Training-Free Multi-Concept Image Editing

Niki Foteinopoulou[1], Ignas Budvytis[2], and Stephan Liwicki[1]

[1] Cambridge Research Laboratory, Toshiba Europe, Cambridge, UK
{niki.foteinopoulou, stephan.liwicki}@toshiba.eu
[2] Independent Researcher, UK
ignas.budvytis@gmail.com

Abstract. Training-free image editing with diffusion models is highly desirable yet is complex and remains a significant challenge. While recent optimisation-based methods achieve strong zero-shot edits from text, they still struggle to preserve identity and capture intricate details, such as facial structure, surface texture, or object-specific geometry, that exist below the level of linguistic abstraction. To address this fundamental gap, we propose Concept Distillation Sampling (CDS). To the best of our knowledge, we are the first to introduce a unified, training-free framework for target-less, multi-concept image editing.
CDS overcomes this linguistic bottleneck of previous methods by anchoring the editing process in the certainty of pretrained LoRA adapters. We integrate a highly stable distillation backbone (featuring ordered timesteps, regularisation, and negative-prompt guidance) with a novel dynamic weighting mechanism. This approach enables the composition and control of multiple visual concepts directly within the diffusion process, utilising spatially-aware priors from pretrained LoRA adapters without causing concept clashing. Our method preserves instance concept identity without requiring reference samples of the desired edit. Extensive quantitative and qualitative evaluations demonstrate that CDS establishes a new state-of-the-art over existing training-free editing and multi-LoRA composition methods on the InstructPix2Pix and ComposLoRA benchmarks. Project Page: https://nickyfot.github.io/cds/.

1 Introduction

Recent advances in conditional image generation have brought us remarkably close to human-level quality in both image creation and editing [17, 27, 28, 32]. This progress is largely driven by diffusion models trained on vast, diverse datasets [33, 39], which enable high-fidelity synthesis guided by flexible conditioning signals such as text, sketches, or images. Among these, text-to-image diffusion models have become particularly influential, thanks to their intuitive natural-language control, which often allows high-level control without requiring task-specific fine-tuning. Training-free approaches are particularly valuable in this context, as they bypass the steep computational costs and data-collection burdens of per-image or per-task fine-tuning. Furthermore, as foundational diffusion models continue to scale and improve, training-free methods inherently

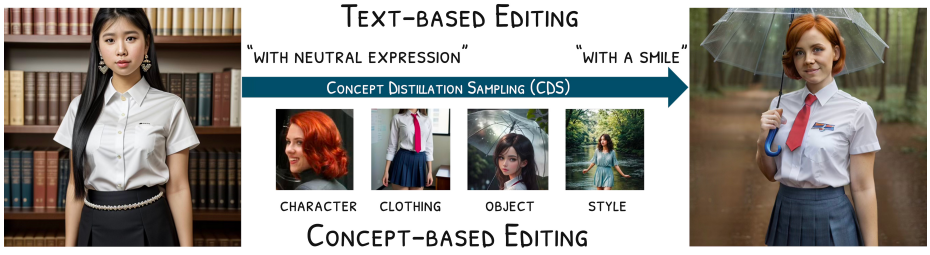


Fig. 1: Our method enables training-free, concept-based image editing with diffusion models. While recent state-of-the-art optimisation methods rely solely on text guidance and often lose instance identity during complex, multi-element changes, we unify a regularised optimisation objective with LoRA-driven concept composition. This enables multiple semantic or visual concepts (*e.g.* facial structure, clothing, objects, or backgrounds captured in Dreambooth-style LoRAs) to be combined and controlled directly within the diffusion process. This approach preserves concept identity and fine-grained details that would be impossible to adequately describe in text alone, bridging the gap between text-based and visual concept-driven control.

inherit their enhanced zero-shot capabilities, offering a highly adaptable and accessible solution for complex image manipulation.

However, editing existing images poses a subtler challenge than generating them from scratch. Effective editing requires not only semantic understanding but also content consistency, ensuring that the edited image remains recognisably the same subject. Recent training-free optimisation-based methods such as Delta Denoising Score (DDS) [8] achieve impressive zero-shot edits by refining the Score Distillation Sampling (SDS) objective [26], providing strong semantic control when an edit can be clearly expressed in natural language. Yet, language alone is an insufficient descriptor for all visual concepts. Many attributes defining identity, such as structure, texture, or object-specific geometry, exist below the level of linguistic abstraction (*e.g.*, the exact curvature of a specific character's jawline, or the unique wear patterns on a specific garment). As a result, state-of-the-art text-only methods often fail when edits involve multiple entities or concept-specific features that cannot be captured purely by text prompts.

In parallel, personalisation methods such as DreamBooth [30] and Low-Rank Adaptation (LoRA) [9] have demonstrated how instance-specific information can be encoded directly into a diffusion model's parameters. LoRAs effectively serve as concept representations, embedding appearance, style, or identity traits that text cannot describe. However, current multi-LoRA composition techniques [6, 23, 40] are primarily designed for text-to-image generation, not for image editing, where spatial alignment and subject consistency must be maintained.

To bridge this gap, we propose Concept Distillation Sampling (CDS), a novel method for preserving identity in multi-LoRA image editing. Our work unifies optimisation-based image editing and LoRA-based concept composition into a single framework for concept-aware, instance-consistent editing. DDS offers

precise control via optimisation, while LoRAs provide low-level, semantically grounded priors that anchor edits to specific concepts or instances. Combining the two enables edits that are both semantically meaningful and visually consistent across multiple concepts (Figure 1). Here, a concept denotes any coherent visual attribute encoded by a LoRA adapter that text alone cannot capture. Each LoRA thus serves as a latent prior and low-level anchor within the optimisation loop (Figure 2), enabling controlled, consistent multi-concept editing.

