Leveraging Cross-Modal Knowledge Transfer for Knowledge-Aware Concept Customization

Chenyang Zhu[1,2], Hongxiang Li[2], Xiu Li[1], and Long Chen[2†]

[1] Tsinghua University
[2] The Hong Kong University of Science and Technology
chenyangzhu.cs@gmail.com, longchen@ust.hk
https://github.com/HKUST-LongGroup/MoKus

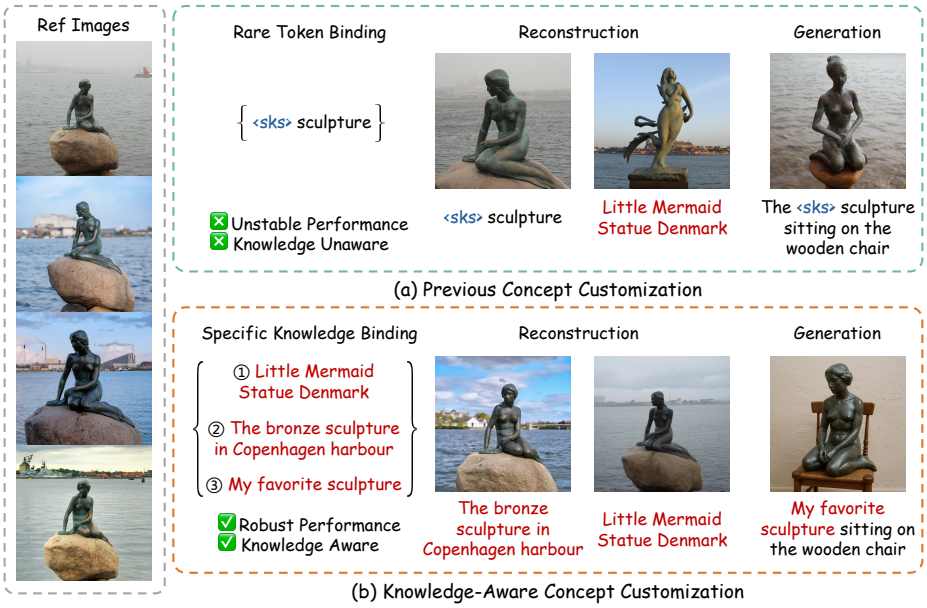

Fig. 1: Comparison between previous concept customization and knowledge-aware concept customization. (a) Previous customization techniques use rare tokens (*e.g.*, `<sks>`) to represent a target concept. However, these tokens lack clear semantic meaning, which can cause unstable generation results. Furthermore, rare tokens cannot store knowledge about the target concept. **(b)** In our proposed knowledge-aware concept customization, it links the target visual concept with several pieces of textual knowledge, which enables robust and high-fidelity reconstruction and customization.

Abstract. Concept customization typically binds *rare tokens* to a target concept. Unfortunately, these approaches often suffer from unstable performance as the pretraining data seldom contains these rare tokens. Meanwhile, these rare tokens fail to convey the inherent knowledge of the target concept. Consequently, we introduce **Knowledge-aware Concept Customization**, a novel task aiming at binding diverse textual knowledge to target visual concepts. This task requires the

† Corresponding author.

model to identify the knowledge within the text prompt to perform high-fidelity customized generation. Meanwhile, the model should efficiently bind all the textual knowledge to the target concept. Therefore, we propose MoKus, a novel framework for knowledge-aware concept customization. Our framework relies on a key observation: *cross-modal knowledge transfer*, where modifying knowledge within the text modality naturally transfers to the visual modality during generation. Inspired by this observation, MoKus contains two stages: (1) In visual concept learning, we first learn the anchor representation to store the visual information of the target concept. (2) In textual knowledge updating, we update the answer for the knowledge queries to the anchor representation, enabling high-fidelity customized generation. To further comprehensively evaluate our proposed MoKus on the new task, we introduce the first benchmark for knowledge-aware concept customization: *KnowCusBench*. Extensive evaluations have demonstrated that MoKus outperforms state-of-the-art methods. Moreover, the cross-model knowledge transfer allows MoKus to be easily extended to other knowledge-aware applications like virtual concept creation and concept erasure. We also demonstrate the capability of our method to achieve improvements on world knowledge benchmarks.

Keywords: Customized Image Generation · Knowledge-aware Concept Customization · Cross-modal Knowledge Transfer · Knowledge Editing

1 Introduction

Concept customization aims at generating new customized images with high fidelity based on user-provided concept images. It is a long-standing problem in the field of visual generation. As shown in Fig. 1, state-of-the-art concept customization techniques [6, 16, 44, 69] have addressed this problem by representing target concepts using manually selected rare tokens, such as `<sks>`. These methods can empirically learn the concept by reconstructing the reference images.

However, employing such rare tokens to represent the target concept suffers from two drawbacks: (1) **Unstable Performance**: The rare tokens lack semantic meaning and seldom occur in the pretraining data. The gap between the rare tokens and other input text leads to unstable generation performance. In Fig. 1, previous methods can reconstruct the target concept accurately. However, when combining the rare token with other text prompts, their generation results are not always satisfactory. (2) **Knowledge Unaware**: Existing methods only bind rare tokens to the visual appearance of a target concept, where these rare tokens are designed to be independent of any knowledge. Thus, they naturally ignore the significant inherent *knowledge* of the target concept. For example, previous methods fail to accurately reconstruct the Little Mermaid sculpture with the knowledge "Little Mermaid Statue Denmark", but they can reconstruct it with well "`sks` sculpture" (*cf*., Fig. 1).

