Dynamic Image Prompt Adapter for Scalable Zero-shot Personalized Text-to-Image Generation

Zhizhong Wang, Tianyi Chu, Zeyi Huang*, Nanyang Wang, and Kehan Li

Central Media Technology Institute, Huawei, Hangzhou, China
{endy.won, chutianyi1, huangzeyi2}@huawei.com

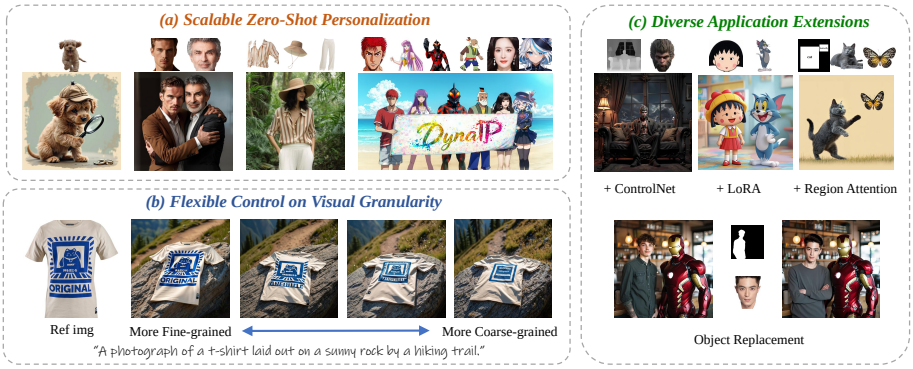


Fig. 1: Representative results showcase the capabilities of DynaIP in: **(a) Scalable zero-shot personalized text-to-image generation**—spanning single-subject to multi-subject—*trained solely on single-subject datasets*. **(b) Flexible control on the visual granularity of concept preservation**, enabled by modulating fusion coefficients for image features across hierarchical levels. **(c) Native compatibility with base model extensions**, unlocking diverse application scenarios.

Abstract. Personalized Text-to-Image (PT2I) generation aims to produce customized images based on reference images. A prominent interest pertains to the integration of an image prompt adapter to facilitate zero-shot PT2I without test-time fine-tuning. However, current methods grapple with three fundamental challenges: **1.** the elusive equilibrium between Concept Preservation (CP) and Prompt Following (PF), **2.** the difficulty in retaining fine-grained concept details in reference images, and **3.** the restricted scalability to extend to multi-subject personalization. To tackle these challenges, we present ***Dyna****mic* ***I****mage* ***P****rompt Adapter (DynaIP)*, a cutting-edge plugin to enhance the *fine-grained concept fidelity*, *CP·PF balance*, and *subject scalability* of state-of-the-art T2I multimodal diffusion transformers (MM-DiT) for PT2I generation. *Our key finding is that MM-DiT inherently exhibit decoupling learning behavior when injecting*

Z. Wang and T. Chu—Equal contribution. *Corresponding author.

reference image features into its dual branches via cross attentions. Based on this, we design an innovative *Dynamic Decoupling Strategy* that removes the interference of concept-agnostic information during inference, significantly enhancing the CP·PF balance and further bolstering the scalability of multi-subject compositions. Moreover, we identify the visual encoder as a key factor affecting fine-grained CP and reveal that *the hierarchical features of commonly used CLIP can capture visual information at diverse granularity levels.* Therefore, we introduce a novel *Hierarchical Mixture-of-Experts Feature Fusion Module* to fully leverage the hierarchical features of CLIP, remarkably elevating the fine-grained concept fidelity while also providing flexible control of visual granularity. Extensive experiments across single- and multi-subject PT2I tasks verify that our DynaIP outperforms existing approaches, while *requiring only single-subject training datasets.*

Keywords: Diffusion Models · Personalized Text-to-Image Generation

1 Introduction

Recent advances in denoising diffusion models have led to remarkable success in Text-to-Image (T2I) generation, where State-of-the-Art (SOTA) models [10, 28] can produce visually plausible, high-quality images based on textual prompts. Subsequent researches have extended T2I to Personalized Text-to-Image (PT2I) generation [11, 48], which aims to synthesize customized images based on both reference images and text prompts, ensuring that objects (*e.g.*, persons, animals) in the generated images maintain specific concept characteristics.

Pioneering PT2I methods, represented by DreamBooth [48] and Textual Inversion [11], primarily relied on fine-tuning models or specific text embeddings, often suffering from low efficiency. To reduce usage cost, recent attention has shifted towards finetuning-free paradigms, as exemplified by IP-Adapter [69] and OminiControl [52], which achieve zero-shot personalization via engineered decoupled cross-attention or multimodal joint attention mechanisms. Furthermore, there are several works [5,22,57,63,65] have extended Single-Subject PT2I (SS-PT2I) to Multi-Subject PT2I (MS-PT2I).

In this work, we are interested in the adapter-based methods [22,57,69] owing to their intrinsic high flexibility and low computational complexity. Methods within this paradigm generally extract features from reference images via a visual encoder [45] and subsequently inject them into the cross-attention layers of a T2I diffusion model. Although substantial advancements have been made, this line of approaches are still confronted with three pivotal limitations:

1. Entanglement of concept-specific information (*e.g.*, ID, shape, and textures) and concept-agnostic information (*e.g.*, pose, perspective, and illumination) in the injected reference image features, leading to an irreconcilable trade-off between Concept Preservation (CP, image & image consistency) and Prompt Following (PF, image & prompt consistency) in generated results, as shown in Fig. 2 (a).

Fig. 2: Limitations of existing adapter-based PT2I methods (*e.g.*, [15, 53, 69]), including (a) irreconcilable trade-off between CP and PF, (b) loss of fine-grained concept details, and (c) restricted scalability to directly extend SS-PT2I to MS-PT2I via mask-guided feature injection. Our proposed DynaIP addresses all these challenges.

