Spanning the Visual Analogy Space with a Weight Basis of LoRAs

Hila Manor[2†,1], Rinon Gal[2†], Haggai Maron[2,1], Tomer Michaeli[1], and Gal Chechik[2,3]

[1]Technion, Israel [2]NVIDIA, Israel [3]Bar-Ilan University, Israel

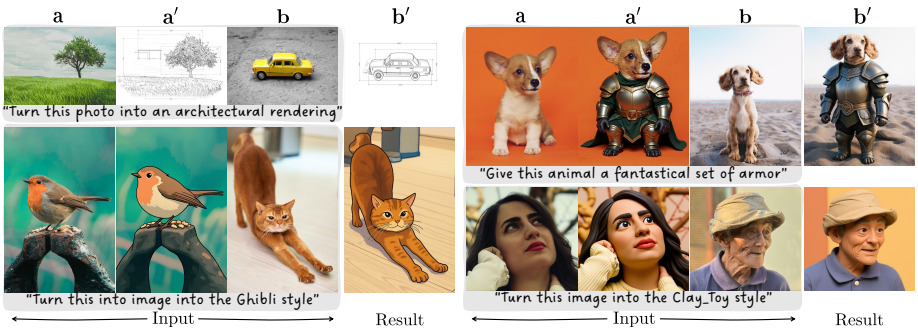

Fig. 1: LoRWeB. We present a novel method for analogy-based editing via learnable mixing of LoRAs. Given a prompt and an image triplet $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$ depicting a desired transformation, LoRWeB dynamically constructs a single LoRA from a learnable basis of LoRAs, and produces an editing result $\mathbf{b}'$ that applies the same analogy to $\mathbf{b}$.

Abstract. Visual analogy learning enables image editing via demonstration rather than textual description, allowing users to specify complex transformations difficult to articulate in words. Given a triplet $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$, the goal is to generate $\mathbf{b}'$ such that $\mathbf{a} : \mathbf{a}' :: \mathbf{b} : \mathbf{b}'$. Recent methods adapt text-to-image models with a single Low-Rank Adaptation (LoRA) module, but they face a fundamental limitation: attempting to capture the diverse space of visual transformations within a fixed module constrains generalization. Inspired by recent work showing that LoRAs in constrained domains span meaningful, interpolatable semantic spaces, we propose LoRWeB, which specializes the model for each analogy task in a single inference pass. LoRWeB dynamically composes learned transformation primitives, informally, choosing a point in a "*space of LoRAs*". We introduce two key components: (1) a learnable basis of LoRAs to span the space of different visual transformations, and (2) a lightweight encoder that dynamically weighs these basis LoRAs given the input analogy pair. Comprehensive evaluations demonstrate state-of-the-art performance and significantly improved generalization to unseen transformations. Our findings suggest LoRA basis decompositions are a promising direction for flexible visual manipulation tasks. See our website for code.

Keywords: Image analogies · Image editing · LoRA · Flow models

† This work was done while HM and RG were working at NVIDIA.

1 Introduction

Text-based image editing models [4, 5, 51, 63, 71] have recently emerged as powerful tools for controllable image generation and manipulation, enabling users to modify images through textual descriptions. However, many visual transformations are inherently difficult to articulate precisely through text alone. For example, consider describing the transformation that converts a photograph into the style of a specific painting, or conveying an exact target pose through text. Such inherent limitations motivate the need for alternative paradigms that can capture and apply complex visual transformations.

Visual analogy learning [25] offers a compelling solution to this challenge by enabling models to understand transformations through examples rather than explicit descriptions. In this paradigm, given a triplet of images $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$, the goal is to generate an image $\mathbf{b}'$ such that the visual relationship $\mathbf{a} : \mathbf{a}' :: \mathbf{b} : \mathbf{b}'$ holds. That is, the transformation applied between $\mathbf{a}$ and $\mathbf{a}'$ should be analogously applied to $\mathbf{b}$ to produce $\mathbf{b}'$. This approach allows users to specify complex visual changes through demonstration, making it possible to capture nuanced transformations that would be difficult or impossible to describe textually.