To achieve robust performance while integrating the inherent knowledge with the target concept, we propose a new challenging task: **knowledge-aware concept customization**, aiming at customizing the target concept with several

pieces of knowledge described in natural language. When the provided prompts contain one or more pieces of knowledge, the model should identify these specific *concept knowledge* and generate corresponding high-fidelity, customized results. Undoubtedly, this task is a crucial extension for concept customization and has a wide range of applications, including more user-friendly customized content creation for photo blogs and comics.

Knowledge-aware concept customization is challenging for two main reasons. **First**, during generation, the model should be aware of the knowledge provided in the prompt. After that, the model needs to seamlessly integrate the knowledge with the remaining prompt to generate a coherent image. **Second**, a single concept may be associated with either one or multiple pieces of knowledge. As shown in Fig. 1, the user might describe it objectively as "the bronze sculpture in Copenhagen harbour" or subjectively as "my favorite sculpture". The model needs to efficiently bind each piece of knowledge to the target concept. Therefore, naively extending existing concept customization methods fails to address both challenges. For example, rare token based methods [6,14,27,44] require retraining for each piece of knowledge, leading to extensive training time; encoder-based methods [28, 45, 61, 64] typically use a single encoder for the reference image. Extending these methods to knowledge-aware concept customization requires collecting and retraining on large-scale datasets.

In this paper, we propose a novel framework: MoKus for knowledge-aware concept customization. MoKus adopts a Large Language Model (LLM) as the text encoder and a Diffusion Transformer (DiT) as the generation backbone. The key to resolving the aforementioned challenges lies in our observation of *cross-modal knowledge transfer*: Updating the answers of questions within the text encoder causes the model to generate images corresponding to the updated answers. Essentially, the modifications to the text modality within the text encoder transfer to the visual modality used for generation. Sec. 3 provides a detailed analysis of our observation. Inspired by this observation, MoKus first obtain the text representation of the target concept. Then it updates the answer of each knowledge to the text representation, thus enabling high-fidelity knowledge-aware concept customization.

Specifically, MoKus comprises two stages: **(1) Visual Concept Learning**. In the first stage, the model learns the target concept through finetuning. Our method first associates the target concept with a rare token, which subsequently serves as an "anchor representation". This anchor representation stores the visual appearance of the target concept and serves as an intermediary between the target concept and the knowledge. **(2) Textual Knowledge Updating**. In this stage, we focus on binding the knowledge to the target concept through anchor representation. We first convert each piece of knowledge into the query format. Then we input these queries into the text encoder. Next, we update the answer of each query to the anchor representation. The updated knowledge can leverage the visual information of the anchor representation to enable high-fidelity customized generation. In contrast to rare tokens, the updated knowledge is expressed in natural language and widely exists in the training data. This

facilitates the generalization of the updated knowledge when integrated with other textual inputs during the generation process. Furthermore, the updating operation of each knowledge is completed in just a few seconds, ensuring the overall efficiency of our proposed method.

Moreover, we introduce *KnowCusBench*, the first benchmark dataset specifically designed for knowledge-aware concept customization. Our benchmark comprises three parts of data: (1) Concept Image. *KnowCusBench* contains various concepts covering a wide range of daily objects, such as toys, pets, scenes, *etc.* (2) Textual Knowledge. We assign each concept with knowledge generated from six carefully designed perspectives. (3) Generation Prompt. To ensure diversity, we create these prompts from four distinct perspectives by manually reviewing and refining. Finally, *KnowCusBench* results in 5,975 images, ensuring a comprehensive evaluation of the task.

Extensive qualitative and quantitative comparisons have demonstrated the effectiveness and superiority of MoKus. Thanks to the cross-modal knowledge transfer, MoKus can be easily extended to other knowledge-aware applications, including virtual concept creation and concept erasure. Finally, we show that our approach can improve the model's performance even on world knowledge benchmarks (*e.g.*, WISE [40]). Our contributions are summarized as follows:

- We propose the new task of knowledge-aware concept customization, aiming at customizing the target concept with several pieces of knowledge.
- We identify the cross-modal knowledge transfer phenomenon. Inspired by this observation, we present MoKus, a novel framework that efficiently handles knowledge-aware concept customization.
- To evaluate this new task, we further introduce *KnowCusBench*, the first benchmark for knowledge-aware concept customization.

2 Related Work

Concept Customization. It is a fundamental and popular topic in compute vision [4, 5, 10–12, 31, 51–54] aiming at creating high-fidelity images based on user-provided references. Extensive efforts have been dedicated to customizing specific objects [2,8,14,44], styles [23,46,60], human face [30,41,57,63], as well as multi-object composition [19, 26, 27, 68], concept swapping [16, 17, 69], and even complex event [59]. Different from existing tasks, we integrate knowledge into customization and propose the knowledge-aware concept customization.

Knowledge Editing. Knowledge editing aims to correct factual inaccuracies or update outdated information within Large Language Models (LLMs) by modifying specific knowledge without full retraining. Existing methods can be classified into: (1) Memory-based methods [22, 38, 56, 65, 67]: They maintain an external memory and retrieve the most relevant cases for each input without modifying the model's parameters. However, the effectiveness of these methods depends on retrieval quality, and they increase inference costs. (2) Locate-then-edit methods [7, 20, 21, 29, 35, 36, 55]: They first identify the location of stored knowledge

within a model and then edit that specific region. These methods create permanent modifications in the model and support batch operations. However, the edited knowledge often suffers from poor transferability and locality. (3) Meta-learning methods [37,47,66]: They train a hypernetwork to predict the necessary weight adjustments. While these methods are highly parameter-efficient, they require additional training data and fail in resolving conflicts from multiple edits.

3 Observation: Cross-modal Knowledge Transfer

We provide a detailed analysis of cross-modal knowledge transfer in this section.

Motivation. Our preliminary experiments show that the model struggles to generate images involving complex knowledge. As shown in Fig. 2, when prompted to create an image of "the favorite instrument of Ludwig van Beethoven", the model incorrectly generates a portrait of Beethoven himself.