2. Insufficient fine-grained feature extraction from reference images, resulting in failure to faithfully recover the intricate object details in customized outputs, as shown in Fig. 2 (b).
3. Significant challenge to directly extend the capability of SS-PT2I to MS-PT2I. Existing approaches often heavily rely on well-curated large-scale multi-subject datasets [57, 65], or encounter pronounced inconsistency when composing multiple subjects via mask-guided feature injection [53, 69], severely impeding their scalability, as shown in Fig. 2 (c).

Facing the aforementioned challenges, in this work, we propose ***Dyna****mic* ***I****mage* ***P****rompt Adapter (DynaIP)*, aiming at improving the *concept fidelity*, *CP·PF balance*, and *subject scalability* of adapter-based methods for PT2I generation tasks. Our DynaIP serves as a cutting-edge plugin for SOTA T2I multi-modal diffusion transformers (MM-DiT) [10, 28] and is compatible with diverse extensions (*e.g.*, ControlNet [71], LoRA [19], and Region Attention [4]) of the same base model, as demonstrated in Fig. 1 (c).

The key insight is that *the latest MM-DiT architecture inherently exhibits decoupling learning behavior when injecting reference image features into its dual branches via cross attentions.* Specifically, the noisy image branch selectively captures the concept-specific information of the reference image, such as the subject's ID and unique appearance, whereas the text branch learns the concept-agnostic information, such as pose, perspective, and illumination. Based on this finding, we design an innovative *Dynamic Decoupling Strategy (DDS)* that removes the interference of concept-agnostic information during the inference phase. This dynamic isolation offers two key merits: (1) *improving the prompt following while maintaining the concept preservation, and thus achieving a better sweet spot between them*, and (2) *mitigating the visual integration inconsistency arising from the retention of concept-agnostic information when extending SS-PT2I to MS-PT2I, thereby significantly boosting the subject scalability.*

Moreover, we believe that the key factor affecting fine-grained CP of reference images lies in the visual encoder. Existing methods [22, 57, 69] typically employ the CLIP [45] image encoder to extract high-level information from deep features of the final or penultimate layers. Unfortunately, these deep features incur substantial loss of detailed information, rendering them inadequate for recovering the fine-grained visual details of reference images. To address this gap, we first systematically investigate the reconstruction capabilities of CLIP's multi-layer features and reveal that *these hierarchical features can capture visual information at diverse granularity levels*. Building on this insight, we further propose a novel *Hierarchical Mixture-of-Experts Feature Fusion Module (HMoE-FFM)* to fully harness CLIP's hierarchical features. HMoE-FFM deploys layer-specific expert networks to process hierarchical features and dynamically calibrates their fusion coefficients via a routing mechanism that adapts to the characteristics of each reference image. This approach not only enhances the fidelity of fine-grained concept details but also facilitates precise modulation of visual granularity. For instance, users can manually adjust the fusion coefficients of experts' outputs, *enabling flexible control over the granularity of concept preservation* according to specific needs, as shown in Fig. 1 (b) and Supplementary Material (SM).

Overall, our contributions are three-fold:

- We identify an intrinsic decoupling learning behavior of the MM-DiT architecture and devise a *Dynamic Decoupling Strategy* to effectively disentangle concept-specific information from concept-agnostic information within reference images. This strategy significantly enhancing the CP·PF balance and further bolstering the scalability of multi-subject compositions.
- We reveal that the hierarchical features of standard CLIP inherently encode visual information across multiple granularities. Based on this, we propose a novel *Hierarchical Mixture-of-Experts Feature Fusion Module* that dynamically integrates these multi-level features to preserve both fine-grained visual details and semantic consistency. This module not only enhances concept fidelity but also enables flexible control over visual granularity.
- Extensive experiments demonstrate the superior performance and broad application potential of DynaIP. Our method outperforms SOTA approaches in attaining *fine-grained concept fidelity* and striking an *optimal balance between CP and PF*. Moreover, it exhibits exceptional *scalability*: being *directly extendable to MS-PT2I scenarios without requiring additional training on multi-subject datasets*, as demonstrated in Fig. 1 (a).

2 Related Work

Text-to-Image Generation. T2I generative models have experienced explosive growth in recent years, especially for denoising diffusion models [10, 18]. Early pioneering models [44, 46, 49] commonly utilized U-Net [47] for noise prediction. Recent studies have replaced it with scalable transformer [54] architectures, giving rise to more advanced models such as Diffusion Transformers (DiT) [7, 42].

Subsequent works [10, 28] have further extended DiT to multimodal DiT (MM-DiT), achieving SOTA T2I generation performance.

Personalized T2I Generation. PT2I generation has garnered significant attention in both academia and industry. Early research predominantly relied on finetuning-based methods [12, 24, 27, 34, 58]. For instance, DreamBooth [48] and Textual Inversion [11] bound visual concepts to text identifiers via finetuning, while LoRA [19] introduced lightweight parameters to enable efficient adjustments. However, these methods are plagued by the heavy burden of per-subject finetuning. In response, IP-Adapter [69] and BLIP Diffusion [30] incorporated additional image encoders and layers to process reference images and injecting image features into diffusion models. These adapter-based techniques enable zero-shot PT2I generation, emerging as the dominant research paradigm and inspiring a wealth of follow-up works [8, 13, 16, 21, 26, 55, 56, 59], including several multi-subject approaches [22, 35, 57, 60, 65, 72]. Aligning with part of our motivations, DisEnvisioner [15] disentangles and enriches concept-specific attributes by learning individual image tokens. In contrast, our method directly leverages the decoupled learning behavior of MM-DiT to separate concept-specific information from concept-agnostic content. Furthermore, DisEnvisioner still suffers from the loss of fine-grained concept details as it relies on deep-layer features from CLIP, as illustrated in Fig. 2 (b). Additionally, most of the aforementioned efforts are built on U-Net-based diffusion models, with limited exploration of adapter-based PT2I techniques for SOTA DiT-based models such as FLUX.1 [28].