Early learning-based approaches trained stand-alone analogy models directly from analogy data [3, 35, 46, 60, 61, 64], but this lead to limited task diversity and image quality, or required extensive compute. More recent work aims to leverage the rich prior of powerful text-to-image backbones by adapting them to the visual analogy task, using a single Low-Rank Adaptation (LoRA) module [17, 36, 52]. While effective, these methods face a fundamental limitation: they attempt to capture the diverse space of possible transformations within a single adaptation module. This constraint may limit the model's ability to generalize across the rich variety of relationships that exist in images.

We hypothesize that specializing the model to each specific analogy task at inference time may improve performance and generalization. While this objective could theoretically be achieved via hypernetworks that generate task-specific LoRAs [52], these are notoriously difficult to train and often suffer from instability [42]. Instead, we draw inspiration from recent work [11] demonstrating that fine-tuned LoRAs (*e.g.*, for personalization tasks) can form an interpretable weight space. In this space, individually-trained LoRAs act as a basis of fundamental and semantic visual traits, and interpolating between the weights of these LoRAs can effectively cover new points in this semantic space, which allows for creating new, blended concepts. Building on this insight, we explore a similar principle for visual analogy learning and propose LoRWeB, a two-component system: (1) a learnable basis of LoRA modules and (2) a lightweight encoder that dynamically combines LoRAs from this basis at inference time based on the input analogy pair. These components are jointly trained, enabling the model to compose appropriate transformations for novel analogies unseen during training.

Existing methods typically encode analogy images using vision-language models such as CLIP [44] or SigLIP [67] and provide these encodings as context to the generative model. This can provide the higher-level semantic understanding needed for understanding the analogy task. However, this might lead to loss of

detail in fine-grained visual detail preservation. Recent advances have shown that diffusion models can extract remarkably accurate visual details through extended attention mechanisms [4, 7]. Thus, we leverage this capability by providing the full analogy triplet directly to the diffusion model through an extended-attention mechanism, while reserving CLIP-based encodings specifically for LoRA selection. This approach allows LoRWeB to balance fine-detail consistency with the higher-level semantics required to understand the analogy task.

We evaluate LoRWeB against established baselines and show it achieves state-of-the-art results. Our contributions include: (1) a novel architecture that decomposes visual analogy learning into a basis of LoRAs with dynamic composition, and (2) a comprehensive evaluation showing improved generalization to unseen transformations compared to existing single-LoRA approaches.

2 Related Work

Visual Analogies. Visual analogies, also known as "Image Analogies" [25], "Visual Prompting" [3] or "Visual Relations" [17], is the task of learning a transformation from a pair of before-and-after exemplars and applying it analogously to new images. Early non-neural methods learned explicit per-pair filters for simpler tasks such as style transfer [25], or per-pair optimization for relighting in 3D [14]. With the advent of network-based methods, initial works proposed models conditioned on image embeddings or NeRF [40] representations to present analogies through simple vector arithmetic [13, 21, 32, 46]. While these methods showed promise on datasets of simple, isolated objects, they struggled with the complexity of real-world images, and still mostly tackled style-transfer analogies. Newer methods instead treat analogy learning as in-context learning, where the model is directly conditioned on the exemplar pair and a reference image, and is trained to successfully synthesize the matching target [3, 60, 61, 64]. More recently, some works adapt pre-trained text-to-image foundation models to the new task, going beyond simple style-transfer analogies. For example, [53] use per-sample optimization, backpropegating through the entire diffusion process. However, such approaches can require dozens of minutes to edit every image, and their memory requirements can be challenging with newer, larger models. Another approach adapts the foundation model directly using a LoRA module [8, 17, 26]. These methods, while showing impressive results, still struggle to generalize to unseen tasks. Our approach aims to tackle this limitation by avoiding the bottleneck of a single LoRA, opting instead to train a basis of adapters which can be mixed to achieve greater flexibility and better generalization.