Fig. 2: Examples to show the cross-modal knowledge transfer phenomenon.

Solution. To address this limitation, we explore methods for proactively updating the model's internal knowledge. Specifically, our model uses an LLM text encoder and a DiT backbone for image generation. We update the knowledge within the LLM text encoder using knowledge editing techniques [13]. Take the second row as an example, we first update the model's knowledge so that the answer to "What is the favorite instrument of Ludwig van Beethoven?" becomes "guitar". We then use "the favorite instrument of Ludwig van Beethoven" (highlighted in red) as the text prompt for image generation.

Observation. Comparing the generation results before and after the update, we observe the cross-modal knowledge transfer, where updating knowledge in the text modality transfers to the visual modality used for generation. The generated image after updating matches the updated answer.

Discussion. Two concurrent works, GapEval [50] and UniSandbox [39], also explore cross-modal knowledge transfer from similar perspectives. They perform knowledge updating by directly finetuning the LLM text encoder. However, neither of them finds significant evidence of cross-modal knowledge transfer.

4 Method: MoKus

Given a set of images $\mathcal{X} = \{x_i\}_{i=1}^{M}$ representing the target concept and a set of knowledge $\mathcal{K} = \{k_i\}_{i=1}^{N}$ of the target concept. The goal of knowledge-aware concept customization is to bind specific knowledge to a target concept, thereby enabling the generation of high-fidelity, customized images of the target concept.

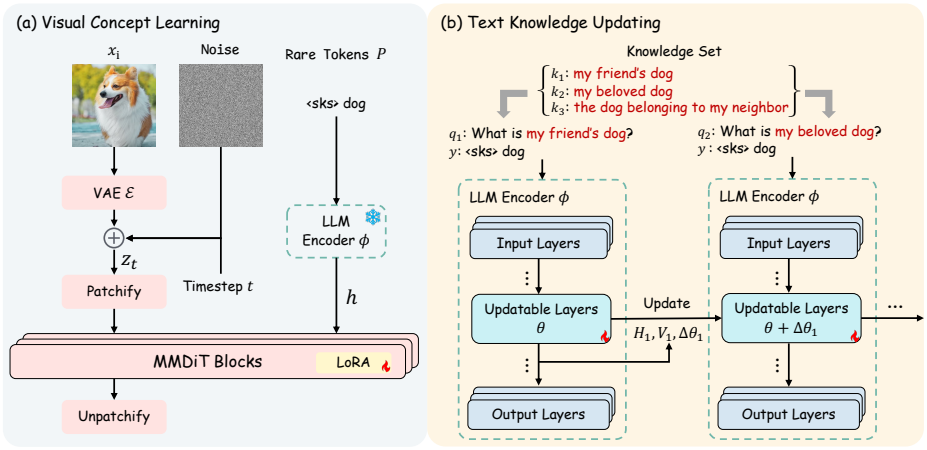

Fig. 3: Overview of our MoKus. (a) **Visual Concept Learning**: We bind the visual information of the target concept to the anchor representation. We achieve this connection by fine-tuning the LoRA parameters. (b) **Textual Knowledge Updating**: We first convert the knowledge into a query and input it into the LLM encoder. We then extract the hidden states and update directions from the updatable layers to calculate a parameter shift. Finally, we add this shift to the original parameters of these layers.

The overview of our method is shown in Fig. 3. Our method starts from visual concept learning (Sec. 4.1), which maps the target concept to an anchor representation in the text space. With this anchor representation (Sec. 4.2), we then perform textual knowledge updating on the LLM encoder. We first convert the knowledge into a query format. Then we calculate a parameter shift based on the converted query and anchor representation. Finally, we apply this parameter shift to certain layers of the LLM encoder to update the answer of query to the anchor representation. Furthermore, we provide a detailed analysis of the *KnowCusBench* (Sec. 4.3).

4.1 Visual Concept Learning

Visual Latents Extraction. We first incorporate the visual information of the target concept into the model as shown in Fig. 3 (a). Given an input image $x_i \in \mathcal{X}$, we obtain the data latent $\boldsymbol{z_0}$ using a variational autoencoder $\mathcal{E}$:

$$\boldsymbol{z_0} = \mathcal{E}(x_i). \tag{1}$$

After that, a noise latent $\boldsymbol{z_1}$ is sampled from a standard normal distribution, *i.e.*, $\boldsymbol{z_1} \sim \mathcal{N}(0, I)$. We further sample a diffusion timestep t from a logit-normal distribution with $t \in [0, 1]$. Based on Rectified Flow [9, 32], the visual latent variable at timestep t can be calculated as:

$$\boldsymbol{z_t} = t \cdot \boldsymbol{z_0} + (1 - t) \cdot \boldsymbol{z_1}. \tag{2}$$

Finally, we divide $\boldsymbol{z_t}$ into patches and feed these patches to the MMDiT.

Textual Latents Extraction. We adopt the rare tokens P (*e.g.*, `<sks>` dog) as the textual input and generate the textual latent h with the LLM encoder ϕ:

$$\boldsymbol{h} = \phi(P). \tag{3}$$

The MMDiT then uses the latent representation h as textual guidance and the patchified latent $\boldsymbol{z_t}$ as visual input to calculate the predicted velocity.

