Mod-Adapter: Tuning-Free and Versatile Multi-concept Personalization via Modulation Adapter

Anonymous Author(s)
Affiliation
Address
email

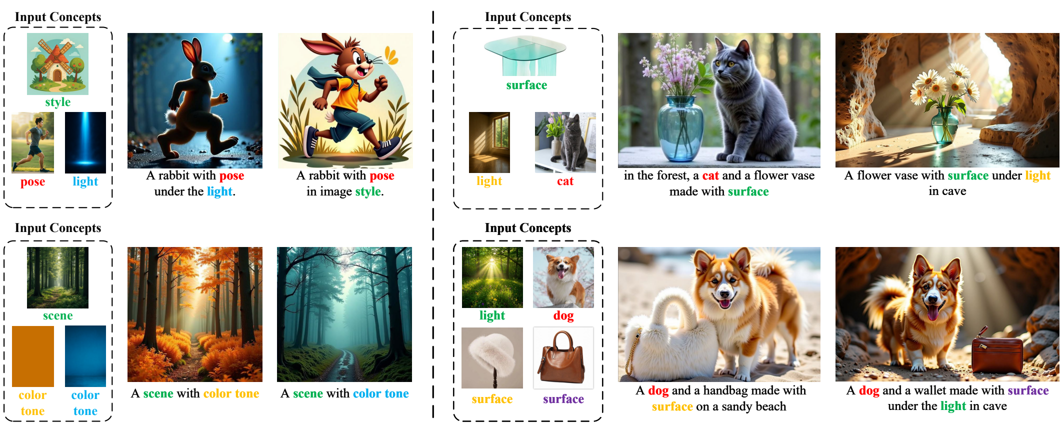

Figure 1: **Results of our multi-concept personalized image generation method.** Our method enables customizing both object and abstract concepts (e.g., pose, light, surface) without test-time fine-tuning. The colored words in the prompt below image indicate concepts to be personalized.

Abstract

Personalized text-to-image generation aims to synthesize images of user-provided concepts in diverse contexts. Despite recent progress in multi-concept personalization, most are limited to object concepts and struggle to customize abstract concepts (e.g., pose, lighting). Some methods have begun exploring multi-concept personalization supporting abstract concepts, but they require test-time fine-tuning for each new concept, which is time-consuming and prone to overfitting on limited training images. In this work, we propose a novel tuning-free method for multi-concept personalization that can effectively customize both object and abstract concepts without test-time fine-tuning. Our method builds upon the modulation mechanism in pre-trained Diffusion Transformers (DiTs) model, leveraging the localized and semantically meaningful properties of the modulation space. Specifically, we propose a novel module, Mod-Adapter, to predict concept-specific modulation direction for the modulation process of concept-related text tokens. It incorporates vision-language cross-attention for extracting concept visual features, and Mixture-of-Experts (MoE) layers that adaptively map the concept features into the modulation space. Furthermore, to mitigate the training difficulty caused by the large gap between the concept image space and the modulation space, we introduce a VLM-guided pretraining strategy that leverages the strong image understanding capabilities of vision-language models to provide semantic supervision signals. For a comprehensive comparison, we extend a standard benchmark by incorporating ab-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

stract concepts. Our method achieves state-of-the-art performance in multi-concept personalization, supported by quantitative, qualitative, and human evaluations.

1 Introduction

Personalized text-to-image generation aims to synthesize images of the concepts specified by user-provided images in diverse contexts. Recently, this technique has attracted increasing research attention due to its broad applications, such as poster design and storytelling. However, existing personalized generation methods primarily focus on object concepts (e.g., common objects and animals) and struggle to personalize non-object/abstract concepts (e.g., pose and lighting), limiting their wider applicability. Recently, TokenVerse [9] proposes a multi-concept personalization framework supporting both object and abstract concepts. However, it requires fine-tuning for each new concept image at test time, which is time-consuming and tends to overfit on the single training image, leading to suboptimal results. In this work, we take the first step toward a tuning-free framework enabling versatile multi-concept personalization for both objects and abstract concepts as shown in Fig.1.

Existing tuning-free personalized generation methods [33, 37, 14, 23] often face two challenges when customizing abstract concepts. First, these methods failed to decouple the object concept and abstract concept from the input image due to the lack of an effective mechanism for extracting abstract features. As a result, they tend to directly replicate the object into the generated image. For example, when personalizing persons' pose concept, the generated person often closely resembles the one in the input concept image, rather than merely reflecting the pose features. This compromises the alignment between the generated image and the input prompt. Second, the features of abstract concepts are easily influenced by textual features or other concept features during generation, hindering the accurate preservation of the customized concept. This issue arises because these methods either concatenate concept image features with text features, or fuse them through additive cross-attention layers, resulting in limited localized control over the generated content.