To the best of our knowledge, we are the first to introduce a unified, training-free framework that combines the tasks of multi-LoRA composition and image editing. Unlike text-only SOTA methods [14, 15] that suffer from gradient ambiguity during complex edits, CDS effectively overcomes the linguistic bottleneck of previous distillation methods by optimising and anchoring the editing process in the certainty of the LoRA adapters. Our proposed CDS framework comprises two synergistic components: First, we propose an optimised improvement upon previous distillation sampling works for 2D text-guided image editing. By revisiting DDS, we identify and resolve key design factors that limit stability and edit fidelity through principled refinements in timestep ordering, regularisation, and negative-prompt guidance. Second, we introduce a mechanism that dynamically weighs the contribution of each concept-LoRA. This mechanism performs patch-wise weighting of multiple LoRA outputs based on their feature similarity with the base model's output, allowing multiple concepts to be seamlessly composed and controlled directly within the diffusion process (Figures 1 and 2).

Furthermore, our method is truly generalisable. Any concept-LoRA can be incorporated into the editing process without requiring reference samples of the desired final edit. Recent works [22] explore the idea of concept-based image editing; however, requiring target images for novel concept composition is inherently counter-intuitive to the goal of creating entirely novel, out-of-distribution compositions (*i.e.*, unique, synthetic edits that combine elements in ways not present in any existing source image)s. CDS bypasses this limitation entirely, yielding stable, spatially-aware, and concept-preserving edits under a strict training-free constraint. Our contributions are summarised as follows:

1. We propose Concept Distillation Sampling (CDS), the first unified, training-free framework that formalises a new task: zero-shot, reference-free, multi-concept editing, and provides the first unified framework that addresses it.
2. The first component of CDS is a refined delta-denoising formulation that improves stability and fidelity in zero-shot editing, coupled with a novel dynamic weighting method that patch-wise balances the contribution of multiple concept-LoRAs without retraining.
3. We are the first to formalise the combined challenge of zero-shot multi-LoRA composition and training-free image editing. Furthermore, we show that CDS generalises across diverse editing scenarios. From text-driven zero-shot edits to complex pose transformations in InstructPix2Pix [1], and multi-element composition on ComposLoRA [40], CDS achieves state-of-the-art results, proving vastly superior to naive adaptations of existing text-to-image composition methods (*e.g.*, strategies like Switch or Merge applied to editing).

2 Related Work

Diffusion-based Editing: Recent diffusion-based editing methods adapt text-to-image models for image modification by exploiting semantic priors. Approaches like Prompt-to-Prompt [24], DiffEdit [3], and MasaCtrl [2] use attention manipulation, masking, or diffusion inversion for localised, text-guided edits, while Null-Text [24] enhances reconstruction fidelity. However, their reliance on text conditioning often compromises instance preservation across edits. Optimisation-based methods such as SDS [26] align parametric representations with diffusion priors but can cause unwanted global changes. DDS [8] mitigates this via differential gradients, and PDS [15] adds regularisation to retain image composition (Sec. 3). Recent work like DreamCatalyst [14] further improves identity preservation, though it remains inherently restricted to 3D scene editing. These methods randomly sample steps and remain confined strictly to text-to-image edits.

Composable Text-to-Image Generation seeks to synthesise images that coherently blend several concepts. Early work used structured conditioning (layouts or scene graphs) to improve compositionality [7, 13, 38]. Diffusion-based methods later refined the generative process or prompting to handle multi-concept inputs [5, 11, 16, 19, 25], but often depend on prompt design. Modular systems [4, 18, 20, 36] combine separately trained models yet require heavy fine-tuning. We instead propose CDS, a unified training-free framework that dynamically composes LoRA adapters directly into the image editing process.

LoRA-Based Skill Composition. LoRA provides lightweight personalisation for large diffusion models [30, 37] in image generation. Approaches such as LoRAHub, ZipLoRA, and related mixture or hypernetwork-based schemes [10, 31, 34, 35, 41] combine multiple LoRAs through learned weighting, but require supervision and offer limited generalisation. Arithmetic or attention-driven merging (LoRA Merge, CLoRA) [12, 23] and inference-time strategies (LoRA Switch, Composite, MultiLFG) [29, 40, 42] avoid retraining yet suffer from instability and inefficiency as LoRAs accumulate. LoRAtorio [6] proposes using the distance of LoRA-enhanced model predictions from that of the base model as a proxy of certainty. Concurrent works [22] target concept-based images with source-target example image pairs. In contrast, we enable multi-concept editing intuitively, which will synthesise novel views without explicit guidance. None of the existing multi-LoRA composition methods is formulated for training-free, target-less image editing. **To our knowledge, CDS is the first framework to formalise the multi-concept editing task, allowing seamless and controllable integration of multiple concepts without requiring reference images of the desired edit.**

3 Preliminaries

Score Distillation Sampling (SDS): Score Distillation Sampling (SDS) [26] leverages pretrained diffusion models to optimise a parametric image or 3D representation x_θ, where θ denotes the optimisable image parameters. This is achieved

by matching the model's predicted noise to that of a frozen teacher network $\epsilon_\phi(z_t, y, t)$ conditioned on a text prompt y. The SDS gradient can be written as:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\left[w(t) \left(\epsilon_\phi(z_t, y, t) - \epsilon\right) \frac{\partial z_0}{\partial \theta} \right], \tag{1}$$

where $w(t)$ is a time-dependent weighting term, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the injected noise, and z_t denotes the noisy latent at timestep t. Here, z_0 represents the clean latent encoding of the generated image, making $\frac{\partial z_0}{\partial \theta}$ the Jacobian of the differentiable encoding and rendering process. The timesteps t are uniformly randomly sampled, ensuring denoising across varying noise magnitudes rather than following the forward diffusion order. Although SDS provides an effective means for optimising image parameters under text guidance, it often introduces undesired global updates when used for editing (Figure 4).