For DiT architectures, methods such as OminiControl [52] adopt unified conditioning strategies to handle text embeddings, latent tokens, and VAE-encoded reference subjects. Although showing promise in PT2I generation [2, 5, 14, 20, 36, 37, 50, 63, 64], they often necessitate constructing complex cross-pair and multi-subject datasets for base model finetuning. In addition, the full attention mechanism's computational complexity escalates exponentially with more reference conditions, particularly in multi-subject scenarios, resulting in lower efficiency and limited flexibility compared to adapter-based methods.

Another category of methods leverages Multimodal Large Language Models (MLLMs) [1,73] to conduct multi-modal training through text-image interleaving to bridge the gap between image and text prompts [9, 31, 40, 41, 51, 61, 62, 66]. Yet, these methods typically require re-training MLLM encoders or generators, imposing substantial demands on training resources.

Multi-layer Feature Fusion. In the context of MLLMs, several empirical studies [3, 6, 68] have shown that multi-layer visual features can enhance model performance. Lin *et al*. [32] categorizes existing fusion strategies into four categories and reveals that direct fusion (addition/concatenation) at the input stage consistently yields superior performance across various configurations. However, due to the fundamental task divergence between understanding-oriented and generation-oriented tasks, these observations may not be generalizable to our PT2I scenario (see comparison results in Sec. 5.4). For the PT2I tasks, while a handful of works [59, 72] have also attempted to fuse multi-layer features for performance improvement, their fusion strategies remain limited to simple addition

or concatenation. In contrast, our HMoE-FFM leverages a MoE architecture to integrate hierarchical visual features, with a dynamic routing mechanism that adaptively calibrates fusion coefficients based on the unique attributes of each input reference image. This approach not only enhances fine-grained visual details and semantic consistency but also enables precise modulation of visual granularity, as validated in Sec. 5.4 and SM.

3 Preliminary

3.1 Multimodal Diffusion Transformer

MM-DiT [10,28] enhances the vanilla DiT [7,42] by using a unified Multi-Modal Attention (MMA) mechanism to jointly process noisy image tokens $X \in \mathbb{R}^{m \times d}$ and text tokens $T \in \mathbb{R}^{n \times d}$. Here, d represents the latent dimension, while m and n represent the sequence lengths of image and text tokens, respectively.

Specifically, the MMA mechanism projects text and image tokens into position-encoded query $\mathbf{Q} = [\mathbf{Q}_T, \mathbf{Q}_X]$, key $\mathbf{K} = [\mathbf{K}_T, \mathbf{K}_X]$, and value $\mathbf{V} = [\mathbf{V}_T, \mathbf{V}_X]$ representations, enabling cross-modal attention computation across all tokens:

$$MMA([T, X]) = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} = [T^{MMA}, X^{MMA}], \tag{1}$$

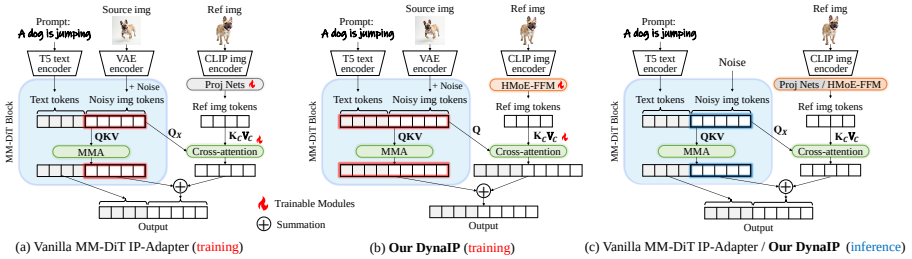
where $[\cdot, \cdot]$ denotes the sequence concatenation. T^{MMA} and X^{MMA} are the new text and noisy image tokens after MMA, respectively. The MMA allows both representations to operate within their respective spaces while still taking the other into account.

3.2 Image Prompt Adapter for MM-DiT

Image Prompt Adapter (IP-Adapter) [69] is a lightweight method to endow pretrained T2I diffusion models with the ability to utilize image prompts. Its core lies in a Decoupled Cross-Attention (DCA) mechanism that distinctly separates the cross-attention layers for processing text features and image features. The majority of prior works have predominantly relied on U-Net-based diffusion models. When transferring to MM-DiT-based models, a vanilla solution [53] is computing CA between noisy image tokens X and reference image tokens $C \in \mathbb{R}^{h \times d}$ (Eq. (2)), as illustrated in Fig. 3 (a, c). Here, h represents the sequence length of reference image tokens, which is typically extracted by a feature extraction model comprising a pretrained image encoder (*e.g.*, CLIP [45]) and a trainable projection network.

$$CA(X, C) = softmax(\frac{\mathbf{Q}_X\mathbf{K}_C^\top}{\sqrt{d}})\mathbf{V}_C, \tag{2}$$

where $\mathbf{Q}_X$ is the query representation of noisy image tokens X in Eq. (1), while $\mathbf{K}_C/\mathbf{V}_C$ are newly learned key/value representations obtained by linearly projecting the reference image tokens C.



(a) Vanilla MM-DiT IP-Adapter (training) (b) **Our DynaIP** (training) (c) Vanilla MM-DiT IP-Adapter / **Our DynaIP** (inference)

Fig. 3: Training/Inference pipeline of **(a, c)** vanilla IP-Adapter and **(b, c)** our DynaIP.

Then, DCA mechanism integrates the output of image CA with the output of text CA:

$$DCA(T, X, C) = X^{MMA} + \lambda \cdot CA(X, C). \tag{3}$$

Here, the output of text CA in MM-DiT is embedded in the new noisy image tokens X^{MMA} after MMA (Eq. (1)). λ is a weight factor to control the influence of the reference image features.

4 Dynamic Image Prompt Adapter

As discussed in Sec. 1, current adapter-based methods face fundamental challenges in fine-grained concept fidelity, CP·PF balance, and subject scalability. To tackle this challenges, we introduce *DynaIP*, which comprises two novel components. First, to improve CP·PF balance and subject scalability, we propose *Dynamic Decoupling Strategy*. It leverages an intrinsic characteristic of MM-DiT to dynamically disentangle concept-specific information from concept-agnostic information in reference images. Second, we present *Hierarchical Mixture-of-Experts Feature Fusion Module*, which dynamically integrates multi-level CLIP features according to different reference inputs. It not only boosts fine-grained concept fidelity but also enables flexible control over visual granularity.