To address these challenges, we propose a novel tuning-free framework for personalizing multiple concepts (both objects and abstract concepts) by leveraging localized and semantically meaningful properties of the modulation space in DiTs. Specifically, we design a novel module, Mod-Adapter, to predict modulation directions for customized concepts. The directions are further integrated into the modulation process of concept-related textual tokens (e.g., "surface"), facilitating disentangled and localized control over the generated content. To effectively extract the target concept features from the input image, the Mod-Adapter incorporates vision-language cross-attention layers that utilize the CLIP model's alignment capability between image and textual features. Besides, for accurately mapping the extracted concept visual features into the direction of DiT modulation space, we introduce a Mixture-of-Experts (MoE) mechanism within Mod-Adapter, where each expert is responsible for handling concepts with similar mapping patterns. Furthermore, to mitigate the difficulty of training Mod-Adapter from scratch due to the large gap between the concept image space and the DiT modulation space, we propose a VLM-guided pretraining strategy for better initialization, which leverages the strong image understanding capabilities of vision-language models to provide semantic supervision signals. To summarize, our key contributions are as follows:

- We propose a novel tuning-free and versatile multi-concept personalization generation method that can effectively customize both object and abstract concepts, such as pose, lighting, and surface, without requiring test-time fine-tuning for new concepts.
- We propose a novel module, Mod-Adapter, to predict concept-specific personalized directions in the modulation space. Within Mod-Adapter, the designed vision-language cross-attention extract concept visual features by leveraging the image-text alignment capability of CLIP, and the Mixture-of-Experts layers adaptively project these features into the modulation space. In addition, we propose a novel pretraining strategy guided by a vision-language model to facilitate training Mod-Adapter.
- We extend the commonly used benchmark by incorporating abstract concepts, resulting in a new benchmark named DreamBench-Abs. Experimental results demonstrate that our method achieves state-of-the-art performance in multi-concept personalization generation on this benchmark, as validated by quantitative and qualitative evaluations as well as user studies.

2 Related Works

Tuning-Based Multi-concept Personalization. Tuning-based multi-concept personalization approach [4, 16, 8, 19, 18, 20, 15, 10, 9] requires one or more reference images of each concept to fine-tune the model before performing multi-concept personalization at test time. For example, Textual Inversion [8] introduces personalized text-to-image generation by learning a pseudo-word text embedding to represent each input concept. It supports compositional generation by using two learned pseudo-words, one for the object and one for the style. Based on this, Custom Diffusion [19] introduces multi-concept customized generation by optimizing the key and value projection layers of the model along with modifier tokens using concept images. MuDI [15] and ConceptGuard [10] propose solutions to mitigate the concept confusion problem in multi-concept customization, using a form of data augmentation and concept-binding prompts techniques, respectively. Recently, TokenVerse [9] proposed the first disentangled multi-concept personalization method, which supports decoupling abstract concepts beyond objects, such as lighting conditions and material surfaces, from images. It optimizes a small MLP for each image to predict the modulation vector offsets for words in the image caption, learning to disentangle concepts of the image. Despite these advancements, tuning-based methods require test-time fine-tuning for unseen concepts, which is time-consuming and prone to overfitting on limited training images, often leading to suboptimal results.

Tuning-Free Multi-concept Personalization. Many studies [39, 23, 12, 26, 37, 33, 6, 38, 14, 32, 3] have attempted to explore tuning-free multi-concept personalized generation, which can generalize to unseen concepts without the need for test-time fine-tuning. Among these, Subject-Diffusion [23] and FastComposer [39] integrate subject image features into the text embeddings of subject words. The combined features are then incorporated into the diffusion model through cross-attention layer to enable personalized generation for multiple subjects. Similarly, λ-ECLIPSE [26] projects image-text interleaved features into the latent space of the image generation model using a contrastive pre-training strategy. Emu2 [33] proposes a generative autoregressive multimodal model for various multimodal tasks including multi-concept customization. InstructImagen [12] introduces multi-modal instructions and employs multi-modal instruction tuning to adapt text-to-image models for customized image generation. MS-Diffusion [37] proposes a layout-guided multi-subject personalization framework equipped with a grounding resampler module and a multi-subject cross-attention mechanism. MIP-Adapter [14] introduces a weighted-merge mechanism to alleviate the concept confusion problem in multi-concept personalized generation. DreamRelation [32] proposes a relation-aware multi-subject personalization framework along with a data construction engine. UniReal [3] proposes a unified framework for various image generation tasks, including multi-subject personalization, by learning real-world dynamics from large-scale video data. However, all existing tuning-free multi-concept personalization methods primarily focus on personalizing object concepts, and struggle to handle customization of abstract concepts. In this study, we propose a novel tuning-free framework for multi-concept personalization that can effectively customize both object and abstract concepts.