Delta Denoising Score (DDS): To mitigate the spurious gradients of SDS in image editing, Delta Denoising Score (DDS) [8] removes the source-conditioned denoising component from the target's score estimate, leaving only the differential ("delta") signal corresponding to the semantic change between prompts. The DDS gradient is given by:

$$\nabla_\theta \mathcal{L}_{\text{DDS}} = \mathbb{E}_{t,\epsilon}\Big[w(t) \left(\epsilon_\phi(z_t^{\text{tgt}}, y_{\text{tgt}}, t) - \epsilon_\phi(z_t^{\text{src}}, y_{\text{src}}, t)\right)\Big] \frac{\partial z_0^{\text{tgt}}}{\partial \theta}, \tag{2}$$

where z_t^{src} and z_t^{tgt} denote the noised versions of the source and target latents, respectively, while y_{src} and y_{tgt} are their associated prompts. When the prompts coincide ($y_{\text{tgt}} = y_{\text{src}}$), the DDS gradient vanishes, thus preserving unedited regions. This formulation leads to more localised targeted edits with reduced blurring, although unintentional environment edits are not entirely eradicated.

To further constrain the distillation process, PDS [15] and DreamCatalyst [14] in 3D/NeRF editing explore aligning the posterior trajectories between source and target domains. In practice, this introduces a regularisation term scaling with the difference between the source and target latents ($\hat{\epsilon}_t^{\text{tgt}} - \hat{\epsilon}_t^{\text{src}}$ or $z_0^{\text{tgt}} - z_0^{\text{src}}$). These formulations rely on time-dependent weighting functions derived strictly from the base model's variance schedule and vanish when ordering timesteps [15].

4 Concept Distillation Sampling

In this section, we formalise the problem of target-less, multi-concept image editing. To solve this, we introduce Concept Distillation Sampling, a unified training-free framework. CDS comprises two core innovations: a regularised, timestep-ordered distillation objective that guarantees structural stability, and a dynamic concept-weighting mechanism that seamlessly composes multiple LoRA adapters without spatial interference. An overview is provided in Figure 2.

Problem Formulation: We address the task of zero-shot concept-based image editing. In this context, a "concept" refers to any semantically meaningful attribute that can be encoded via a LoRA adapter (*e.g.* a specific character

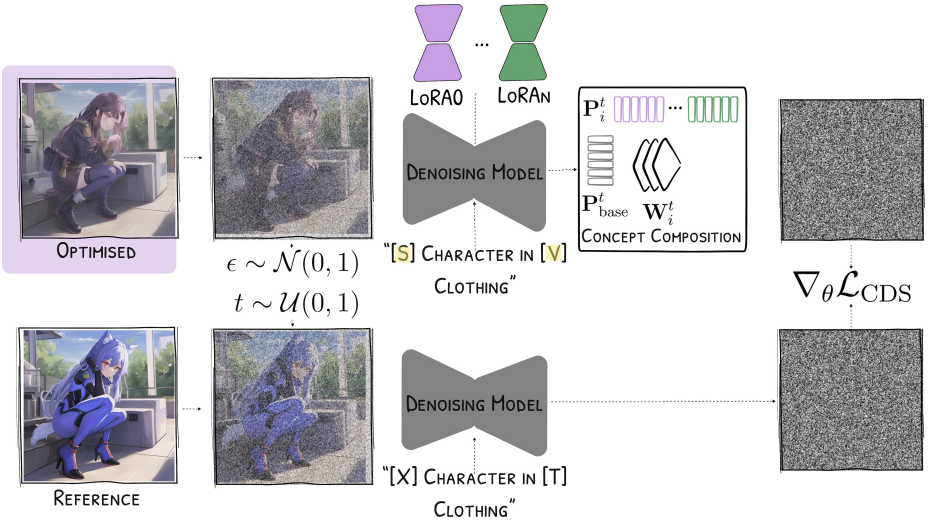


Fig. 2: Overview of our CDS image editing framework. We integrate multiple LoRA adapters into the image optimisation loop, enabling spatially-aware, concept-preserving edits across multiple concepts. The denoising trajectory follows ordered timesteps, allowing precise control over pose, style, and semantic attributes while maintaining overall composition.

identity, object geometry, or artistic style) that is encoded within a personalised LoRA adapter, but which cannot be adequately isolated or described by natural language alone. Each concept is assumed to be learned independently, to reflect truly in-the-wild zero-shot settings.

Given a source image x_{src}, a target text prompt y^{tgt}, and a set of N concept-encoding LoRA modules $\{lora_1, lora_2, \ldots, lora_N\}$, our objective is to synthesise a target image x_{tgt}. This image must preserve the unedited structural foundations of x_{src} while seamlessly integrating the non-linguistic concepts dictated by the LoRAs. We enforce a strict zero-shot constraint, that is, no reference samples of the desired final edit are available to guide the spatial composition. To achieve this, the editing process must satisfy two conditions simultaneously. First, the parametrised target x_0^{tgt} must be optimised such that the differential noise aligns with the desired semantic shift:

$$\epsilon_{\text{base}}(x_t^{\text{tgt}}, y^{\text{tgt}}, t) - \epsilon_{\text{base}}(x_t^{\text{src}}, y^{\text{src}}, t) \approx x_0^{\text{tgt}} - x_0^{\text{src}}. \tag{3}$$

Second, the standard base noise estimate ϵ_{base} must be dynamically replaced by a spatially aware, composite noise prediction derived from the N independent LoRA modules, preventing concept clash.