Diffusion-Based Image Editing. The unprecedented semantic control offered by large scale text-to-image diffusion models [4, 45, 47] has inspired extensive work leveraging them as priors for image editing. Early works add noise to an image and remove it conditioned on a novel prompt [39], though such methods significantly change image structure. Subsequent work improved content preservation by manipulating internal feature representations [23, 43, 59]

or the model's denoising trajectory [10, 22, 29, 30]. Recent works go beyond text and incorporate different control modalities for enhanced precision, such as ControlNet [7, 69], or attention-sharing [1, 16, 24, 57]. Others explore text-free editing to enable modifications that cannot be textually described [20, 37], though without direct control. Transformer-based diffusion models further popularized attention-sharing for maintaining subject consistency in personalization [15, 48] and editing [6, 54, 56]. Among these, Flux.1-Kontext [4] was specifically trained for text-based editing, incorporating input images via extended attention mechanisms. Our work extends this model's capabilities to visual analogies.

LoRA and Weight Bases. LoRA [26] is a parameter-efficient fine-tuning method that modifies a model using low-rank matrices learned on top of the existing weights. Its success lead to a range of downstream approaches trying to improve on the original formula. Of these, a line of work explores the combination of multiple LoRa modules, either to combine them post-tuning [50, 68], or as a means of turning an existing model into a mixture of experts [12, 38, 62]. In visual content generation, a recent work [11] showed that independently trained LoRA weights can span a semantic basis, and interpolations between them can be meaningful. However, to use their findings on a practical task such as face personalization, their approach required 65,000 independently trained LoRAs, and further employed PCA to span this basis, using test-time tuning per subject. Similar observations on weight bases of LoRAs were made in language processing: LoraHub [27] independently trained LoRAs and combined them post-tuning with test-time optimization to enable new tasks at inference given multiple reference outputs of the new tasks, and Sci-LoRA [9] combined LoRAs for tasks like text simplification across different scientific domains. We propose to further expand on this idea by learning a joint basis of LoRAs, along with the router to mix and match between them efficiently at inference time. Thus, we can learn a basis that is more amenable to interpolations, and enable better downstream generalization. Our approach does not necessitate the extensive inference-time compute required by prior approaches, rather creating an input dependent LoRA with a single forward pass at inference time, without test-time optimization.

3 Method

3.1 Preliminaries

Low-Rank Adaption. LoRA [26] offers a parameter-efficient alternative to conventional fine-tuning of large models by learning low-rank matrices that adapt the pre-trained weights. Specifically, starting from a frozen pre-trained weight matrix $\boldsymbol{W}_0 \in \mathbb{R}^{m \times n}$, the update of the weights is represented as the product of two learned low-rank matrices $\boldsymbol{\Delta W} = \boldsymbol{BA}$, where $\boldsymbol{B} \in \mathbb{R}^{m \times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r \times n}$, and the rank r is typically $r \ll \min(m, n)$. This formulation drastically reduces the number of trainable parameters, while typically maintaining model performance. The final weights of the model are then updated to $\boldsymbol{W} = \boldsymbol{W}_0 + \frac{\alpha}{r}\boldsymbol{BA}$, where α is a scaling constant.