4.1 Dynamic Decoupling Strategy

As introduced in Sec. 3.2, the vanilla IP-Adapter holistically injects reference image features into the noisy image, inevitably introducing concept-agnostic information that undermines personalization quality, as shown in Fig. 2. To address this issue, our key strategy is to enable reference image features to perform CA with both the noisy image branch and the text branch of MM-DiT in the training phase, as illustrated in Fig. 3 (b) and formalized below:

$$CA([T, X], C) = softmax(\frac{\mathbf{Q}\mathbf{K}_C^\top}{\sqrt{d}})\mathbf{V}_C. \tag{4}$$

Then, the output of CA is added to the complete output of MMA (Eq. (1)):

$$DCA'(T, X, C) = [T^{MMA}, X^{MMA}] + \lambda \cdot CA([T, X], C). \tag{5}$$

The core insight behind this strategy is that the *text branch in MM-DiT interacts with the noisy image branch via MMA (Eq. (1)) to regulate the overall semantics (*e.g.*, pose and perspective) and visual presentation (*e.g.*, illumination) of the generated image—all of which are independent of specific concept information*. As a result, the text branch primarily learns concept-agnostic knowledge from the reference image, enabling the noisy image branch to focus on selectively encoding the concept-specific information, such as the unique visual characteristics of the reference subject.

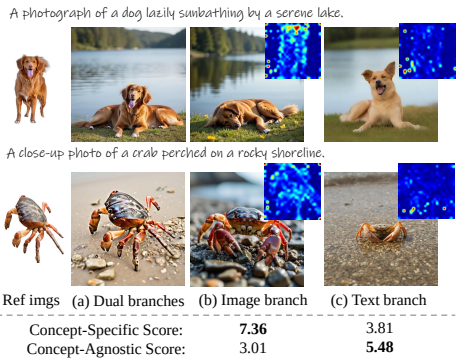


Fig. 4: Illustration of the decoupling learning behavior of MM-DiT.

To demonstrate this intrinsic decoupling learning behavior of MM-DiT, we first visualize the attention maps that illustrate the interactions between each branch's features and the reference image features, as shown in the upper right corners of Fig. 4 (b-c). As observed, the noisy image branch focuses on the core appearance and semantic regions of the reference subject, whereas the text branch exhibits relatively scattered attention, with higher activation values only in areas irrelevant to the subject's appearance. It is well aligned with the phenomena manifested in the generated results in Fig. 4 (a-c): (b) the noisy image branch focuses on capturing the concept-specific information like the subject's ID and unique appearance; (c) the text branch specializes in learning the concept-agnostic information, such as pose and perspective in the first row, and illumination in the second row; (a) combining text branch with noisy image branch yields outputs where concept-specific and concept-agnostic information are entangled, leading to copy-paste issues that compromise the quality and controllability of generated results. Besides, we also leverage powerful MLLMs (InternVL3-78B [74] and Qwen3-VL-32B [67]) to quantify the preservation of concept-specific and -agnostic information across over 1,000 image pairs (*see evaluation prompt in SM*). The quantitative scores shown in the bottom rows of Fig. 4 further corroborate the decoupling learning behavior of MM-DiT.

Based on the above findings, *during inference, we can dynamically eliminate concept-agnostic information by performing CA exclusively with the noisy image branch*, as shown in Fig. 3 (c) and Eqs. (2, 3). This simple yet effective strategy substantially improves the CP·PF balance and bolsters the scalability of multi-subject compositions, as will be further validated in later Sec. 5.4.

4.2 Hierarchical Mixture-of-Experts Feature Fusion

The image encoder used in IP-Adapter plays a critical role to extract concept information from reference images. Current methods typically utilize CLIP [45], extracting features from its final or penultimate layers. While a few works [26,53] have explored more powerful alternatives such as SigLIP [70] and DINOv2 [39],

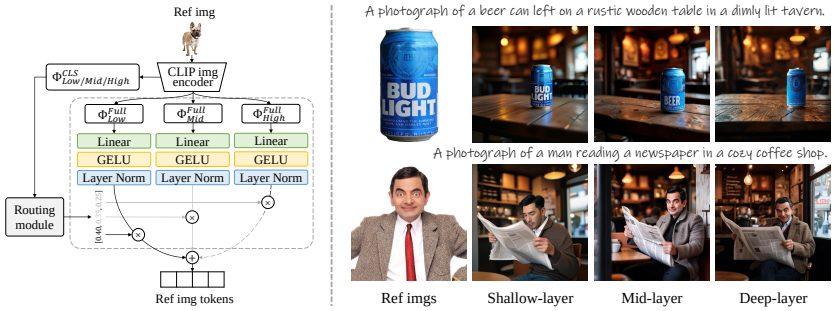


Fig. 5: Left: Architecture of our proposed HMoE-FFM. **Right**: Personalization results generated by injecting features from different layers of CLIP via cross-attentions, demonstrating that CLIP's hierarchical features can capture visual information at diverse granularity levels.

they still rely on deep-layer features. These deep-layer features suffer from substantial loss of detailed information, making them inadequate for recovering fine-grained visual details of reference images. To address this limitation, we first investigate CLIP's feature extraction capabilities by systematically visualizing the reconstruction performance of its multi-layer features—achieved by injecting each layer's features individually via cross-attentions. As demonstrated in Fig. 5-right, we made a striking observation: *CLIP's hierarchical features excel at capturing visual information across varying granularity levels*. Specifically, shallow-layer features (e.g., layer 10) capture low-level patterns such as lines and text; mid-layer features (e.g., layer 17) capture fine-grained textures and structures like facial details; and deep-layer features (e.g., layer 24) capture semantic information alongside coarse-grained textures and structures (*results across more layers can be found in SM*). Unfortunately, these multi-granularity hierarchical features have not been fully utilized in existing works.