4.1 CDS Optimisation Objective

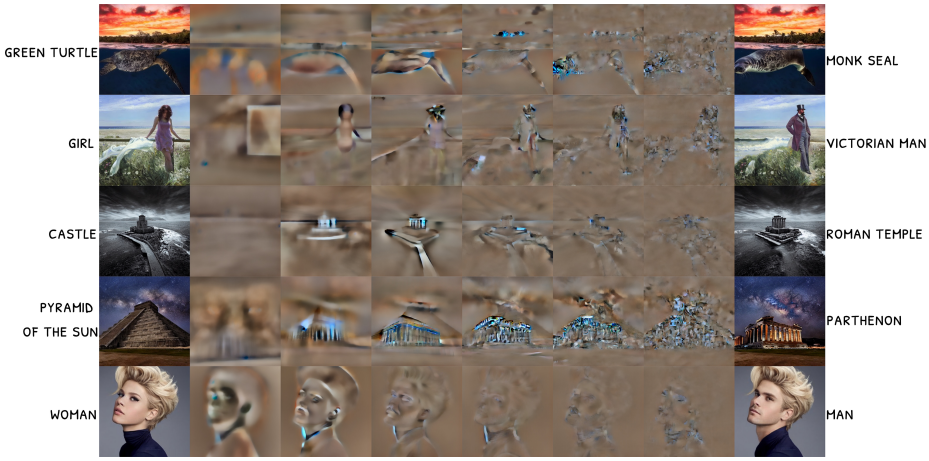


Fig. 3: Visualisation of timestep-ordered denoising. Comparison between the source image, $\nabla_\theta \mathcal{L}_{\text{CDS}}$ at intermediate optimisation steps, and the final target image. Unlike DDS [8], which samples timesteps uniformly at random, our method enforces a strict descending timestep order, enabling a coarse-to-fine denoising trajectory as is evident from the visualised gradients. Early steps capture high-frequency structural details such as edges, while later steps refine lower-frequency and stylistic components, resulting in a more coherent and stable edit.

Standard distillation objectives sample timesteps $t \sim \mathcal{U}(0,1)$ uniformly at random. While this promotes generalisation across noise levels, it fundamentally ignores the temporal structure of the diffusion reverse process, where early steps dictate high-frequency structural edges and later steps govern low-frequency stylistic rendering. To anchor the concept editing process, CDS enforces a strict descending timestep order, $1 > u > \cdots > v > 0$. This forces a coarse-to-fine denoising trajectory (Figure 3). However, deterministic ordering introduces destabilising gradients if unconstrained. To counter this without suffering the vanishing gradients of prior trajectory alignment works, we propose a training schedule-independent regularisation objective:

$$\nabla_\theta \mathcal{L}_{\text{CDS}} = \mathbb{E}_{t,\epsilon_t,\epsilon_{t-1}}\Big[\big(\eta\,|x_0^{\text{tgt}} - x_0^{\text{src}}| + (\hat{\epsilon}_t^{\text{tgt}} - \hat{\epsilon}_t^{\text{src}})\big)\frac{\partial x_0^{\text{tgt}}}{\partial \theta}\Big], \tag{4}$$

where $\hat{\epsilon}_t^{\text{src}}$ and $\hat{\epsilon}_t^{\text{tgt}}$ are the respective predicted noises, and η is a static hyper-parameter governing the strength of latent alignment. This explicit formulation prevents coefficients from decaying during sequential stepping, as opposed to [15] where ordered timesteps result to a zero derivative and thus no edit. To further guide the optimisation against degenerate visual modes commonly induced by aggressive LoRA conditioning, we integrate negative prompt guidance directly into the optimisation loop:

$$\hat{\epsilon}_{\text{guided}}(z_t, y^+, y^-, t) = (1+\lambda)\epsilon(z_t, y^+, t) - \lambda\epsilon(z_t, y^-, t), \tag{5}$$

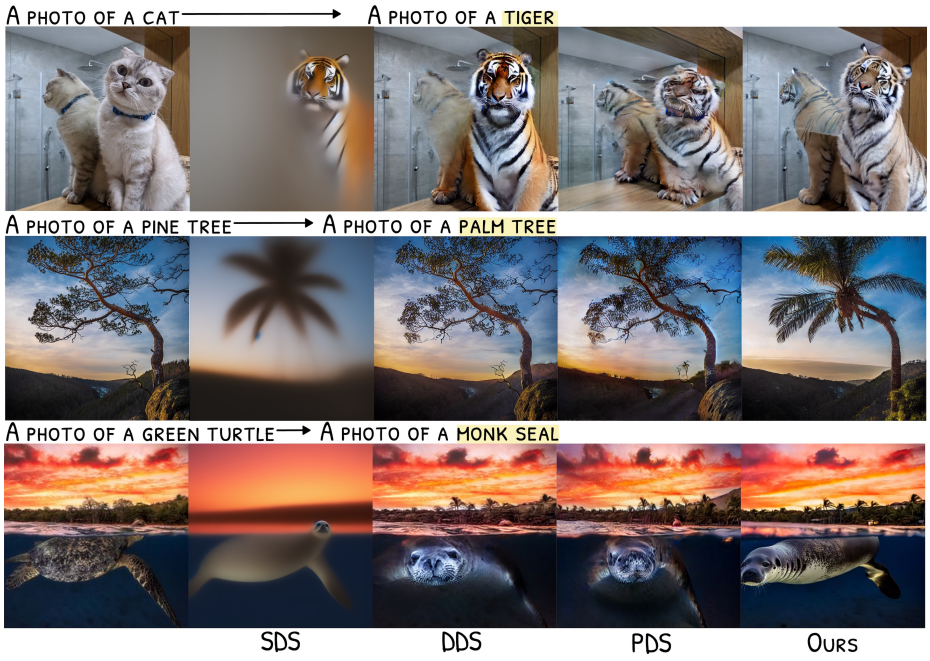


Fig. 4: Comparison of our optimisation objective vs baselines. Our method produces better results for the target subject with similarly perceived changes in the background.

where λ dictates the classifier-free guidance scale. Together, timestep ordering, explicit regularisation, and negative guidance form a highly stable distillation backbone capable of supporting complex conceptual shifts.

4.2 Dynamic Concept Weighting

While the $\mathcal{L}_{\text{CDS}}$ objective provides the vehicle for stable text-guided editing, the core of concept-based editing lies in composing multiple LoRA adapters without concept clashing. A naïve merge of LoRA weights often affects instance fidelity, especially as multiple LoRAs are composed [40], and creates severe spatial artefacts, even in the simpler image generation task. We propose a dynamic, inference-time weighting mechanism that assesses a proxy for the confidence of each concept at every denoising step. Our core intuition is that if a LoRA-enhanced model predicts noise highly similar to the unmodified base model, that LoRA is not contributing meaningful concept-specific information. On the contrary, where the prediction diverges strongly from the base model, the LoRA is actively injecting its encoded concept.