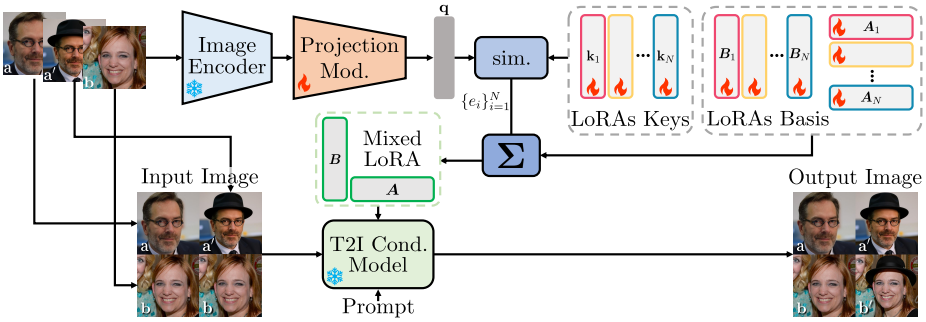

Fig. 2: LoRWeB Overview. We first encode $\mathbf{a}$ and $\mathbf{a}'$, that describe a visual transformation (*e.g.* adding a hat to the man), and $\mathbf{b}$, which should be edited analogously (*e.g.* adding a hat to the woman) with CLIP [44], and a small learned projection module. The similarity between the encoded vector and a set of learned keys determines the linear coefficients for combining the learned LoRAs into a single, mixed LoRA. This mixed LoRA is injected into a conditional flow model (*e.g.* Flux.1-Kontext [4]). Next, we build a 2×2 composite image from $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$. The conditional flow model gets this composite image as its input, along with a guiding edit prompt, and produces a composite image with the edited results $\mathbf{b}'$ in the bottom-right quadrant.

Flow Models. Flow-based generative models [2, 33, 34] learn a series of transformations to map samples from one probability distribution $\mathbf{x}_1 \sim p$, to samples from another $\mathbf{x}_0 \sim q$. In the generative context, p is typically taken as the standard normal distribution, while q is the data distribution in a latent space [47]. Then, These models learn a time-dependent velocity field $v_\theta(\mathbf{z}_t, t)$ that models the direction from a noisy sample towards the data manifold. The noisy sample $\mathbf{z}_t$ is a linearly interpolated latent between the two data distributions, and is given as $\mathbf{z}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. The rectified flow-matching training loss for a conditional model conditioned on a text prompt c is given by

$$\mathcal{L} = \mathbb{E}_{t\sim p(t),\mathbf{x}_0,\mathbf{x}_1,\mathbf{y},c}\left[\|v_\theta(\mathbf{z}_t, t, \mathbf{y}, c) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2\right]. \tag{1}$$

Here, the velocity field is optionally conditioned on a context image $\mathbf{y}$.

3.2 LoRWeB

Our objective is to perform visual analogy completion [25], where the model infers a proposed edit from a given image pair and applies it to a new image. Formally, two reference images, $\mathbf{a}, \mathbf{a}' \in \mathbb{R}^D$, are related by an unknown transformation $\mathcal{T} : \mathbb{R}^D \to \mathbb{R}^D$ such that $\mathbf{a}' = \mathcal{T}(\mathbf{a})$. Given a new image $\mathbf{b} \in \mathbb{R}^D$, the goal is to generate $\mathbf{b}' \in \mathbb{R}^D$ such that $\mathbf{b}' \approx \mathcal{T}(\mathbf{b})$.

Naive Solutions and Limitations. Using a pre-trained conditional generative model, such as FLUX.1-Kontext [4], existing solutions for this task fine-tune the

model using a single LoRA [49]. For example, given the input triplet $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$, one can construct a composite 2×2 image $\mathbf{y} = [\mathbf{a}, \mathbf{a}'; \mathbf{b}, \mathbf{b}]$, as shown in the bottom-left part of Fig. 2, which serves as the conditioning input. The goal of the model is to output $\mathbf{x}_0 = [\mathbf{a}, \mathbf{a}'; \mathbf{b}, \mathbf{b}']$, such that the bottom-right quadrant was transformed from $\mathbf{b}$ to $\mathbf{b}'$, by training over Eq. (1). While these approaches perform well when the transformation $\mathcal{T}$ is constrained to the analogy types seen in the training set, they struggle to generalize to new, diverse transformations. We propose this arises in part because the single adapter struggles to capture the wide range of analogical relationships, from different style transfers to objects insertion or layout modifications.