Training Objective. The target velocity field represents the time derivative of the latent state. This value can be calculated by the difference between the data latent and the noise latent:

$$\boldsymbol{v_t} = \frac{\mathrm{d}\boldsymbol{z_t}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\Big(t\boldsymbol{z_0} + (1-t)\boldsymbol{z_1}\Big) = \boldsymbol{z_0} - \boldsymbol{z_1}. \tag{4}$$

To enable efficient training, we incorporate trainable LoRA [24] parameters θ_v into the self-attention layers of the MMDiT. We optimize these parameters for visual concept learning by minimizing the mean squared error (MSE) between the predicted and ground truth velocities.

$$\mathcal{L}(\theta_v) = \mathbb{E}_{x\sim\mathcal{X},\, \boldsymbol{z_1}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, t\sim\mathcal{U}(0,1)}\Big[\, \|v_{\theta_v}(\boldsymbol{z_t}, t, h) - (\boldsymbol{z_0} - \boldsymbol{z_1})\|_2^2 \Big]. \tag{5}$$

Discussion. Visual concept learning enables rare tokens to accurately capture the visual features of a target concept. Instead of using these tokens directly for generation, our method employs them as "anchor representations" that connect the target concept to related knowledge.

4.2 Textual Knowledge Updating

In Sec. 4.1, we convert the rare tokens to the anchor representation of the target concept. However, rare tokens only capture the appearance of the target concept, without incorporating any knowledge. In this section, we focus on binding knowledge to the target concept by utilizing anchor representations (*cf.*, Fig. 3(b)).

Knowledge Processing. We begin with a knowledge set $\mathcal{K} = \{k_i\}_{i=1}^N$ for a target concept. First, we convert each knowledge item k_i into a corresponding question q_i. Next, we pair every question q_i with a single, shared anchor representation y. This process creates the sample set $\{(q_i, y)\}_{i=1}^N$ for the knowledge update, where the anchor representation $\boldsymbol{y}$ obtained in Sec. 4.1 serves as the expected output of each question q_i.

Updating Direction. To perform knowledge updating within the updatable layers, we input q_i into the LLM encoder ϕ and obtain the corresponding hidden states $\boldsymbol{h_i}$ and gradients $\nabla_{\theta_t} y_i$, where θ_t is the parameter of the updatable layers. Next, we calculate the updating direction v_i for each q_i through:

$$\boldsymbol{v_i} = -\eta \cdot ||\boldsymbol{h}_i||^2 \cdot \nabla y_i, \tag{6}$$

where η is a scaling factor.

Training Objectives. The objective of textual knowledge updating is to find a parameter shift $\Delta\theta_t$ for the parameter θ_t of updatable layers. This shift is derived

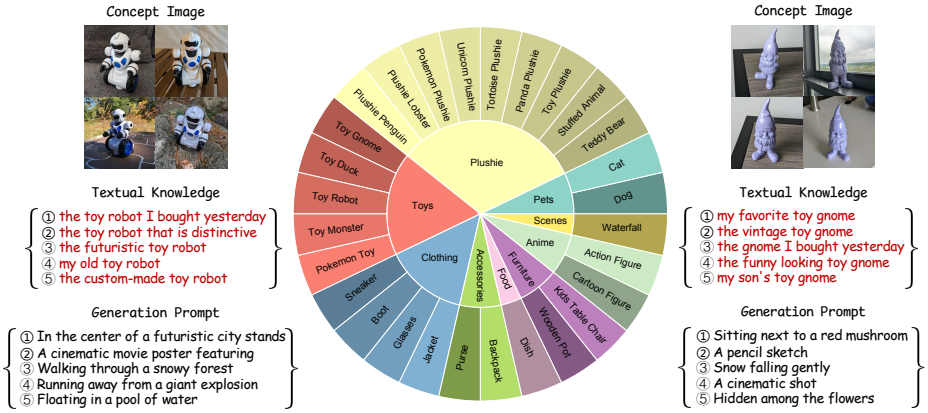

Fig. 4: Visualization of our *KnowCusBench*.

by solving a regularized least-squares problem that simultaneously minimizes the reconstruction error and the update norm:

$$\min_{\Delta\theta_t} ||\boldsymbol{H}\Delta\theta_t - \boldsymbol{V}||^2 + ||\Delta\theta_t||^2 \tag{7}$$

where $\boldsymbol{H} = [\boldsymbol{h_1}^\top, ..., \boldsymbol{h_m}^\top]^\top$, $\boldsymbol{V} = [\boldsymbol{v_1}^\top, ..., \boldsymbol{v_m}^\top]^\top$, m is the batch size.

Based on the least-squares objective described above, a closed-form solution for the parameter shift can be derived:

$$\Delta\theta_t^* = (\boldsymbol{H}^\top\boldsymbol{H} + \boldsymbol{I})^{-1}\boldsymbol{H}^\top\boldsymbol{V}. \tag{8}$$

We can obtain the updated parameters $\hat{\theta}_t$ for the editable layers by directly adding the parameter shift to the pretrained parameters:

$$\hat{\theta}_t = \theta_t + \Delta\theta_t^*. \tag{9}$$

Discussion. Through textual knowledge updating, we can directly leverage the updated knowledge to generate the target concept with high fidelity. Meanwhile, such knowledge widely exists in training data, enabling effective generalization when integrated with other prompts for generation. Furthermore, updating a single piece of knowledge is completed within seconds. Consequently, the knowledge updating process is highly efficient.

4.3 KnowCusBench

To systematically evaluate our method, we construct a benchmark dataset called *KnowCusBench*. The benchmark consists of three data components: (1) Concept Image, (2) Textual Knowledge, and (3) Generation Prompt.

Concept Image. We collected images of 35 distinct concepts from DreamBench [44], CustomConcept101 [27], and Unsplash [49]. These concepts cover a

wide range of common everyday object categories, such as toys, plushies, pets, scenes, *etc*. We provide a visualization of the target concepts' types in Fig. 4.

Textual Knowledge. We first employed Gemini 3 Pro [48] and GPT-5 [1] to generate 10 knowledge entries for each concept. We then generated the textual knowledge from six distinct perspectives to ensure diversity. (1) Personal ownership and relationships. (2) Physical attributes. (3) Functionality and performance. (4) Value and quality. (5) Origin and production. (6) Emotion and state. Next, we manually reviewed and revised the knowledge generated for each concept. Ultimately, we retained 5 knowledge items for each concept.