Let $\epsilon_{\text{base}}(z_t, t, c)$ be the noise prediction of the frozen base model, and the prediction of the i-th LoRA adapter be $\epsilon_{lora_i}(z_t, t, c)$. At each timestep t, we extract the spatial feature maps and partition them into P non-overlapping

patches. We define a flattening function ψ that maps these spatial features into a set of vectors:

$$\mathbf{P}^t_{\text{base}} = \psi(\epsilon_{\text{base}}(z_t, t, c)), \quad \mathbf{P}^t_i = \psi(\epsilon_{lora_i}(z_t, t, c)). \tag{6}$$

For every patch $p \in \{1, \ldots, P\}$, we compute the cosine similarity between the base model's patch and the corresponding patch from the i-th LoRA:

$$S^t_{i,p} = \langle \mathbf{P}^t_{\text{base},p}, \mathbf{P}^t_{i,p} \rangle_{\cos}. \tag{7}$$

High similarity indicates low concept injection. Therefore, we compute the adaptive spatial weight $\omega^t_{i,p}$ for each LoRA patch by applying a temperature-scaled SoftMin operation across all N concepts:

$$\omega^t_{i,p} = \text{SoftMin}_\tau(S^t_{i,p}) = \frac{\exp(-S^t_{i,p}/\tau)}{\sum_{j=1}^{N} \exp(-S^t_{j,p}/\tau)}, \tag{8}$$

where the temperature τ modulates the sharpness of the spatial isolation. These patch-wise weights are then upsampled back to the original spatial dimensions via nearest-neighbour interpolation, denoted as $\mathbf{W}^t_i$.The final, multi-concept conditional noise prediction used within the $\mathcal{L}_{\text{CDS}}$ optimisation loop is thus constructed dynamically:

$$\tilde{\epsilon}_{\text{concept}}(z_t, t, c) = \sum_{i=1}^{N} \mathbf{W}^t_i \odot \epsilon_{lora_i}(z_t, t, c), \tag{9}$$

where $\odot$ denotes the Hadamard (element-wise) product. By evaluating concept confidence iteratively and spatially, CDS ensures that distinct elements, such as a specific character's face from one LoRA and a unique clothing style from another, can seamlessly be composed in the edited image without concept confusion, while maintaining the source image structure.

5 Results

We summarise the core insights of our evaluation. First, we demonstrate that the foundational CDS distillation objective achieves state-of-the-art text-guided editing by finding a balance between semantic edit strength and structural preservation. Second, in the novel and highly challenging task of multi-concept editing (up to 5 LoRAs), our full CDS framework successfully prevents the severe concept erasure and spatial artefacts that plague existing composition baselines. Finally, comprehensive human and GPT-4V evaluations validate that CDS yields vastly superior visual harmony and instance fidelity, confirming that our spatially-aware dynamic weighting mechanism successfully bridges the gap between text-driven and concept-driven control.

Table 1: Quantitative Evaluation of our optimisation method against previous SoTA, on InstructPix2Pix. Our method achieves statistically significant improvements in CLIPScore while maintaining comparable LPIPS; the CLIPScore gain reflects a genuine semantic edge under a deliberate semantic–fidelity trade-off.

	CLIPScore↑	LPIPS↓
DiffusionClip‡	0.251 ± 0.022	0.572 ± 0.0594
PnP‡	0.221 ± 0.036	0.310 ± 0.075
InstructPix2Pix‡	0.219 ± 0.037	0.322 ± 0.215
DDS‡	0.225 ± 0.031	0.104 ± 0.061
PDS†	0.298 ± 0.044	**0.096 ± 0.033**
Ours	**0.308 ± 0.042***	0.100 ± 0.042

‡Reported in [8]
†PDS [15] is a 3D method adapted to RGB
* statistically significant at 99% confidence interval.

Table 2: Quantitative Evaluation of Concept Distillation Sampling (CDS). Performance comparison across four LoRA composition strategies for different numbers of LoRAs on *ComposLoRA* testbed.
Bold values indicate statistical significance at the 99% confidence interval.

	$N=2$		$N=3$		$N=4$		$N=5$		Total	
	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓
CDS	0.340 ± 0.030	**0.401 ± 0.174**	**0.361 ± 0.022**	**0.461 ± 0.158**	**0.370 ± 0.019**	0.534 ± 0.104	**0.368 ± 0.059**	0.503 ± 0.082	**0.359 ± 0.032**	**0.474 ± 0.129**
Composite	**0.351 ± 0.021**	0.542 ± 0.150	**0.369 ± 0.019**	0.613 ± 0.129	**0.377 ± 0.018**	0.644 ± 0.077	0.363 ± 0.057	0.664 ± 0.039	**0.365 ± 0.028**	0.615 ± 0.098
Switch	**0.350 ± 0.021**	0.456 ± 0.145	**0.361 ± 0.021**	0.485 ± 0.147	0.367 ± 0.020	**0.467 ± 0.128**	0.363 ± 0.057	**0.487 ± 0.108**	**0.360 ± 0.029**	**0.473 ± 0.132**
Merge	0.346 ± 0.016	0.547 ± 0.133	0.352 ± 0.018	0.601 ± 0.112	0.354 ± 0.021	0.634 ± 0.067	0.346 ± 0.061	0.662 ± 0.038	0.349 ± 0.029	0.610 ± 0.087