A more advanced solution could be to span the diverse set of possible analogies using multiple adapters. Recently, [11] demonstrated that LoRAs trained for model personalization can span a semantic basis. Inspired by this, we propose to learn such a basis for *task LoRAs*. A naïve adaptation of [11] to analogy tasks would require us to first optimize a single adapter for each of N analogy types seen during training, such that each LoRA module i excels at a different subset of visual edits. Once the specialized adapters are trained, they can be linearly combined to obtain an equivalent single "novel" adapter

$$\boldsymbol{A} = \sum e_i \boldsymbol{A}_i, \quad \boldsymbol{B} = \sum e_i \boldsymbol{B}_i, \tag{2}$$

where the coefficients e_i are optimized for each analogy task separately through the use of Eq. (1) and the reference pair of images $\{\mathbf{a}, \mathbf{a}'\}$. The model using the combined LoRA is then used to transform $\mathbf{b}$ to $\mathbf{b}'$.

However, this approach requires training a large number of models, and a test-time tuning phase for every new analogy. Indeed, [11] required $65,000$ LoRAs to capture the constrained space of faces, and collecting a significant number of different analogy pairs is more difficult.

Our Approach. Instead, we propose LoRWeB (**Lo**w-**R**ank **We**ight **B**asis). Rather than training individual LoRAs and combining them only at inference time, we propose to simultaneously train a basis of LoRA adapters, jointly with an encoder that predicts linear-combination coefficients for each input analogy pair. Specifically, we maintain a set of N rank-r LoRAs, and associate each $\boldsymbol{A}_i, \boldsymbol{B}_i$ pair where $i \in \{1, \ldots, N\}$ with a learnable key vector $\mathbf{k}_i \in \mathbb{R}^d$, as depicted in the right part of Fig. 2. Next, we define an encoder network based on a frozen, pre-trained ViT [66], $\mathcal{E}$, *e.g.* CLIP [44]. The encoder takes as input the conditioning image triplet, $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$, passes them through the ViT, concatenates the results and projects them through a small learnable projection module $\mathcal{P}$ that outputs the results as a query vector $\mathbf{q} \in \mathbb{R}^d$:

$$\mathbf{q}(\mathbf{a}, \mathbf{a}', \mathbf{b}) = \mathcal{P}\Big(\big[\mathcal{E}(\mathbf{a}), \mathcal{E}(\mathbf{a}'), \mathcal{E}(\mathbf{b})\big]\Big). \tag{3}$$

Then, based on the conditioning query, we compute N coefficients with

$$e_i(\mathbf{a}, \mathbf{a}', \mathbf{b}) = \left[\text{softmax}\left(\frac{\mathbf{q}(\mathbf{a}, \mathbf{a}', \mathbf{b})\boldsymbol{K}^T}{\sqrt{d}}\right)\right]_i, \tag{4}$$

where $K \in \mathbb{R}^{d \times N}$ contains the key vectors $\{\mathbf{k}_i\}_{i=1}^{N}$ in its columns. The final LoRA combination follows

$$\Delta \boldsymbol{W} = \boldsymbol{B}\boldsymbol{A} = \sum e_i (\boldsymbol{B}_i \boldsymbol{A}_i), \tag{5}$$

and is marked as "Mixed LoRA" in Fig. 2.

Importantly, a single rank-r' LoRA is a special case of LoRWeB. Specifically, if the learned mixing router collapses to a constant output for any input $\{\mathbf{a}, \mathbf{a}'\mathbf{b}\}$, then e_i will be constant for all inputs. This will result in the same static mixed N rank r LoRA combination. This mixture can yield a matrix of rank $r' = Nr$, which is equivalent to finetuning a single rank r' LoRA. However, as LoRWeB differently combines LoRAs for different inputs, this dramatically increases the expressive power of our representation. Indeed, a single rank r' LoRA is just a single point in the space of all rank r' LoRAs, which LoRWeB aims to span.