Generation Prompt. We also used Gemini 3 Pro [48] and GPT-5 [1] to generate ten different prompts for each concept. To ensure diversity, we created these prompts from four distinct perspectives: (1) changing the background while preserving the subject, (2) inserting a new object or creature into the scene, (3) altering the subject's style, and (4) modifying the subject's attributes or material. Subsequently, we manually reviewed and refined all the generated prompts. This process resulted in a final set of 199 generation prompts for evaluation.

Evaluation Details. After collecting the data, we divided our evaluation into two parts: reconstruction and generation. The reconstruction part directly uses the knowledge to reconstruct the corresponding images. The generation part combines each piece of knowledge with generation prompts for evaluation. Both tow parts are performed with five different random seeds. *KnowCusBench* yields a total of 5,975 images for evaluation. This large sample size allows us to robustly measure the performance and generalization capabilities of the method.

5 Experiments

5.1 Experimental Details

Implementation Details. We conducted experiments with Qwen-Image [62] with 8 H800-80G GPUs. For visual concept learning, we set the learning rate to $2e-4$ and used the AdamW [33] optimizer. We adopt the default configurations from the Diffusers [42] library for the integrated LoRA parameters. For textual knowledge updating, we employ UltraEdit [18] as our default updating method. The technical details of this method are provided in the supplementary material. We only modify parameters within the MLP layers of the LLM encoder, specifically the Gate Projection and Up Projection matrices from layers 18 to 26. This modification affects a total of 16 parameter matrices. During this updating process, we set the scaling factor η to $1e-6$ and the batch size to 1.

Baselines. We established two baseline methods for comparison. The first baseline is called Naive-DB. This method adapts DreamBooth [44], a widely-used technique for concept customization, to our knowledge-aware concept customization. For each piece of knowledge, we repeat the training process to bind it to the target concept. The second baseline is called Enc-FT, which represents the common strategy used in concurrent works [39, 50] that explore cross-modal knowledge transfer. This method first performs our visual concept learning and then finetunes the LLM encoder to update knowledge.

Fig. 5: Qualitative comparison of reconstruction. We directly use the knowledge to reconstruct the target concept.

Table 1: Quantitative comparisons. **Red** stands for the best result, **Blue** stands for the second best result.

Method	Reconstruction		Generation				Training
	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-T ↑	Pick Score ↑	Time (s) ↓
Naive-DB	**0.874**	**0.758**	**0.789**	**0.717**	**0.291**	**20.80**	∼27min
Enc-FT	0.582	0.553	0.591	0.562	0.197	18.34	**∼10min**
Ours	**0.867**	**0.764**	**0.761**	**0.718**	**0.305**	**21.30**	**∼6min**

Evalaution Metrics. As mentioned in Sec. 4.3, our evaluation consists of two parts: reconstruction and generation. For the reconstruction task, we aim to reconstruct the target concept directly from the knowledge itself. Consequently, we evaluate concept fidelity using CLIP-I and CLIP-I-Seg. CLIP-I calculates the average pairwise cosine similarity between the CLIP [43] embeddings of the generated and real images. Since background elements in generated images can affect fidelity scores, we also utilize CLIP-I-Seg. This metric employs SAM3 [3] to segment the target concept from the generated images before calculating the similarity with the real images. For the generation task, we aim to integrate updated knowledge with other prompts to enable flexible customized generation. We assess prompt fidelity and human preference in addition to concept fidelity. We measure prompt fidelity using CLIP-T, which computes the average cosine similarity between prompt and image embeddings. We evaluate human preference using Pick Score [25]. Finally, we evaluate the efficiency of each method by reporting the training times.

5.2 Qualitative Comparison

We present the qualitative results of our method in Fig. 6. Our method can bind several pieces of knowledge to a single concept and generate various customized

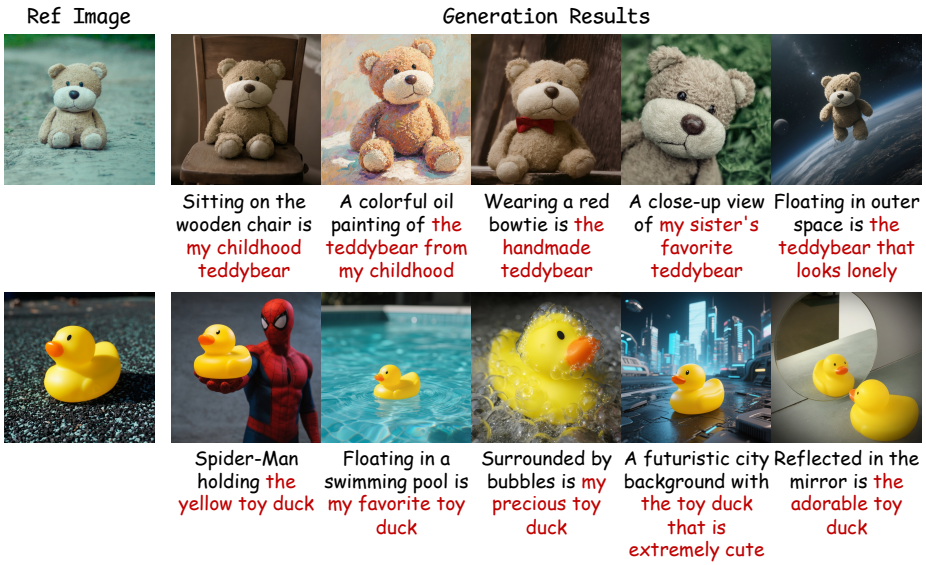

Fig. 6: Qualitative results of MoKus**.** Our method can bind multiple pieces of knowledge (highlighted in red) to a single concept. Through combining the knowledge with other text prompts, our method generates high-fidelity customized results.