To bridge this gap and fully harness the multi-granularity potential of CLIP's hierarchical features, we further propose a novel *Hierarchical Mixture-of-Experts Feature Fusion Module (HMoE-FFM)*, as depicted in Fig. 5-left. Specifically, let Φ_{Low}, Φ_{Mid}, and Φ_{High} denote the image features from CLIP's shallow, middle, and deep layers, respectively. $\Phi^{CLS} \in \mathbb{R}^{1\times d_1}$ represents the class tokens of these features, and $\Phi^{Full} \in \mathbb{R}^{g\times d_1}$ represents their full tokens, where g denotes the sequence length and d_1 denotes the feature dimension.

HMoE-FFM first deploys layer-specific expert networks $Expert_l$—each comprising a linear layer with GELU [17] activation and layer normalization—to process the full tokens of hierarchical features, formalized as follows:

$$e_l = Expert_l(\Phi_l^{Full}), \quad l \in [Low, Mid, High]. \tag{6}$$

Additionally, a routing module (a two-layer Multi-Layer Perceptron with Tanh as the middle activation function) dynamically calibrates the fusion coefficients

for each expert's output based on the class tokens of hierarchical features:

$$w_l = Route_l(\Phi_l^{CLS}), \quad l \in [Low, Mid, High], \quad \sum_l w_l = 1. \tag{7}$$

Finally, the output feature is the weighted fusion of all experts' outputs:

$$\Phi_{Fused} = \sum_l w_l \cdot e_l, \quad l \in [Low, Mid, High]. \tag{8}$$

Notably, this approach not only leverages the complementary strengths of multi-level features to preserve both fine-grained visual details and semantic consistency, but also enables precise modulation of visual granularity. For instance, users can manually adjust the fusion coefficients (Eq. (7)) of the experts' outputs, thereby achieving flexible control over the granularity of concept preservation tailored to specific needs—as illustrated in Fig. 1 (b) and *more results in SM*. Comparisons between our HMoE-FFM and other widely used feature fusion approaches will be presented in later Sec. 5.4.

4.3 Multi-Subject Personalization

Without requiring additional training on multi-subject datasets, our DynaIP can be directly extended to multi-subject personalization scenarios via a straightforward mask-guided region-level feature injection. Concretely, given N reference images corresponding N masks, we first compute the CA between noisy image tokens X and each reference image tokens C_i as specified in Eq. (2). These CA outputs are then added to the corresponding regions of the text CA output features, guided by their respective masks:

$$DCA''(T, X, C) = X^{MMA} + \sum_{i=1}^{N} \lambda_i \cdot M_i \cdot CA(X, C_i), \tag{9}$$

where X^{MMA} implies the output of text CA as defined in Eq. (3), M_i is a binary mask specifying the regions in the generated image where the subject needs to be replaced by the subject from the i_{th} reference image; it can either be manually annotated by users or automatically generated using existing grounding and segmentation tools [25,33]. In our experiments, we adopt the automatic approach or preset masks for multi-subject personalization. λ_i is the weight factor to regulate the influence of the i_{th} reference image features. As will be shown in later Sec. 5, thanks to our proposed Dynamic Decoupling Strategy, our results can significantly reduce inconsistencies when composing different reference subjects and greatly enhance the overall harmony of multi-subject generation—even when these subjects possess distinct styles and visual attributes. Furthermore, this method facilitates effective object replacement, which in turn supports flexible editing of the generated results, as shown in the bottom case of Fig. 1 (c).

5 Experimental Results

5.1 Experiment Settings

Implementation Details. We build our model on FLUX.1-Dev [28], with OpenCLIP ViT-L/14-336[1] adopted as the image encoder. The backbone includes 57 MM-DiT blocks; we augment each block with a new image cross-attention layer. For the HMoE-FFM, we utilize features from CLIP's layer 10, 17, and 24 as the input features for the low-, mid-, and high-level expert networks, respectively. The total number of trainable parameters in our DynaIP is approximately 1.4B. During training, we only optimize the augmented layers while keeping the parameters of the pretrained models fixed. The training objective is the flow-matching loss [28]. The training process is divided into two sequential stages: The initial stage (20,000 iterations) establishes fundamental capabilities through exclusive intra-pair training, developing robust subject-specific adaptation. Building upon this foundation, the second stage (80,000 iterations) leverages the cross-pair dataset to mitigate copy-paste artifacts [57]. For optimization, we employ the AdamW optimizer with a cosine learning rate scheduler initialized at 0.00002, combined with a weight decay of 0.0001. All experiments were conducted on 8 Ascend 910B NPUs, with a total batch size of 16 and a training resolution of 1024×1024. To enable classifier-free guidance, we use a 5% probability to drop text and image individually, and a 5% probability to drop text and image simultaneously. For the training of HMoE-FFM, each expert's feature is dropped independently with a 5% probability. Finally, we set the image weight factor λ to 1.0 for both training and inference stages.

Datasets. *We train our model only on single-subject datasets, please refer to SM for details.* For evaluation, we assess SS-PT2I performance on DreamBench++ [43], which contains 1,350 test samples across various categories—including animals, humans, and objects—and covers both photorealistic and non-photorealistic styles. For MS-PT2I evaluation, we follow previous studies [22, 57] to establish a new benchmark named DynaIP-Bench. Specifically, we source diverse subjects and prompts from existing works [5, 43, 57] to construct 642 two-subject pairs and 246 three-subject triplets, forming a comprehensive set of 888 multi-subject test samples (*details are provided in SM*).