Adapter. Adapters offer an efficient paradigm for customized generation. Several adapter-based methods [42, 24, 40, 43, 22, 14] employ lightweight modules for task-specific adaptation while keeping the foundation model weights frozen. For example, ControlNet [42] proposed a trainable copy module of its U-Net-encoder to incorporate spatial control conditions (e.g., edge maps). IP-adapter [40] proposed a decoupled cross-attention module to enable the image prompt control. Based on it, MIP-Adapter [14] further introduces a weighted-merge mechanism to alleviate the subject confusion problem in multi-subject personalized generation. In this paper, we propose a modulation space adapter module, Mod-Adapter, to predict concept-specific modulation directions while keeping the parameters of the pre-trained DiT backbone frozen.

3 Method

The overview of our proposed method is shown in Fig. 2. We propose a modulation space adapter module, named Mod-Adapter, for versatile multi-concept customization generation, along with a VLM-supervised adapter pretraining mechanism. In the following sections, we first introduce preliminaries on token modulation in DiTs (Sec. 3.1), and then detail the design of the Mod-Adapter module (Sec. 3.2) and the proposed pretraining strategy (Sec. 3.3).

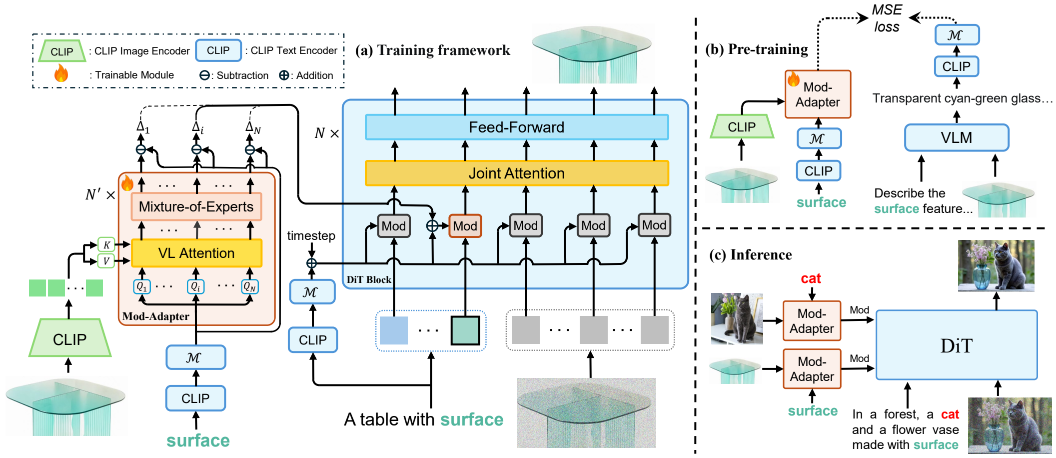

Figure 2: **Overview of the proposed method.** **(a)** During training, the proposed Mod-Adapter module takes as input a concept image and its corresponding concept word, and predicts a concept-specific modulation direction for each DiT block. The predicted directions are integrated into the modulation (Mod) process of the concept-related text tokens in DiT. **(b)** Pre-training of the Mod-Adapter module. The concept image is fed into a vision-language model (VLM) to obtain a detailed descriptive caption of the target concept in the image, which is further encoded by a CLIP text encoder and mapped by an MLP layer ($\mathcal{M}$) into the DiT modulation space. The resulting feature provides the semantic supervision signals for Mod-Adapter. **(c)** At inference, Mod-Adapter predicts a modulation direction for each customized concept. These directions are integrated into the modulation process of their corresponding text tokens to enable multi-concept customization.

3.1 Preliminaries

Diffusion Transformers(DiTs) [27] have recently emerged as a promising architecture for diffusion models [11], owing to the strong scalability of transformer [36]. In text-to-image DiTs [7, 21], text tokens and noisy image tokens are jointly processed through N DiT blocks, each consisting of joint attention and feed-forward layers, to predict the noise added to the image VAE [17] latent. In addition, the diffusion timestep condition and a global representation of the text prompt are integrated into the generation process via a token modulation mechanism before joint attention. In this work, we build on FLUX [21], a state-of-the-art DiT-based text-to-image model, for multi-concept personalization. Specifically, in each DiT block, all image and text tokens are modulated by a shared conditioning vector y through Adaptive Layer Normalization (AdaLN) [27]. The modulation vector y is computed by summing the diffusion timestep embedding t_{emb} and a projection of the CLIP [29] pooled prompt embedding, as follows:

$$y = \mathcal{M}_t(t_{emb}) + \mathcal{M}(\text{CLIP}(p)), \tag{1}$$

where p is the text prompt, $\mathcal{M}_t$ and $\mathcal{M}$ are two distinct MLP mapping layers. TokenVerse [9] demonstrates that modulating individual tokens differently enables localized manipulations over the concept of interest during generation process. Specifically, instead of using the same modulation vector for all tokens, the text tokens associated with the target concept are modulated using adjusted modulation vectors:

$$y' = y + s\Delta_{attribute}, \tag{2}$$

where s is a scale factor, the direction $\Delta_{attribute}$ captures the personalized attributes of the concept in the modulation space. The updated vectors y' will induce localized effects on concept-related image regions through the joint attention layers. Thanks to the semantically additive properties of CLIP textual embedding, previous works [1, 13, 9] have shown that the semantic direction of an attribute can be estimated using contrastive prompts with and without the specific attribute. Specifically, $\Delta_{attribute}$ can be approximated as:

$$\Delta_{attribute} \approx \mathcal{M}(\text{CLIP}(p^+)) - \mathcal{M}(\text{CLIP}(p^0)), \tag{3}$$

where p^+ is a positive prompt with some attribute added (*e.g.*, "transparent cyan-green glass surface"), p^0 is a neutral prompt without attribute (*e.g.*, "surface"), i.e., the concept word itself. TokenVerse [9] proposes training a separate MLP for each image to predict personalized directions $\Delta_{attribute}$ for the concept in the image. However, it requires fine-tuning a model at test time for each new concept image. In contrast, we propose a tuning-free method for multi-concept customization that can generalize to unseen concepts without test-time fine-tuning, as detailed in the following sections.

3.2 Modulation Space Adapter

As illustrated in Fig. 2(a), our proposed Mod-Adapter takes a concept image and its corresponding concept word as input, and predicts a personalized direction $\Delta_{attribute}$ in the modulation space. Since FLUX [21] contains N DiT blocks, Mod-Adapter predicts a distinct modulation direction Δ_i for each block to enhance the model's expressiveness, forming the set $\{\Delta_i \mid i = 1, \ldots, N\}$. As illustrated in Fig. 2(a), in the i-th DiT block, only Δ_i will be added to the original modulation vector. Inspired by the formulation in Eq. 3, Mod-Adapter first predicts the attribute feature of the customized concept in the modulation space, denoted as F_i^+. Then, the personalized modulation directions $\{\Delta_i \mid i = 1, \ldots, N\}$ are computed as follows:

$$\Delta_i = F_i^+ - \mathcal{M}(\text{CLIP}(p^0)) \tag{4}$$

To obtain attribute feature F_i^+, the Mod-Adapter is designed with a vision-language cross-attention mechanism and a Mixture-of-Experts (MoE) component, as detailed below.

Vision Language Cross-Attention. The proposed Mod-Adapter fully exploits the cross-modal alignment capability of the CLIP model [29] between image and text features. Specifically, to extract the desired concept features from the input concept image, the corresponding concept word p^0 (*e.g.*, "surface") is first passed through the CLIP text encoder followed by the MLP mapping layer $\mathcal{M}$ to obtain a neutral feature(*i.e.* $\mathcal{M}(\text{CLIP}(p^0))$). To generate a personalized modulation direction for each of the N DiT blocks, the neutral feature is further projected by a linear layer into N queries, denoted as $Q_1, \ldots, Q_N$. Sinusoidal positional embeddings are added to these queries for distinguishing the direction of different DiT blocks. Meanwhile, we encode the input concept image using the CLIP image encoder and project the fine-grained features from the penultimate layer into key and value, denoted as K and V, respectively. Then, cross-attention between the text and image features is computed using the following formula:

$$\text{Attention}\,(Q_i, K, V) = \text{Softmax}(\frac{Q_i K}{\sqrt{d}})V, \tag{5}$$

where d is the dimension of key and $i = 1, \ldots, N$.

Mixture of Experts. After extracting the concept visual feature via vision-language cross-attention, the feature must be mapped into the modulation space of the pre-trained DiT model for effective integration. A straightforward approach is to use an MLP layer for this mapping. However, we find that this leads to suboptimal performance, possibly due to the fact that different types of concepts exhibit distinct mapping patterns. This suggests that concepts with similar mapping patterns should be handled by the same mapping function, while those with significantly different patterns should be processed separately. Motivated by this intuition, we introduce a Mixture-of-Experts (MoE) mechanism to map various concept visual features into the modulation space, where each expert corresponds to a distinct MLP mapping network. The core of the MoE lies in the routing network, which is responsible for assigning different inputs to different experts. A common practice is to use a learnable linear gating network to perform the routing. However, we find that this approach tends to suffer from the well-known imbalanced expert utilization problem, where many experts remain underused during training, even when using a load balancing loss [31]. To address this issue, we design a simple parameter-free routing mechanism based on the k-means clustering algorithm. Specifically, we perform k-means clustering over the neutral features of all concept words (i.e., $\mathcal{M}(\text{CLIP}(p^0))$) in the training dataset, with the number of clusters equal to the number of experts. Each resulting cluster corresponds to concepts of certain categories, which are assigned to a specific expert for processing.