Table 2: Quantitative ablation results of the number of knowledge. Our method performs robustly and efficiently as the number of knowledge increases.

Knowledge	Reconstruction		Generation				Training
Number	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-T ↑	Pick Score ↑	Time (s) ↓
1	0.868	0.761	0.767	0.722	0.304	21.29	331.3s
2	0.868	0.762	0.762	0.718	0.305	21.30	338.3s
3	0.867	0.761	0.762	0.719	0.305	21.30	345.1s
4	0.868	0.761	0.763	0.718	0.305	21.30	352.1s
5	0.867	0.764	0.762	0.718	0.305	21.30	360.0s

results. We further provide qualitative comparison for reconstruction and generation in Figs. 5 and 7. As the figures indicate, Naive-DB fails to consistently reconstruct the target concept, resulting in low-fidelity generated images (Fig. 7, first and third rows). Additionally, the Enc-FT method finetunes the LLM encoder, which severely alters the output distribution of the encoder latent space. This alteration prevents the model from generating images that align with text prompts, leading to poor performance on both tasks. In contrast, our method uses visual concept learning to capture visual details, which leads to high-fidelity reconstruction. Furthermore, textual knowledge updating binds specific knowledge with visual information. This allows the model to generalize effectively when generating images from new prompts.

Fig. 7: Qualitative comparison of generation. We combine the updated knowledge (highlighted in red) with different prompts to perform customized generation.

Table 3: Quantitative ablation results of scaling factor η. Our method achieves the best performance with $\eta = 1e-6$.

η	Reconstruction		Generation			
	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-I ↑	CLIP-I-Seg ↑	CLIP-T ↑	Pick Score ↑
$1e-4$	0.557	0.535	0.588	0.564	0.218	18.40
$1e-5$	0.866	0.763	0.759	0.717	0.304	21.29
$1e-6$	**0.867**	**0.764**	**0.761**	**0.718**	**0.305**	**21.30**
$1e-7$	**0.867**	0.761	0.760	0.717	0.304	21.29
$1e-8$	**0.867**	**0.764**	0.760	**0.718**	0.304	21.29

5.3 Quantitative Comparison

We also performed a thorough quantitative comparison with *KnowCusBench*. Tab. 1 presents the results. For reconstruction, our method performs slightly worse than Naive-DB on the CLIP-I metric but surpasses all baseline methods on CLIP-I-Seg. We believe that CLIP-I-Seg more accurately reflects concept fidelity as it focuses on evaluating the segmented target concept. For generation, our method similarly surpasses all baselines on the CLIP-I-Seg metric. Furthermore, our approach achieves the best results in prompt fidelity and human preference while maintaining the highest efficiency.

5.4 Ablation Studies

Scaling Factor. We conduct an ablation study on the scaling factor η. The visualization in Fig. 8 suggests that $\eta = 1e-6$ yields favorable results. To verify this, we further conduct a quantitative evaluation, as shown in Tab. 3. The results confirm that our method achieves the best performance across all metrics with $\eta = 1e-6$. Therefore, we set η to $1e-6$ in our final model.

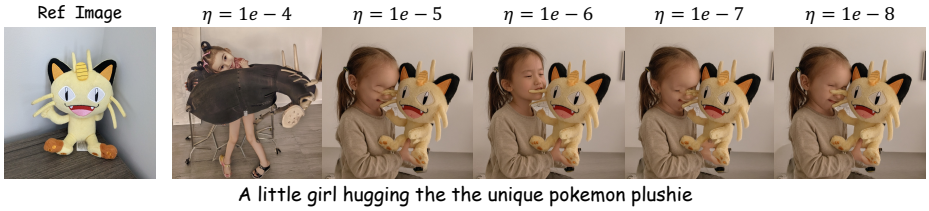

Fig. 8: Qualitative ablation results of scaling factor η. Our method achieves the best performance with $\eta = 1e-6$.

Fig. 9: Qualitative application results of MoKus. (a) MoKus can create a virtual concept within the model by describing the visual attributes of the concept. (b) MoKus can erase certain concepts by modifying the appearance description in the model. (c) MoKus can improve the model's performance on world knowledge benchmarks.

Number of Knowledge. We conduct an ablation study on the number of knowledge during the textual knowledge updating process. Tab. 2 presents the detailed quantitative results. Our method maintains robust and stable performance as the number of knowledge increases. Meanwhile, each additional knowledge item increases training time by only about 7 seconds, demonstrating the efficiency of our approach.

5.5 Extension to Knowledge-Aware Applications

Virtual Concept Creation. Our method enables the creation of virtual concepts within a generative model. The process involves describing a concept's visual attributes and incorporating this information into the model through textual knowledge updating. This procedure effectively establishes a new, usable virtual concept within the model. As illustrated in Fig. 9(a), we introduce a concept of an old white gentleman using the identifier `vfx`. The model then faithfully generates this virtual concept when prompted with the assigned name.

Concept Erasure. Concept erasure [15, 34, 58] aims to prevent the generation of unwanted concepts within a generative model. This task has gained increasing attention with the advancement of generative models. We can easily apply our method to concept erasure. As shown in Fig. 9(b), We update the answer to the question "What is the hair color of `Taylor Swift`?" to "black". We also applied similar modifications to other visual attributes. Then, we used "a photo of `Taylor Swift`" as the generation prompt. The results show that the model fails to generate accurate images of `Taylor Swift` after knowledge updating.