Evaluation Metrics. Previous works [48, 57] typically compare feature-based similarity metrics from models like CLIP [45] and DINO [39]. However, as highlighted in [2, 43], these metrics only capture global semantic similarity and suffer from extreme noise and bias. Therefore, we adhere to the evaluation protocol outlined in [43], which systematically assesses PT2I performance through two key dimensions: Concept Preservation (CP) and Prompt Following (PF), using an MLLM [23]. This protocol demonstrates better alignment with human preferences compared to traditional metrics. For reproducibility and mitigating model-specific biases, we employ two SOTA open-source MLLMs—InternVL3-78B [74] and Qwen3-VL-32B [67]—and report the average of their scores. The

[1] https://huggingface.co/openai/clip-vit-large-patch14-336

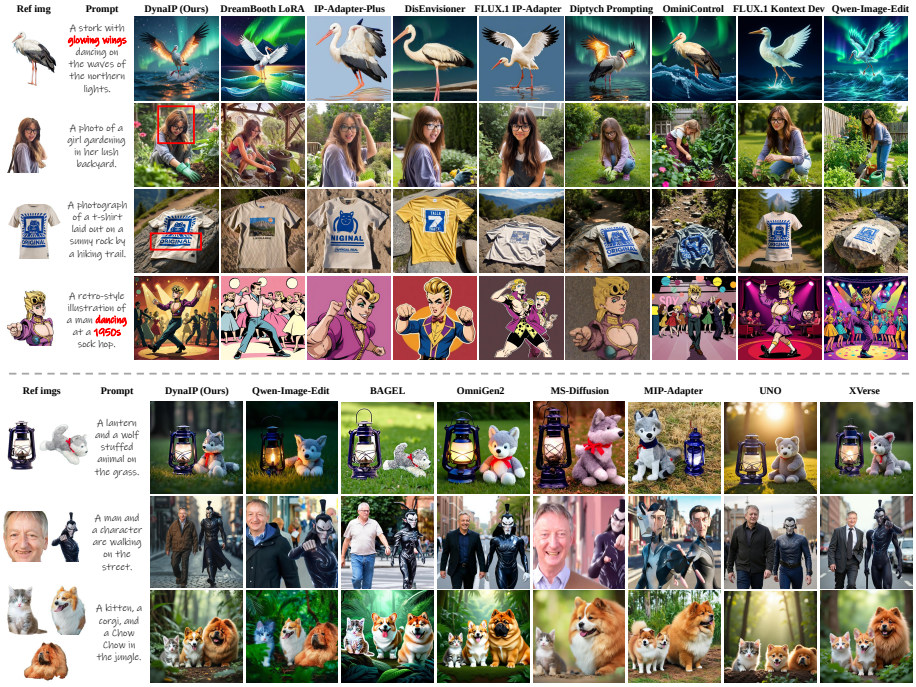


Fig. 6: Qualitative comparisons on single- and multi-subject PT2I generation.

goal of the PT2I task is to maximize the Nash utility [38], defined as the product of CP and PF scores. Note that for MS-PT2I scenarios, we use the average CP score across all subjects. *The detailed evaluation prompts can be found in SM.*

Compared Baselines. For SS-PT2I, we compare our model with leading open-source methods such as finetuning-based DreamBooth [48], DreamBooth LoRA [19], and Textual Inversion [11]; zero-shot UNet-based IP-Adapter-Plus [69] and DisEnvisioner [15]; DiT-based FLUX.1 IP-Adapter [53], Diptych Prompting [50], OminiControl [52], Self-Distillation [2], and FLUX.1 Kontext Dev [29]. For MS-PT2I, we compare our model with MLLM-based Qwen-Image-Edit [61], BAGEL [9], and OmniGen2 [62], and other finetuning-free methods like MS-Diffusion [57], MIP-Adapter [22], UNO [63], and XVerse [5]. For fair comparison, we adapt input prompts to the officially recommended format of each compared method to avoid biases from mismatched prompt formats.

5.2 Qualitative Comparison

We present qualitative comparison results for SS-PT2I in Fig. 6 (top). Our DynaIP preserves fine-grained subject details with high fidelity—such as the facial features in Row 2 and the legible text in Row 3—while simultaneously achieving strong prompt alignment, as evidenced by the glowing wings in Row 1 and

Table 1: Quantitative comparisons with SOTA methods. CP: Concept Preservation, PF: Prompt Following. The **highest** and second-highest scores are highlighted.

Method	Single-subject			Multi-subject		
	CP	PF	CP · PF	CP	PF	CP · PF
DreamBooth	0.458	0.721	0.330	-	-	-
DreamBooth LoRA	0.594	0.840	0.499	-	-	-
Textual Inversion	0.348	0.633	0.220	-	-	-
IP-Adapter-Plus	**0.738**	0.668	0.493	-	-	-
DisEnvisioner	0.559	0.664	0.371	-	-	-
FLUX.1 IP-Adapter	0.681	0.600	0.408	-	-	-
Diptych Prompting	0.616	0.839	0.517	-	-	-
OminiControl	0.596	0.895	0.534	-	-	-
Self-Distillation	0.513	0.870	0.447	-	-	-
FLUX.1 Kontext Dev	0.718	0.893	0.641	-	-	-
Qwen-Image-Edit	0.686	0.930	0.638	0.612	0.984	0.602
BAGEL	0.603	0.926	0.558	0.566	0.885	0.500
OmniGen2	0.622	0.922	0.574	0.552	0.952	0.526
MS-Diffusion	0.686	0.828	0.568	0.584	0.850	0.496
MIP-Adapter	0.692	0.649	0.449	0.388	0.713	0.276
UNO	0.721	0.799	0.576	0.509	0.857	0.436
XVerse	0.643	0.869	0.559	0.548	0.890	0.488
Our DynaIP	0.696	**0.934**	**0.650**	**0.617**	**0.997**	**0.615**

the 1950s-style dancing in Row 4. In contrast, existing methods either fail to maintain fine-grained subject details or struggle to adhere to input prompts.

Furthermore, we showcase qualitative comparisons for MS-PT2I in Fig. 6 (bottom). Notably, while all competing methods are trained on carefully curated multi-subject datasets, our DynaIP delivers superior performance in both subject consistency and visual harmony—*despite being trained exclusively on single-subject data*. Specifically, our method generates plausible, harmonious, and naturally coordinated compositions, even when reference subjects exhibit distinct styles and visual attributes (e.g., Row 2). By contrast, existing approaches frequently produce copy-paste artifacts or inconsistent subject renderings, resulting in reduced concept fidelity and visually disjointed compositions.