(a) Reality

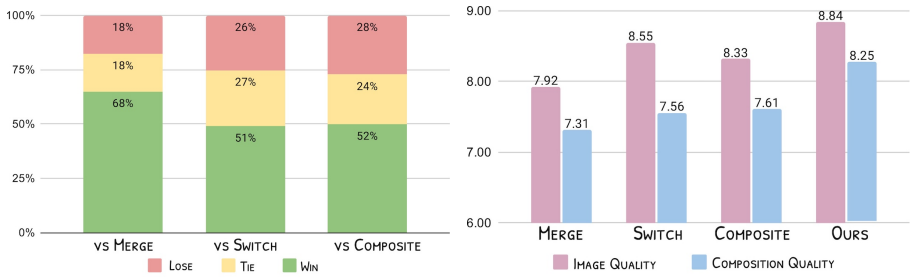
	$N=2$		$N=3$		$N=4$		$N=5$		Total	
	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓	CLIPScore↑	LPIPS↓
CDS	0.324 ± 0.026	**0.321 ± 0.096**	0.327 ± 0.027	**0.313 ± 0.084**	0.331 ± 0.027	**0.314 ± 0.079**	0.318 ± 0.029	**0.306 ± 0.067**	0.325 ± 0.027	**0.314 ± 0.082**
Composite	0.324 ± 0.029	0.658 ± 0.057	0.320 ± 0.031	0.67 ± 0.068	0.327 ± 0.026	0.668 ± 0.058	0.310 ± 0.022	0.663 ± 0.039	0.320 ± 0.027	0.665 ± 0.055
Switch	0.331 ± 0.028	0.378 ± 0.07	0.335 ± 0.03	0.368 ± 0.053	0.339 ± 0.025	**0.316 ± 0.047**	**0.331 ± 0.024**	0.319 ± 0.043	**0.334 ± 0.027**	0.345 ± 0.053
Merge	**0.341 ± 0.026**	0.388 ± 0.091	**0.341 ± 0.03**	0.428 ± 0.091	**0.340 ± 0.026**	0.473 ± 0.094	**0.331 ± 0.022**	0.498 ± 0.084	**0.338 ± 0.026**	0.447 ± 0.090

(b) Anime

5.1 Implementation Details

For a fair comparison with previous works, we select *stable-diffusion-v1.5* [28] as the backbone for all experiments. For experiments on InstructPix2Pix, we select 1000 images from the train set for all zero-shot text-guided comparisons of our optimisation objective against baseline DDS. We use a guidance scale $s = 10$, and image size 512×512 and checkpoint "runwayml/stable-diffusion-v1-5".

For our multi-concept editing experiments, we build upon the setup of [40], using *ComposLoRA* tests that include 22 pre-trained LoRAs spanning characters, clothing, styles, backgrounds, and objects, and modify from text-to-image generation to image-editing (details in Sec. F). We use the "Realistic_Vision_V5.1" and "Counterfeit-V2.5" checkpoints for realistic and anime-style images, respectively. For the realistic subset, a guidance scale $s = 7$, and image size 512×512; for the anime subset we use $s = 10$. DPM-Solver++ [21] is used as the sampler, with all LoRAs scaled by a weight of 0.8. We empirically set the temperature



(a) Win Rate (pairwise comparison). **(b)** Image and Composition Quality.

Fig. 5: Comparison of our CDS framework against previous LoRA-composition methods, evaluated using GPT-4V. Our method achieves higher perceived image and composition quality, with consistently higher pairwise win rates.

$\tau = 0.002$. For all experiments, we set the size of each patch to 2×2. For all experiments, we select 300 denoising steps and set $\eta = 0.5$. Since our method operates at inference time, all experiments are run on a single RTX A6000 GPU. Results are averaged over three runs. We provide an extended sensitivity analysis in Sec. A.

5.2 Quantitative Evaluation of Image Editing

Text-Guided Editing: We first evaluate the core distillation objective of CDS on the InstructPix2Pix benchmark in Tab. 1. By isolating the optimisation backbone (ordered timesteps, regularisation, and negative prompts), we demonstrate its stability over prior distillation methods. Our method achieves a statistically significant improvement in CLIPScore over previous SoTA at 99% confidence interval, while maintaining comparable LPIPS, visualised in Figure 4. **Multi-Concept Editing (Full CDS):** To evaluate the effectiveness of our method in multi-concept editing, we evaluate the full CDS framework, which integrates our regularised distillation objective with dynamic concept weighting. We compare our proposed CDS against three existing multi-LoRA strategies: *Composite* [40], *Switch* [40], and *Merge* [12], selected as they have publicly available code that can be adopted to image editing, without multiple inference steps. For each configuration, we generate scenes containing N distinct LoRAs, each representing a different concept from the *ComposLoRA* testbed (character, clothing, object, background or style). We then apply our method to simultaneously edit all N entities into different but semantically consistent counterparts (*e.g.*, $\text{character}_1 \rightarrow \text{character}_2$, $\text{clothing}_1 \rightarrow \text{clothing}_2$). This process is repeated across all valid concept combinations in the testbed to ensure a comprehensive evaluation of both cross-concept consistency and edit controllability.

CDS achieves the lowest LPIPS across most configurations Tab. 2; on the Reality total split, LPIPS is statistically tied with Switch, but CDS retains a clear advantage on Anime (0.314 vs. 0.345) and on every per-N slice for $N \geq 3$. CLIPScore remains on par or higher than baselines but saturates with more concepts,