World Knowledge Benchmark. Our method can also improve model performance on world knowledge benchmarks. To evaluate this, we select a subset of WISE [40] that requires explicit world knowledge for generation. Fig. 9 (c) shows the visualization results. Through textual knowledge updating, we inject world knowledge into the model. This enables the model to use complex information for accurate generation. We also conduct quantitative experiments. As shown in Tab. 4, our approach effectively improves performance across all metrics.

Table 4: Quantitative results on the subset of WISE [56]. Our method can improve the model's performance on the world knowledge benchmark.

Metric	Pre-Update	Post-Update
Consistency	0.76	**1.26**
Realism	0.70	**1.42**
Aesthetic Quality	1.40	**1.68**
WiScore	0.81	**1.33**

Multiple Knowledge Combination. Our method can combine multiple pieces of knowledge during generation. As shown in Fig. 10, our method generates several target concepts faithfully based on the complex text prompt.

Fig. 10: Visualization of multiple knowledge combination. Our method can generate several concepts with high fidelity based on multiple knowledge inputs (highlighted in red).

6 Conclusion

In this paper, we introduce knowledge-aware concept customization, a new task that uses multiple knowledge to represent a target concept. We observe the phenomenon of cross-modal knowledge transfer, where changes in the text modality can transfer to visual modality and directly influence the generated visual output. Inspired by this observation, we propose MoKus. Our method first binds the visual appearance of the target concept to an anchor representation through visual concept learning. During textual knowledge updates, we project each knowledge to the obtained anchor representation through modifying certain parameters within the LLM encoder. To systematically evaluate this new task, we also introduce *KnowCusBench*, the first benchmark for knowledge-aware

concept customization. The impressive performance of MoKus demonstrates its effectiveness. Future work could explore three main directions: (1) extending knowledge-aware concept customization to the video domain, (2) developing more accurate and comprehensive evaluation metrics for knowledge-aware concept customization task, and (3) creating end-to-end knowledge-aware concept customization methods.

Acknowledgment. This work was supported by Hong Kong SAR Research Grants Council (RGC) General Research Fund (16219025), National Natural Science Foundation of China (NSFC) Young Scientists Fund Category B (62522216), Hong Kong SAR Research Grants Council (RGC) Early Career Scheme (26208924), and National Natural Science Foundation of China (NSFC) Young Scientists Fund Category C (62402408).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 9
2. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023) 4
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 10
4. Chen, C., Hu, S., Zhu, J., Wu, M., Chen, J., Li, Y., Huang, N., Fang, C., Wu, J., Chu, X., et al.: Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. arXiv preprint arXiv:2512.24146 (2025) 4
5. Chen, C., Zhu, J., Feng, X., et al.: S^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models. arXiv preprint arXiv:2508.12880 (2025) 4
6. Chen, H., Zhang, Y., Wu, S., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023) 2, 3
7. Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., Wei, F.: Knowledge neurons in pretrained transformers. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8493–8502 (2022) 4
8. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. arXiv preprint arXiv:2211.11337 (2022) 4
9. 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: Forty-first international conference on machine learning (2024) 6
10. Fang, C., Guo, H., Jiang, Z., He, C., Li, X., Xu, M.: Photon: Speedup volume understanding with efficient multimodal large language models. In: The Fourteenth International Conference on Learning Representations (2026) 4

11. Fang, C., He, C., Tang, L., Zhang, Y., Zhu, C., Shen, Y., Chen, C., Xu, G., Li, X.: Integrating extra modality helps segmentor find camouflaged objects well. arXiv preprint arXiv:2502.14471 (2025) 4
12. Fang, C., He, C., Xiao, F., Zhang, Y., Tang, L., Zhang, Y., Li, K., Li, X.: Real-world image dehazing with coherence-based pseudo labeling and cooperative unfolding network. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024) 4
13. Fang, J., Jiang, H., Wang, K., Ma, Y., Jie, S., Wang, X., He, X., Chua, T.S.: Alphaedit: Null-space constrained knowledge editing for language models. arXiv preprint arXiv:2410.02355 (2024) 5
14. 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. arXiv preprint arXiv:2208.01618 (2022) 3, 4
15. Gao, D., Lu, S., Zhou, W., Chu, J., Zhang, J., Jia, M., Zhang, B., Fan, Z., Zhang, W.: Eraseanything: Enabling concept erasure in rectified flow transformers. In: Forty-second International Conference on Machine Learning (2025) 14
16. Gu, J., Wang, Y., Zhao, N., Fu, T.J., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., et al.: Photoswap: Personalized subject swapping in images. Advances in Neural Information Processing Systems **36**, 35202–35217 (2023) 2, 4
17. Gu, J., Zhao, N., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., Wang, Y., Wang, X.E.: Swapanything: Enabling arbitrary object swapping in personalized image editing. In: European Conference on Computer Vision. pp. 402–418. Springer (2024) 4
18. Gu, X., Chen, G., Li, J., Gu, J.C., Hu, X., Zhang, K.: Ultraedit: Training-, subject-, and memory-free lifelong editing in large language models. arXiv preprint arXiv:2505.14679 (2025) 9
19. 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. Advances in Neural Information Processing Systems **36**, 15890–15902 (2023) 4
20. Gupta, A., Baskaran, S., Anumanchipalli, G.: Rebuilding rome: Resolving model collapse during sequential model editing. arXiv preprint arXiv:2403.07175 (2024) 4
21. Gupta, A., Rao, A., Anumanchipalli, G.: Model editing at scale leads to gradual and catastrophic forgetting. arXiv preprint arXiv:2401.07453 (2024) 4
22. Hartvigsen, T., Sankaranarayanan, S., Palangi, H., Kim, Y., Ghassemi, M.: Aging with grace: Lifelong model editing with discrete key-value adaptors. Advances in Neural Information Processing Systems **36**, 47934–47959 (2023) 4
23. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4775–4785 (2024) 4
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 7
25. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023) 10
26. Kong, Z., Zhang, Y., Yang, T., Wang, T., Zhang, K., Wu, B., Chen, G., Liu, W., Luo, W.: Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In: European Conference on Computer Vision. pp. 253–270. Springer (2024) 4

27. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023) 3, 4, 8
28. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023) 3
29. Li, X., Li, S., Song, S., Yang, J., Ma, J., Yu, J.: Pmet: Precise model editing in a transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18564–18572 (2024) 4
30. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8640–8650 (2024) 4
31. Liu, H., Huang, H., Wang, J., Liu, C., Li, X., Ji, X.: Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. arXiv preprint arXiv:2512.21514 (2025) 4
32. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) 6
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 9
34. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024) 14
35. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022) 4
36. Meng, K., Sharma, A.S., Andonian, A., Belinkov, Y., Bau, D.: Mass-editing memory in a transformer. arXiv preprint arXiv:2210.07229 (2022) 4
37. Mitchell, E., Lin, C., Bosselut, A., Finn, C., Manning, C.D.: Fast model editing at scale. arXiv preprint arXiv:2110.11309 (2021) 5
38. Mitchell, E., Lin, C., Bosselut, A., Manning, C.D., Finn, C.: Memory-based model editing at scale. In: International Conference on Machine Learning. pp. 15817–15831. PMLR (2022) 4
39. Niu, Y., Jin, W., Liao, J., Feng, C., Jin, P., Lin, B., Li, Z., Zhu, B., Yu, W., Yuan, L.: Does understanding inform generation in unified multimodal models? from analysis to path forward. arXiv preprint arXiv:2511.20561 (2025) 5, 9
40. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025) 4, 14
41. Peng, X., Zhu, J., Jiang, B., Tai, Y., Luo, D., Zhang, J., Lin, W., Jin, T., Wang, C., Ji, R.: Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27080–27090 (2024) 4
42. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022) 9
43. 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: International conference on machine learning. pp. 8748–8763. PmLR (2021) 10

44. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023) 2, 3, 4, 8, 9
45. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8543–8552 (2024) 3
46. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023) 4
47. Tan, C., Zhang, G., Fu, J.: Massive editing for large language models via meta learning. arXiv preprint arXiv:2311.04661 (2023) 5
48. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 9
49. Unsplash: Unsplash. https://unsplash.com/ (2024) 8
50. Wang, C., Chen, Y., Hu, Z., Chen, D., Chen, W., Wiegreffe, S., Zhou, T.: Quantifying the gap between understanding and generation within unified multimodal models. arXiv preprint arXiv:2602.02140 (2026) 5, 9
51. Wang, J., Lai, Z., Chen, J., Guo, J., Guo, H., Li, X., Yue, X., Guo, C.: Elastic diffusion transformer. arXiv preprint arXiv:2602.13993 (2026) 4
52. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the diffusion feature correspondence for consistent video editing. Advances in Neural Information Processing Systems **37**, 96541–96565 (2024) 4
53. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746 (2024) 4
54. Wang, J., Zhao, K., Guo, J., Wang, J., Guo, H., Zhu, C., Yue, X., Li, X.: Precisecache: Precise feature caching for efficient and high-fidelity video generation. In: The Fourteenth International Conference on Learning Representations 4
55. Wang, M., Zhang, N., Xu, Z., Xi, Z., Deng, S., Yao, Y., Zhang, Q., Yang, L., Wang, J., Chen, H.: Detoxifying large language models via knowledge editing. arXiv preprint arXiv:2403.14472 (2024) 4
56. Wang, P., Li, Z., Zhang, N., Xu, Z., Yao, Y., Jiang, Y., Xie, P., Huang, F., Chen, H.: Wise: Rethinking the knowledge memory for lifelong model editing of large language models. Advances in Neural Information Processing Systems **37**, 53764–53797 (2024) 4, 14
57. 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) 4
58. Wang, Y., Li, O., Mu, T., Hao, Y., Liu, K., Wang, X., He, X.: Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28759–28768. IEEE (2025) 14
59. Wang, Z., Jiang, Y., Zheng, D., Xiao, J., Chen, L.: Event-customized image generation. In: International Conference on Machine Learning (2025) 4
60. Wang, Z., Wang, X., Xie, L., Qi, Z., Shan, Y., Wang, W., Luo, P.: Styleadapter: A unified stylized image generation model. arXiv preprint arXiv:2309.01770 (2023)
61. 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:

Proceedings of the IEEE/CVF international conference on computer vision. pp. 15943–15953 (2023) 3
62. 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) 9
63. Yan, Y., Zhang, C., Wang, R., Zhou, Y., Zhang, G., Cheng, P., Yu, G., Fu, B.: Facestudio: Put your face everywhere in seconds. arXiv preprint arXiv:2312.02663 (2023) 4
64. 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) 3
65. Yu, L., Chen, Q., Zhou, J., He, L.: Melo: Enhancing model editing with neuron-indexed dynamic lora. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19449–19457 (2024) 4
66. Zhang, N., Tian, B., Cheng, S., Liang, X., Hu, Y., Xue, K., Gou, Y., Chen, X., Chen, H.: Instructedit: Instruction-based knowledge editing for large language models. arXiv preprint arXiv:2402.16123 (2024) 5
67. Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J., Chang, B.: Can we edit factual knowledge by in-context learning? arXiv preprint arXiv:2305.12740 (2023) 4
68. Zhu, C., Li, K., Ma, Y., He, C., Li, X.: Multibooth: Towards generating all your concepts in an image from text. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10923–10931 (2025) 4
69. Zhu, C., Li, K., Ma, Y., Tang, L., Fang, C., Chen, C., Chen, Q., Li, X.: Instantswap: Fast customized concept swapping across sharp shape differences. arXiv preprint arXiv:2412.01197 (2024) 2, 4