5.3 Quantitative Comparison

Automatic Scores. Quantitative comparison results for both SS- and MS-PT2I are presented in Tab. 1. For SS-PT2I, our method achieves the highest PF and CP·PF scores. The PF score quantifies adherence to text prompts, while the CP·PF score reflects balanced performance between subject consistency and prompt alignment. As explicitly emphasized in [43], this balanced CP·PF performance is the ultimate objective of the PT2I task. Although our CP score is marginally lower than that of certain competing methods (*e.g.*, IP-Adapter-Plus [69]), we attribute this discrepancy primarily to two factors: first, these baselines fail to effectively decouple concept-agnostic information, leading generated images closely resemble the input without meaningful transformation, as illustrated in Fig. 6 and further validated by their substantially lower PF scores; second, the CP score itself exhibits limited sensitivity to both such copy-paste artifacts and fine-grained subject details. Therefore, to more effectively

Fig. 7: Qualitative ablation study results of **Left**: Dynamic Decoupling Strategy (DDS) and **Right**: different feature fusion approaches and token-concat baseline.

demonstrate the superiority of our DynaIP in terms of CP, *we further conduct additional user studies in SM, where our method outperforms all competitors.*

For MS-PT2I, our method outperforms all baselines by achieving the highest scores across all metrics, demonstrating its superiority in both maintaining subject consistency and prompt adherence. Notably, our CP score in MS scenarios surpasses those of all competitors—a stark contrast to the SS scenarios. We analyze this is because existing approaches frequently suffer from flaws such as subject omission, identity confusion, or unnatural subject fusion—issues that directly degrade the performance of CP, as shown in Fig. 6. In contrast, our method effectively mitigates these problems and generates visually harmonious multi-subject compositions.

5.4 Ablation Studies

Effects of DDS. As shown in Fig. 7-left, without DDS (*i.e.*, dual-branch training followed by dual-branch inference), the concept-agnostic information of the reference images, such as pose, perspective, and illumination, is retained in the generated results, leading to degraded editability and disharmonious multi-subject compositions. These issues are effectively mitigated by DDS, confirming its ability to disentangle concept-specific information from concept-agnostic information. Additionally, quantitative results in Tab. 2 (1-2) further validate its effectiveness in enhancing the CP·PF balance and multi-subject scalability.

Superiority of HMoE-FFM. We compare our HMoE-FFM with two widely adopted feature fusion methods [32, 59, 72]: element-wise addition (add) and channel-wise concatenation (concat). For fair comparison, we first project each hierarchical CLIP feature by the same layers as HMoE-FFM, and then simply add or concatenate them without MoE routing. As illustrated in Fig. 7 (a-c) and Tab. 2 (3-4), fusing multi-layer CLIP features via add or concat yields suboptimal performance on both CP and PF—particularly in terms of fine-grained fidelity and style transfer. This highlights the superiority of our HMoE-FFM, which more effectively integrates hierarchical features through layer-specific expert networks and a dynamic routing mechanism that adapts to the characteristics

of each reference image. Additionally, as shown in Fig. 1 (b) and SM, our HMoE-FFM provides greater flexibility to enable run-time control over the granularity of concept preservation, a capability unattainable with add and concat approaches. On the other hand, to complement Fig. 5-right, we further present quantitative results for HMoE-FFM using different CLIP layers individually in Tab. 2 (5-7). Evidently, while shallow-layer features excel at capturing low-level patterns (e.g., lines and text), their limited capacity to grasp mid- and high-level attributes results in lower CP scores. Similarly, deep-layer features yield lower CP scores due to their lack of low-level and fine-grained information; however, they achieve the highest PF scores by virtue of their stronger ability to capture semantic information. In contrast, mid-layer features primarily focus on fine-grained details while incorporating a certain amount of semantic information, thus striking a more favorable balance between CP and PF. By harnessing the strengths of these multi-layer features, our HMoE-FFM integrates low-level, fine-grained, and semantic information, thereby achieving an optimal balance between CP and PF. *More results and detailed analyses can be found in SM.*

Table 2: Quantitative ablation study results.

Setting	Single-subject			Multi-subject		
	CP	PF	CP · PF	CP	PF	CP · PF
(1) **Full Model**	0.696	0.934	**0.650**	**0.617**	**0.997**	**0.615**
(2) w/o DDS	**0.785**	0.799	0.627	0.499	0.545	0.272
(3) Add Fusion	0.691	0.916	0.633	0.609	0.995	0.606
(4) Concat Fusion	0.692	0.909	0.629	0.607	0.992	0.602
(5) Only Shallow	0.627	0.924	0.579	0.464	0.991	0.460
(6) Only Mid	0.670	0.928	0.622	0.603	0.993	0.599
(7) Only Deep	0.480	**0.950**	0.456	0.474	0.995	0.471
(8) Token-Concat	0.709	0.886	0.628	-	-	-

Comparison with Token-Concat. We further implement a token concatenation baseline [52] under the same MM-DiT backbone and training datasets as our DynaIP. For fair comparison, we adopt LoRA training [63] with the same parameter budget as our method. As shown in Fig. 7 (d) and Tab. 2 (8), the comparison clearly demonstrates our method outperforms native concatenation.