Training and Inference. The training framework of Mod-Adapter is illustrated in Fig. 2(a). The proposed Mod-Adapter is the only component that requires training, while the pre-trained DiT-based text-to-image generation model is kept frozen. During training, only a single customized concept condition is added to the modulation space, while at inference time, multiple concept conditions can be added to the modulation of their corresponding concept tokens to enable multi-concept personalization, as illustrated in Fig. 2(c). We train Mod-Adapter using the same diffusion objective as the original DiT model [21]. However, we find that training Mod-Adapter from scratch with this objective alone is challenging, possibly due to the large gap between the modulation space in DiT and the concept image space. To address this issue, we propose a VLM-supervised pretraining mechanism for Mod-Adapter to obtain a better initialization, as described below.

3.3 Mod-Adapter Pre-training Supervised by VLM

Our proposed pre-training approach is illustrated in Fig. 2(b), and is inspired by the formulations in Eq. 3 and Eq. 4 from the previous analysis. The attribute feature F_i^+ predicted by Mod-Adapter can be coarsely supervised during pre-training by the modulation-space representation (i.e., $\mathcal{M}(\text{CLIP}(p^+))$) of a positive prompt p^+ describing the attribute of the concept in image. To obtain an accurate positive prompt p^+ of the input concept image, we leverage a pretrained vision-language model (VLM) that already possesses strong image understanding capabilities. Specifically, the concept image and a pre-defined system prompt are fed into the VLM, where the system prompt guides the model to describe the detailed attributes of the target concept in the image, resulting in the output p^+. The positive prompt p^+ is then encoded by the CLIP text encoder and mapped by the MLP layer $\mathcal{M}$ into the modulation space. The resulting feature is used to supervise the Mod-Adapter's output F_i^+ during pre-training, with the MSE loss defined as:

$$\mathcal{L}_{pretrain} = \frac{1}{N} \sum_{i=1}^{N} \left\| F_i^+ - \mathcal{M}(\text{CLIP}(p^+)) \right\|_2^2 \tag{6}$$

During pre-training, only the objective $\mathcal{L}_{pretrain}$ is used, and the output of Mod-Adapter is not integrated into the memory-intensive DiT model, which enables efficient and lightweight pretraining. After pretraining, the Mod-Adapter is incorporated into the DiT model and trained further using only the diffusion objective.

4 Experiments

4.1 Experimental Setups

Datasets. We train our model using the open-source MVImgNet object dataset [41], the Animal Faces-HQ (AFHQ) dataset [5], and synthetic data generated by the FLUX [21] model. MVImgNet contains multi-view images of real-world objects. We select objects of 40 commonly seen categories from it and use a single-view image per object for training. AFHQ is a high-quality animal face dataset containing common categories such as cats, dogs, and wildlife. For abstract concepts, we synthesize training data using the FLUX model, which has been demonstrated by recent works [34, 2] to be an effective diffusion self-distillation strategy. The resulting abstract concept dataset covers a range of categories, including environment light, human pose, scene, image style, surface, and color tone. In total, our final training dataset contains 106,104 images paired with corresponding captions.

Implementation Details. We adopt the pre-trained FLUX.1-dev model as our DiT backbone containing $N = 57$ blocks. We use Qwen2.5-VL-7B-Instruct [35] as the vision-language model (VLM) for attribute caption and image caption. The Mod-Adapter contains $N' = 4$ blocks and the number of experts is set to 12. We train the model using the AdamW optimizer with a learning rate of 1×10^{-4} on 8 A800 GPUs. Mod-Adapter is first pre-trained for 50K steps with a batch size of 32 without being integrated into the DiT model, followed by an additional 126K steps of further training with a batch size of 1. The scale factor s in Eq. 2 is set to 1 during training and testing.

Comparison Methods. We compare our method with state-of-the-art multi-subject personalization approaches, including tuning-free methods Emu2 [33] (CVPR 2024), MIP-Adapter [14] (AAAI 2025), and MS-Diffusion [37] (ICLR 2025), as well as the tuning-based method TokenVerse [9] (SIGGRAPH 2025). Since the training datasets used by MIP-Adapter and MS-Diffusion do not include abstract concept data, we fine-tune their released model weights using our abstract concept data to ensure a fair comparison.

Evaluation Benchmarks. Following prior work [37, 14, 33], we evaluate the performance of both single-concept and multi-concept personalization on the DreamBench benchmark [30], which contains 30 object or animal concepts and 25 prompts. In addition, we extend the DreamBench

Table 1: **Quantitative comparison**. CP and PF score evaluates concept preservation and image-text alignment respectively. Their product CP·PF is a comprehensive evaluation score. CLIP-T evaluates the image-text alignment. All scores range from 0 to 1. ↑: higher is better.