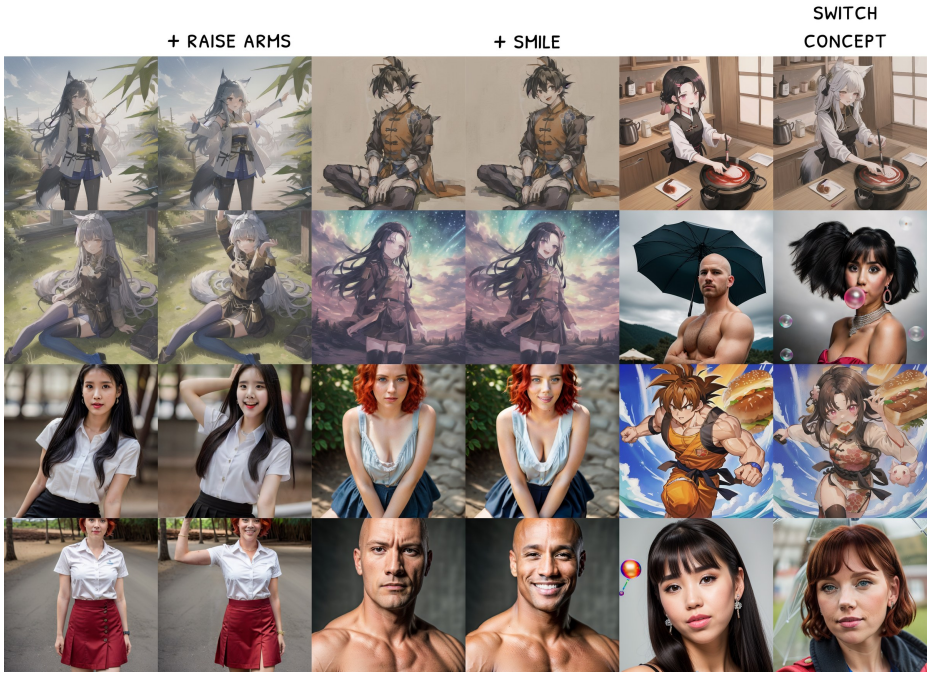


Fig. 6: Examples of our method under different editing conditions. Our approach enables consistent and coherent edits across diverse scenarios, including pose changes, facial expression modifications, and element or object swaps (*e.g.*, between different characters and environmental components). Each pair demonstrates the method's ability to preserve subject and visual fidelity while adapting to the intended transformation.

likely due to modality gaps and prompt ambiguity, suggesting that alone it's insufficient to capture perceptual and compositional quality in multi-concept edits. We see, however, that all methods have CLIPScores in similar ranges, which indicates that CLIPScore is adequate to measure whether the generic object is included, but not if the specific concept and instance are represented, thus highlighting the need for qualitative evaluation.

5.3 Qualitative Evaluation via GPT-4V and Human Ranking

Since CLIPScore alone cannot capture compositional quality, we complement it with qualitative evaluations using GPT-4V and human studies. GPT-4V scores and pairwise win rates against Switch, Composite, and Merge are shown in Figure 5b and Figure 5a. Human evaluators likewise ranked our method highest for image quality and concept integration, with the lowest average rank in Tab. 3, followed closely by Switch. Additional results can be seen in Sec. C.

Table 3: Human evaluation comparing our CDS framework to previous LoRA-composition methods. In line with GPT-4V assessments, human raters consistently preferred our method, which achieved the lowest average rank and most frequent first-place selections (highest win rate).

	Avg. Rank↓	Win Rate↑
CDS	**1.90**	**38%**
Compose	2.62	31%
Switch	2.49	25%
Merge	3.00	5%

Table 4: CDS Core Objective Ablation. We establish a baseline with standard DDS (300 steps unless indicated) and iteratively introduce our objective components to isolate their effects. Our unified formulation achieves the optimal balance between executing semantic changes and maintaining structural integrity.

	CLIPScore↑	LPIPS↓
DDS (200 steps)	0.225 ± 0.031	0.104 ± 0.061
DDS	0.305 ± 0.039	0.264 ± 0.061
Strict Timesteps (+DDS)	**0.332 ± 0.032**	0.461 ± 0.076
Regularisation (+DDS)	0.291 ± 0.041	**0.081 ± 0.033**
Negative Prompts (+DDS)	0.318 ± 0.040	0.319 ± 0.076
Full CDS	0.308 ± 0.042	0.100 ± 0.042

5.4 Pose Changes Under Concept Preservation

By unifying dynamic concept weighting with our distillation objective, CDS successfully handles complex edits involving simultaneous pose and semantic changes. Shown in Figure 6, CDS maintains concepts across pose and expression variations while preserving key composition in multi-LoRA edits.

5.5 Ablation

As shown in Tab. 4, enforcing ordered timesteps increases both metrics, reflecting stronger structural changes consistent with more pronounced edits. In contrast, adding regularisation lowers both metrics, indicating overly conservative optimisation that can suppress edits entirely. Strict Timesteps alone maximise CLIPScore (0.332) but at the cost of LPIPS (0.461); the L1 regulariser alone overconstrains the edit (CLIPScore 0.291). The two work in cooperation; full CDS recovers structural fidelity (LPIPS 0.100) without sacrificing semantic strength. Balancing these competing forces within the CDS objective produces the best trade-off between edit strength and perceptual fidelity, aligning with empirical visual inspection. Additional ablations on the effect of chosen hyperparameters can be found in Sec. A.

Fig. 7: Samples generated from the base model using text descriptions. We observe distortions in the face, additional limbs, missing concepts, and inconsistent knowledge of characters and objects, all of which affect downstream tasks.

6 Limitations and Error Analysis

Despite its training-free design, our method incurs computational cost. Runtime increases approximately linearly with the number of LoRA adapters when evaluated sequentially, as each adapter contributes a separate prediction. On an RTX A6000, a 512×512 edit with five LoRAs requires 44s sequentially, compared to 27s for a text-guided edit using only the base objective. However, the noise predictions for each LoRA are computationally independent operations and can therefore be evaluated in a single batched forward pass, where runtime drops to **33s**, comparable to a single-LoRA edit. Performance also depends on the quality and mutual alignment of LoRA adapters. Variability in publicly available LoRAs (*e.g.*, inconsistent training data or rank) can introduce imbalance, occasionally biasing the composite output toward dominant adapters, an inherent issue of all training-free methods. Finally, generation quality remains bounded by the base model: artefacts such as duplicated limbs or entangled concepts (*e.g.*, prompts such as "raise arm" that also cause a smile) stem from the base model's inherent priors, and are not explicitly addressed by our method (Figure 7).

7 Conclusion

We present Concept Distillation Sampling (CDS), a novel, training-free framework for zero-shot, concept-based image editing. By integrating a robust, regularised distillation objective with a dynamic spatial concept-weighting mechanism, our method enables fine-grained, multi-concept edits. CDS successfully overcomes the linguistic limitations of text-only editing, maintaining strict subject fidelity even under complex transformations, such as simultaneous pose and semantic changes. Extensive quantitative and qualitative evaluations demonstrate SoTA performance across both text-guided and multi-concept edits.