We use the same pre-trained encoder across different network layers, but train individual LoRWeB modules, including LoRAs, keys and projections for each targeted weight matrix $\boldsymbol{W}_0$ in the network. This enables capturing different semantic elements for each weight and layer in the model.

4 Experiments

Settings. We evaluate our approach using Flux.1-Kontext [4] as the pre-trained conditional flow model and CLIP [44] as the image encoder backbone. For our LoRAs Basis, we match the capacity of prior work [17], using $N = 32$ adapters, each of rank $r = 4$, with $d = 128$ as the learned key dimension. We project the CLIP-encoder's output to $\mathbb{R}^d$ using a single fully-connected layer. To save on compute, during training we set the resolution to a maximum of 512×512 images, resizing on the long-edge of images. Additional implementation details are in App. A.1. Importantly, the use of a very lightweight CLIP encoder, combined with efficient Einsum matrix multiplications, ensure the inference overhead of LoRWeB is minimal. Indeed, LoRWeB inference takes 33.4 seconds, compared to 32.4 seconds for a single $r = 128$ LoRA. See further details on our inference efficiency in App. A.2. We compare LoRWeB to four recent baselines: A standard Flux.1-Kontext LoRA of similar parameter capacity (equivalent to LoRWeB with $N = 1, r = 128$), as well as three prior visual analogy methods based on Flux.1-Dev (RelationAdapter [17], VisualCloze [31] and EditTransfer [8]). We additionally compare to Diffusion Image Analogies (DIA) [53] and PairEdit [36], in App. B.1. DIA relies on inversion into CLIP space and expensive per-sample backpropagation through the entire diffusion process of Stable Diffusion 1.4 [47]. Nevertheless, for a fuller assessment, we compare their method with ours, including an additional variant we design to adapt DIA to Flux.1-Kontext. PairEdit also uses per-sample optimization, training three Flux LoRAs for each input triplet: A concept LoRA and a content LoRAs trained in parallel to describe the transformation from $\mathbf{a}$ to $\mathbf{a}'$, and an inversion LoRA to reconstruct $\mathbf{b}$. Once all LoRAs are trained they can be used to generate $\mathbf{b}'$.

Fig. 3: LoRWeB visual analogy results. Using a LoRA Basis allows LoRWeB to generalize to a wide variety of new analogy tasks, from adding objects to transferring specific styles or makeup or copying pose changes. Please zoom in for more details.

Dataset. We train our model using the public Relation252k [17] set, which contains 16K analogy image pairs across 208 tasks. Since the train-set split of Relation252k is not fully publicly available, and only 10 unseen analogy tasks were released, we extend it with a custom validation set to evaluate visual analogies. Specifically, we focus on analogies that were not found in the training set, which we create in the following manner: First, we collect over 100 Unsplash[1] photos covering diverse concepts from three categories: animals, persons, and general objects. Next, we create analogy pairs with a focus on two categories: transformations which are in-domain for the base text-to-image model, and transformations that are not. For in-domain transformations, we first use an LLM to summarize the training prompts for each task in the training-set of Relation252k, yielding 208 representative prompts. Next, we ask the LLM to generate novel prompts