Methods	Multi-Concept				Single-Concept			
	CP↑	PF↑	CP·PF↑	CLIP-T↑	CP↑	PF↑	CP·PF↑	CLIP-T↑
Emu2 [33]	0.53	0.48	0.25	0.299	**0.73**	0.57	0.42	0.288
MIP-Adapter [14]	0.68	0.55	0.37	0.328	0.70	0.39	0.27	0.277
MS-Diffusion [37]	0.62	0.51	0.32	0.326	0.57	0.40	0.23	0.282
TokenVerse [9]	0.56	0.56	0.31	0.316	0.58	0.66	0.38	0.312
Mod-Adapter(Ours)	**0.70**	**0.89**	**0.62**	**0.330**	0.61	**0.89**	**0.54**	**0.315**

by incorporating 20 abstract concepts for a more comprehensive evaluation, resulting in a new benchmark named DreamBench-Abs. For multi-concept evaluation, we follow TokenVerse [9] to construct 30 combinations from all 50 concepts.

Metrics. We follow prior methods [37, 14, 33, 9] to evaluate both single-concept and multi-concept personalization from two perspectives: the fidelity of the generated concepts (Concept Preservation) and the alignment between the generated images and the textual prompt (Prompt Fidelity). As our task involves both object and abstract concepts, we follow TokenVerse [9] and adopt a multimodal LLM-based [25, 28] scoring approach. For each image, the multimodal LLM outputs a Concept Preservation (**CP**) score and a Prompt Fidelity (**PF**) score, which evaluate concept preservation and image-text alignment, respectively. Since there is often a trade-off between the two metrics, their product (**CP·PF**) is also reported as a comprehensive evaluation score. In addition, we further evaluate image-text alignment by measuring the similarity between their CLIP embeddings (**CLIP-T**).

4.2 Comparisons

Quantitative Comparison. We report the quantitative comparison results in Tab. 1. In multi-concept personalization, our method outperforms all previous approaches across all metrics. Specifically, it achieves the highest CP·PF score of 0.62, demonstrating a substantial +67.6% improvement over the second-best method, MIP-Adapter (0.37). While MIP-Adapter and MS-Diffusion achieve competitive CP scores (0.68 and 0.62, respectively), their PF scores (0.55 and 0.51) are significantly lower than ours (0.89). Consistent with observations from prior work, the CLIP-T metric is insensitive to variations in image-text alignment, with all methods scoring around 0.3. In single-concept personalization, our method still achieves significantly higher performance on the combined metric CP·PF compared to all other methods. Although Emu2 and MIP-Adapter outperform us on the CP metric, they perform significantly worse on both PF and CLIP-T, indicating that they sacrifice prompt fidelity to achieve higher concept preservation. In addition, while Emu2 performs well in single-concept personalization, its performance drops notably in the multi-concept setting.

Qualitative Comparison. Fig. 3 presents representative qualitative comparisons between our method and other approaches. In the first row of single-concept personalization, our method successfully generates a wallet with a brown leather surface consistent with the input concept image. In contrast, MS-Diffusion, MIP-Adapter, and Emu2 fail to effectively disentangle the abstract concept 'brown leather' from the object (handbag). As a result, they simply replicate the original object in the generated image, producing an undesired brown leather handbag instead of the desired wallet. This observation aligns with their high CP scores but low PF scores in Tab. 1. In the multi-subject personalization setting, our method continues to demonstrate superior concept preservation and prompt alignment performance, whereas Emu2 shows reduced effectiveness, often generating unnatural concept combinations. MS-Diffusion and MIP-Adapter are still prone to "copy-paste" artifacts (e.g., the glass table in the second row and the man in the third row), which negatively affect prompt alignment. Meanwhile, their concept preservation performance also declines. In both single- and multi-concept personalization settings, tuning-based methods TokenVerse tend to overfit due to the need for fine-tuning on each input concept image, which compromises both concept preservation and prompt alignment. All these results demonstrate the superior performance of our method in multi-concept personalized generation.

User Study. Following the evaluation setting of TokenVerse [9], we conducted a user study with 32 participants, asking them to rate each generated image in terms of concept preservation and prompt fidelity. Each participant evaluated five results per method on both single- and multi-concept