Bibliography

[1] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 18392–18402. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01764, https://doi.org/10.1109/CVPR52729.2023.01764

[2] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 22503–22513. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.02062, https://doi.org/10.1109/ICCV51070.2023.02062

[3] Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/pdf?id=3lge0p5o-M-

[4] Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/49856ed476ad01fcff881d57e161d73f-Abstract.html

[5] Feng, W., He, X., Fu, T., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/pdf?id=PUIqjT4rzq7

[6] Foteinopoulou, N., Budvytis, I., Liwicki, S.: Loratorio: An intrinsic approach to lora skill composition. ArXiv preprint **abs/2508.11624** (2025), https://arxiv.org/abs/2508.11624

[7] Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)

[8] Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 2328–2337. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00221, https://doi.org/10.1109/ICCV51070.2023.00221

[9] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In:

The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=nZeVKeeFYf9

[10] Huang, C., Liu, Q., Lin, B.Y., Pang, T., Du, C., Lin, M.: Lorahub: Efficient cross-task generalization via dynamic lora composition. In: Conference on Language Modeling (2024)

[11] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research, vol. 202, pp. 13753–13773. PMLR (2023), https://proceedings.mlr.press/v202/huang23b.html

[12] Hugging Face: Merging loras. https://huggingface.co/docs/diffusers/en/using-diffusers/merge_loras (2024), accessed: 2025-06-09

[13] Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1219–1228. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00133, http://openaccess.thecvf.com/content_cvpr_2018/html/Johnson_Image_Generation_From_CVPR_2018_paper.html

[14] Kim, J., Lee, S., Shin, J., Choi, J., Shim, H.: Dreamcatalyst: Fast and high-quality 3d editing via controlling editability and identity preservation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=FA5ZAJlv96

[15] Koo, J., Park, C., Sung, M.: Posterior distillation sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 13352–13361. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01268, https://doi.org/10.1109/CVPR52733.2024.01268

[16] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.: Multi-concept customization of text-to-image diffusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 1931–1941. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00192, https://doi.org/10.1109/CVPR52729.2023.00192

[17] Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742

[18] Li, S., Du, Y., Tenenbaum, J.B., Torralba, A., Mordatch, I.: Composing ensembles of pre-trained models via iterative consensus. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali,

Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/pdf?id=gmwDKo-4cY
[19] Lin, K., Yang, Z., Li, L., Wang, J., Wang, L.: Designbench: Exploring and benchmarking dall-e 3 for imagining visual design. ArXiv preprint **abs/2310.15144** (2023), https://arxiv.org/abs/2310.15144
[20] Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 23166–23178 (2021), https://proceedings.neurips.cc/paper/2021/hash/c3008b2c6f5370b744850a98a95b73ad-Abstract.html
[21] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. ArXiv preprint **abs/2211.01095** (2022), https://arxiv.org/abs/2211.01095
[22] Manor, H., Gal, R., Maron, H., Michaeli, T., Chechik, G.: Spanning the visual analogy space with a weight basis of loras (2026), https://arxiv.org/abs/2602.15727
[23] Meral, T.H.S., Simsar, E., Tombari, F., Yanardag, P.: Clora: A contrastive approach to compose multiple lora models (2024)
[24] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 6038–6047. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00585, https://doi.org/10.1109/CVPR52729.2023.00585
[25] Ouyang, Z., Li, Z., Hou, Q.: K-lora: Unlocking training-free fusion of any subject and style loras. ArXiv preprint **abs/2502.18461** (2025), https://arxiv.org/abs/2502.18461
[26] Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/pdf?id=FjNys5c7VyY
[27] Ramesh, A., Pavlov, P., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Hierarchical text-conditional image generation with clip latents (dall·e 2) (2022), https://arxiv.org/abs/2204.06125
[28] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
[29] Roy, A., Suin, M., Shah, K., Chellappa, R.: Multlfg: Training-free multi-lora composition using frequency-domain guidance. ArXiv preprint **abs/2505.20525** (2025), https://arxiv.org/abs/2505.20525
[30] 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: IEEE/CVF Conference on Computer Vision and Pattern

Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 22500–22510. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.02155, https://doi.org/10.1109/CVPR52729.2023.02155

[31] Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 6527–6536. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00624, https://doi.org/10.1109/CVPR52733.2024.00624

[32] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html

[33] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: an open large-scale dataset for training next generation image-text models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[34] Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. In: European Conference on Computer Vision. pp. 422–438. Springer (2024)

[35] Shenaj, D., Bohdal, O., Ozay, M., Zanuttigh, P., Michieli, U.: LoRA.rar: Learning to Merge LoRAs via Hypernetworks for Subject-Style Conditioned Image Generation (2024), https://arxiv.org/abs/2412.05148

[36] Simsar, E., Hofmann, T., Tombari, F., Yanardag, P.: Loraclr: Contrastive adaptation for customization of diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13189–13198 (2025)

[37] 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. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 66860–66889 (2023)

[38] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations.

In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=PxTIG12RRHS

[39] Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 893–911. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.51, https://aclanthology.org/2023.acl-long.51

[40] Zhong, M., Wang, S., Lu, Y., Jiao, Y., Ouyang, S., Yu, D., Han, J., Chen, W., et al.: Multi-lora composition for image generation. In: Transactions on Machine Learning Research (2024)

[41] Zhu, J., Chen, Y., Ding, M., Luo, P., Wang, L., Wang, J.: Mole: Enhancing human-centric text-to-image diffusion via mixture of low-rank experts. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/3415a8f8127d5b0ceb7fd321180b1954-Abstract-Conference.html

[42] Zou, X., Shen, M., Bouganis, C.S., Zhao, Y.: Cached multi-lora composition for multi-concept image generation. In: 13th International Conference on Learning Representations (2025)