We further devise a Layer-wise Dynamic Adjustment Mechanism to enhance multiple reward objectives during the reinforcement learning stage. Extensive experiments demonstrate that our method achieves state-of-the-art performance and exhibits strong practical potential.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 4, 6

[2] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 4, 7

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 4

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 6, 7

[5] 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. 4, 7

[6] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 2, 6, 7

[7] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025. 2

[8] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 2, 4, 7

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 4

[12] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 4, 6

[13] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image–text data created by the people, for the people. In *NeurIPS, Datasets and Benchmarks Track*, 2021. 6

[14] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*, 2020. 2

[15] Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. Discovering salient neurons in deep nlp models. *Journal of Machine Learning Research*, 24(362):1–40, 2023. 2

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 6

[17] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 6

[18] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[19] Jacky He and contributors. text-to-image-2M: A high-quality, diverse text–image training dataset. https://huggingface.co/datasets/jackyhate/text-to-image-2M, 2024. 6

[20] Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. *arXiv preprint arXiv:2010.02812*, 2020. 2

[21] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5:976235, 2022. 3

[22] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 8

[23] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 3

[24] Hilbert Yuen In Lam, Xing Er Ong, and Marek Mutwil. Large language models in plant biology. *Trends in Plant Science*, 29(10):1145–1155, 2024. 2

[25] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:1008–1019, 2022. 2

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4

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

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

[29] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024. 2, 4

[30] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 4

[31] Kevin Li, Manuel Brack, Sudeep Katakol, Hareesh Ravi, and Ajinkya Kale. Unifusion: Vision-language model as unified encoder in image generation. *arXiv preprint arXiv:2510.12789*, 2025. 2, 3, 4, 7

[32] Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. Unifork: Exploring modality alignment for unified multimodal understanding and generation. *arXiv preprint arXiv:2506.17202*, 2025. 4

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

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 4

[36] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 6

[37] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. *URL https://arxiv. org/abs/2505.05470*, . 4

[38] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. 4, 7

[39] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing, 2025. *URL https://arxiv. org/abs/2504.17761*, . 2, 4

[40] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. 4

[41] Ollin Matsubara and Draw Things AI Team. Megalith-10M: A dataset of 10 million public-domain photographs. `https://huggingface.co/datasets/madebyollin/megalith-10m`, 2024. CC0/Flickr-Commons images; Florence-2 captions available in the *megalith-10m-florence2* variant. 6

[42] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. 5, 6, 7

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 4, 6, 7

[44] 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*, pages 8748–8763. PmLR, 2021. 2, 3

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 4

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 6

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 6, 7

[49] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023. 5

[50] 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. 6

[51] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 2, 4, 6

[52] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 6

[53] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 4, 6

[54] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 4

[55] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4, 6, 7

[56] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 4

[57] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 2, 4, 6, 7

[58] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025. 6, 7, 8

[59] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 2, 4

[60] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. 6

[61] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2, 4, 6, 7

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 4

[63] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 4

[64] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 4

[65] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 4

[66] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025. 2

[67] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024. 2, 4

[68] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2, 4

[69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 4

[70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 6

[71] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun

Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*, 2024. 4, 7