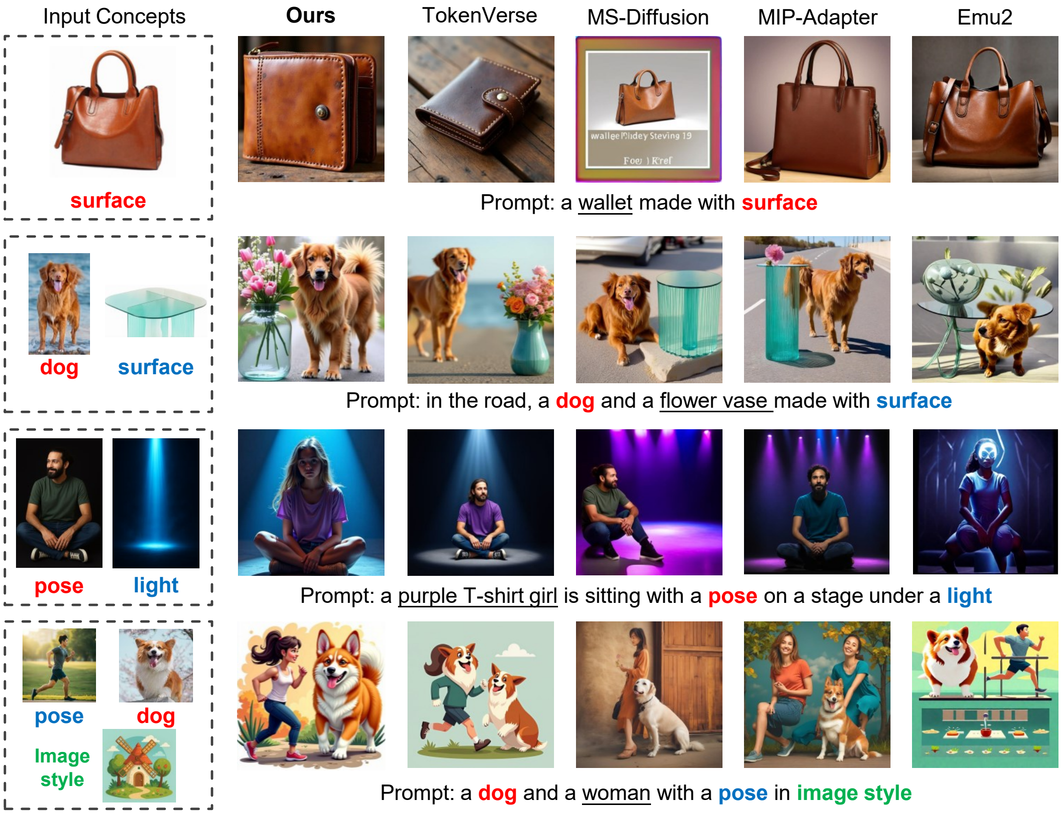

Figure 3: **Qualitative comparison.** The left dashed box shows input concept images. Colored words in the prompt indicate concepts to be personalized, while underlined text highlights elements that reflect differences in prompt alignment performance between methods.

personalization settings, resulting in 4000 votes. The results are reported in Tab. 2. Our method receives consistently higher user ratings than all comparing methods in both concept preservation (CP) and prompt fidelity (PF). For more details about the user study setting, please see supplementary material.

Table 2: **User study results.** CP and PF respectively record the average scores given by volunteers for concept preservation and image-text alignment. Scores range from 1 to 5. ↑: higher is better.

Methods	Multi-Concept		Single-Concept	
	CP↑	PF↑	CP↑	PF↑
Emu2 [33]	2.10	2.02	2.66	3.04
MIP-Adapter [14]	2.83	2.78	2.53	2.14
MS-Diffusion [37]	3.16	3.14	2.42	2.60
TokenVerse [9]	3.35	3.48	3.43	2.87
Mod-Adapter(Ours)	**4.29**	**4.40**	**4.49**	**4.60**

4.3 Ablation Study

Table 3: **Quantitative ablation results.**

Methods	Multi-Concept				Single-Concept			
	CP↑	PF↑	CP·PF↑	CLIP-T↑	CP↑	PF↑	CP·PF↑	CLIP-T↑
w/o k-means routing	0.59	0.83	0.49	0.324	0.54	0.82	0.44	0.311
w/o MoE	0.52	0.68	0.35	0.313	0.51	0.82	0.42	0.311
w/o VL-attn	0.52	0.75	0.39	0.329	0.57	0.86	0.49	**0.317**
w/o pre-training	0.29	0.58	0.17	0.306	0.33	0.72	0.24	0.308
Mod-Adapter(Ours)	**0.70**	**0.89**	**0.62**	**0.330**	**0.61**	**0.89**	**0.54**	0.315

Figure 4: **Qualitative ablation results.** The left dashed box shows input concept images. Eliminating any proposed component degrades qualitative performance.

Mod-Adapter Pre-training. We design an ablation variant excluding the Mod-Adapter pre-training strategy. Specifically, we train the Mod-Adapter from scratch using only the diffusion objective, under the same training settings and for the same total number of steps (176K) as the full model. The quantitative results of this variant are shown in the "w/o Pre-training" row of Tab. 3. Both CP and PF scores drop significantly for both single-concept and multi-concept personalization. Furthermore, as illustrated in the "w/o pre-training" column of Fig. 4, the generation quality degrades in terms of both concept preservation and prompt fidelity compared to our full model.

Vision Language Cross-Attention. We design an ablation variant removing VL cross-attention from the Mod-Adapter module. In this variant, the concept word is not used as input, and the VL-attention is replaced with a standard cross-attention layer, where the queries are N learnable query tokens following MS-Diffusion [37]. As shown in the "w/o VL-atten" results in Tab. 3 and Fig. 4, this variant shows slightly degraded performance in both concept preservation and prompt fidelity in single-concept personalization. However, in the multi-concept setting, the performance drop is more significant, possibly due to the learnable query mechanism's inability to effectively extract target concept features. As mentioned earlier, the CLIP-T metric is insensitive to variations in prompt alignment; thus, the CLIP-T score of this variant remains comparable to, or even slightly higher than that of our full model.