6 Conclusion

In this paper, we present DynaIP, a plug-and-play adapter to empower diffusion transformers for scalable zero-shot personalized text-to-image generation. Our DynaIP introduces two key innovations. The first is a *Dynamic Decoupling Strategy* to disentangle concept-specific information from concept-agnostic information within reference images. This strategy significantly enhances the critical balance between concept preservation and prompt following, while simultaneously bolstering the scalability of multi-subject compositions. The second is a *Hierarchical Mixture-of-Experts Feature Fusion Module* to dynamically integrates the multi-level CLIP features to preserve both fine-grained visual details and semantic consistency. It not only enhances the concept fidelity of reference images but also provides flexible control over visual granularity. Extensive experiments show the superiority of our DynaIP in both single- and multi-subject personalization, while *requiring only single-subject training datasets.*

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Cai, S., Chan, E.R., Zhang, Y., Guibas, L., Wu, J., Wetzstein, G.: Diffusion self-distillation for zero-shot customized image generation. In: CVPR. pp. 18434–18443 (2025)
3. Cao, Y., Liu, Y., Chen, Z., Shi, G., Wang, W., Zhao, D., Lu, T.: Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. arXiv preprint arXiv:2410.11829 (2024)
4. Chen, A., Xu, J., Zheng, W., Dai, G., Wang, Y., Zhang, R., Wang, H., Zhang, S.: Training-free regional prompting for diffusion transformers. arXiv preprint arXiv:2411.02395 (2024)
5. Chen, B., Zhao, M., Sun, H., Chen, L., Wang, X., Du, K., Wu, X.: Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. NeurIPS (2025)
6. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: Lion: Empowering multimodal large language model with dual-level visual knowledge. In: CVPR. pp. 26540–26550 (2024)
7. Chen, J., Jincheng, Y., Chongjian, G., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
8. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: CVPR. pp. 6593–6602 (2024)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
10. 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. In: ICML (2024)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
12. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. NeurIPS **36**, 15890–15902 (2023)
13. Guo, Z., Wu, Y., Zhuowei, C., Zhang, P., He, Q., et al.: Pulid: Pure and lightning id customization via contrastive alignment. NeurIPS **37**, 36777–36804 (2024)
14. Guo, Z., Zhang, P., Wu, Y., Mou, C., Zhao, S., He, Q.: Musar: Exploring multi-subject customization from single-subject dataset via attention routing. arXiv preprint arXiv:2505.02823 (2025)
15. He, J., Haodong, L., Shen, G., Yingjie, C., Qiu, W., Chen, Y.C., et al.: Disenvisioner: Disentangled and enriched visual prompt for customized image generation. In: ICLR (2025)
16. He, Q., Yao, A.: Conceptrol: Concept control of zero-shot personalized image generation. arXiv preprint arXiv:2503.06568 (2025)
17. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
19. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
20. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
21. Huang, M., Mao, Z., Liu, M., He, Q., Zhang, Y.: Realcustom: narrowing real text word for real-time open-domain text-to-image customization. In: CVPR. pp. 7476–7485 (2024)
22. Huang, Q., Fu, S., Liu, J., Jiang, H., Yu, Y., Song, J.: Resolving multi-condition confusion for finetuning-free personalized image generation. In: AAAI. pp. 3707–3714 (2025)
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
24. Jiang, J., Zhang, Y., Feng, K., Wu, X., Li, W., Pei, R., Li, F., Zuo, W.: Mc2: Multi-concept guidance for customized multi-concept generation. arXiv preprint arXiv:2404.05268 (2024)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
26. Kong, L., Wu, K., Hu, X., Han, W., Peng, J., Xu, C., Luo, D., Li, M., Zhang, J., Wang, C., Fu, Y.: Custany: Customizing anything from a single example. In: CVPR (2025)
27. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR. pp. 1931–1941 (2023)
28. Labs, B.F.: Flux: Official inference repository for flux.1 models. `https://github.com/black-forest-labs/flux` (2024)
29. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
30. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. NeurIPS **36**, 30146–30166 (2023)
31. Li, W., Xu, X., Liu, J., Xiao, X.: Unimo-g: Unified image generation through multimodal conditional diffusion. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6173–6188 (2024)
32. Lin, J., Chen, H., Fan, Y., Fan, Y., Jin, X., Su, H., Fu, J., Shen, X.: Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In: CVPR. pp. 4156–4166 (2025)
33. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024)
34. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: concept neurons in diffusion models for customized generation. In: ICML. pp. 21548–21566 (2023)
35. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)

36. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487 (2025)
37. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. arXiv preprint arXiv:2504.16915 (2025)
38. Nash, J.F., et al.: The bargaining problem. Econometrica **18**(2), 155–162 (1950)
39. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR pp. 1–31 (2024)
40. Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-g: Generating images in context with multimodal large language models. In: ICLR (2024)
41. Patel, M., Jung, S., Baral, C., Yang, Y.: λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. TMLR (2024)
42. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
43. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: ICLR (2025)
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
47. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
48. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023)
49. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS **35**, 36479–36494 (2022)
50. Shin, C., Choi, J., Kim, H., Yoon, S.: Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. arXiv preprint arXiv:2411.15466 (2024)
51. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: CVPR. pp. 14398–14409 (2024)
52. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: ICCV. pp. 14940–14950 (October 2025)
53. Team, I.: Instantx flux.1-dev ip-adapter page. `https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter` (2024)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)

55. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
56. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
57. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In: ICLR (2025)
58. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: ICCV. pp. 7677–7689 (2023)
59. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: ICCV. pp. 15943–15953 (2023)
60. Wei, Z., Su, Q., Qin, L., Wang, W.: Mm-diff: High-fidelity image personalization via multi-modal condition integration. arXiv preprint arXiv:2403.15059 (2024)
61. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
62. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
63. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: ICCV. pp. 18682–18692 (October 2025)
64. Wu, T., Jiang, Y., Lu, Y., Wang, Z., Huang, Z., Qin, Z., Li, X.: Multicrafter: High-fidelity multi-subject generation via disentangled attention and identity-aware preference alignment. In: CVPR. pp. 36691–36702 (2026)
65. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. IJCV pp. 1–20 (2024)
66. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024)
67. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
68. Yao, H., Wu, W., Yang, T., Song, Y., Zhang, M., Feng, H., Sun, Y., Li, Z., Ouyang, W., Wang, J.: Dense connector for mllms. NeurIPS **37**, 33108–33140 (2024)
69. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
70. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
71. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
72. Zhang, Y., Song, Y., Liu, J., Wang, R., Yu, J., Tang, H., Li, H., Tang, X., Hu, Y., Pan, H., et al.: Ssr-encoder: Encoding selective subject representation for subject-driven generation. In: CVPR. pp. 8069–8078 (2024)
73. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
74. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)