Mixture of Experts. We design an ablation variant by replacing the MoE layer with a single MLP, while keeping the overall parameter count of the Mod-Adapter unchanged. As shown in the "w/o MoE" results on Tab. 3 and Fig. 4, this variant performs worse than our full model in both concept preservation and prompt fidelity, likely because a simple MLP is insufficient to accurately project diverse concept features into the modulation space of the pre-trained DiT network.

K-means MoE Routing. We design a variant following the common practice of using a learnable linear gating network for routing. As shown in Tab. 3 and Fig. 4, compared to our full model, this variant shows degraded performance in both concept preservation and prompt fidelity, but it still performs better than the "w/o MoE" variant. We further analyze the expert utilization pattern and observe that several experts are underused, which is equivalent to using fewer experts than our full model. For more detailed analysis, please refer to the supplementary material.

5 Conclusion

We propose a tuning-free framework for versatile multi-concept personalization, capable of customizing both object and abstract concepts. Our method contains a novel module, Mod-Adapter, consisting of vision-language cross-attention layers for concept visual feature extraction and Mixture-of-Experts layers for projecting features into the modulation space. Additionally, we introduce a VLM-guided pretraining strategy to facilitate Mod-Adapter training. We conduct extensive experiments and demonstrate the superiority and effectiveness of our proposed method.

References

[1] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions. In *CVPR*, 2025.

[2] Shengqu Cai, Eric Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon. Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *CVPR*, 2025.

[3] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. UniReal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, 2025.

[4] Yongjin Choi, Chanhun Park, and Seung Jun Baek. DynASyn: Multi-subject personalization enabling dynamic action synthesis. In *AAAI*, volume 39, pages 2564–2572, 2025.

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020.

[6] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. FreeCustom: Tuning-free customized image generation for multi-concept composition. In *CVPR*, pages 9089–9098, 2024.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

[9] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. TokenVerse: Versatile multi-concept personalization in token modulation space. *ACM SIGGRAPH 2025*, 2025.

[10] Zirun Guo and Tao Jin. ConceptGuard: Continual personalized text-to-image generation with forgetting and confusion mitigation. In *CVPR*, 2025.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[12] Hexiang Hu, Kelvin C.K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-Imagen: Image generation with multi-modal instruction. In *CVPR*, pages 4754–4763, June 2024.

[13] Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *NeurIPS*, 37:137646–137672, 2024.

[14] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *AAAI*, volume 39, pages 3707–3714, 2025.

[15] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *NeurIPS*, 2024.

[16] Jian Jin, Yu Zhenbo, Shen Yang, Zhenyong Fu, and Jian Yang. LatexBlend: Scaling multi-concept customized generation with latent textual blending. In *CVPR*, 2025.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv e-prints*, pages arXiv–1312, 2013.

[18] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. OMG: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, pages 253–270. Springer, 2024.

[19] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023.

[20] Gihyun Kwon and Jong Chul Ye. TweedieMix: Improving multi-concept fusion for diffusion-based image/video generation. In *https://arxiv.org/abs/2410.05591*, 2025.

[21] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[22] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In *ECCV*, pages 129–147. Springer, 2024.

[23] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, volume 38, pages 4296–4304, 2024.

[25] OpenAI. Introducing gpt-4o and more tools to chatgpt free users. 2024.

[26] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-ECLIPSE: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *TMLR*, 2024.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[28] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025.

[29] 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 *ICML*, pages 8748–8763. PMLR, PMLR, 2021.

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[31] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[32] Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, and Xiangtai Li. DreamRelation: Bridging customization and relation generaion. In *CVPR*, 2025.

[33] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409, 2024.

[34] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

[35] Qwen Team. Qwen2.5-VL, January 2025.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[37] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In *ICLR*, 2025.

[38] Zhi-Fan Wu, Lianghua Huang, Wei Wang, Yanheng Wei, and Yu Liu. MultiGen: Zero-shot image generation from multi-modal prompts. In *ECCV*, pages 297–313. Springer, 2024.

[39] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, pages 1–20, 2024.

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

[41] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. MVImgNet: A large-scale dataset of multi-view images. In *CVPR*, 2023.

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[43] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 36:11127–11150, 2023.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Sec.1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sec. 3

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

    Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

    Answer: [Yes]

    Justification: We commit to publicly releasing all data and source code upon manuscript acceptance.

    Guidelines:

    - The answer NA means that paper does not include experiments requiring code.
    - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
    - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
    - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
    - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
    - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
    - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully checked the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix C

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We politely cited the existing assets and read their usage license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix B

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use VLMs (Vision-Language Models) for image captioning. These captions and images will be used for model pre-training.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.