[1] `https://unsplash.com/`

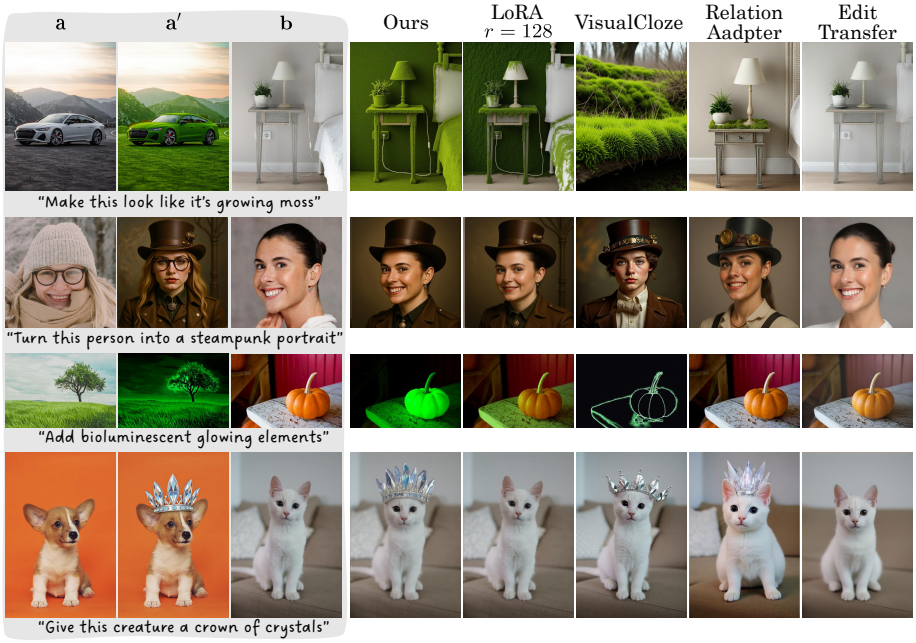

Fig. 4: Comparisons with baseline methods on unseen tasks. We compare LoRWeB with four recent baselines: RelationAdapter [17], VisualCloze [31] and Edit-Transfer [8], as well as a standard Flux.1-Kontext LoRA of similar parameter capacity. Our approach generalizes across more diverse tasks, and better maintains the visual details of both the subject and the analogy.

that differ from the training set's prompts and manually verify that they match the given concept categories. We filter prompts where Flux.1-Kontext fails to produce a meaningful edit, and further randomly select 15 prompts per concept category from the remainder. We generate three images per prompt, obtaining a total of 135 analogy pairs. For out-of-domain analogies, we collect 18 community LoRAs for Flux.1-Kontext from HuggingFace, which were trained to enable edits the base model failed with. We use these pre-trained LoRAs, and repeat the previous random sampling strategy to get 135 analogy pairs. Finally, we randomly select as the input images $\mathbf{b}$ two images from the matching concept category, with a similar aspect ratio to $\mathbf{a}$ and $\mathbf{a}'$, and crop them to the exact size. Our resulting set contains 540 analogy triplets across 90 tasks and 3 concept categories. Including the unseen set of Relation252K, this gives 100 tasks across 840 analogy triplets. On all experiments, we first aggregate the results per analogy task, and then aggregate over all tasks. More details appear in App. A.3.

4.1 Qualitative Evaluations

Figures 1 and 3 include results of analogy-based editing using LoRWeB. Notably, the model generalizes to new tasks covering style transfer, background

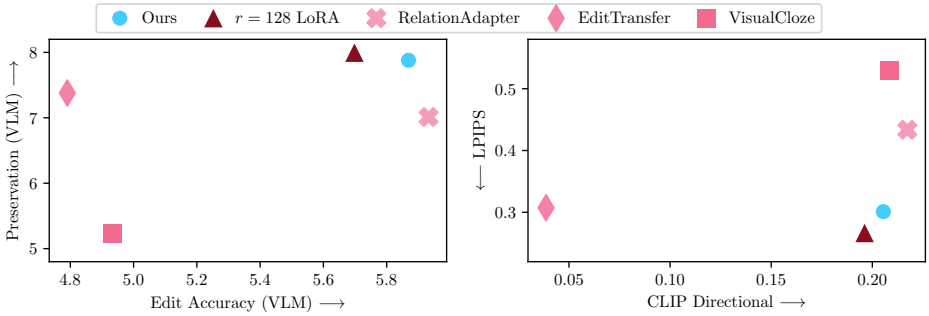

Fig. 5: Quantitative comparisons. (left) Accuracy of the applied edit and preservation of $\mathbf{b}$ in $\mathbf{b}'$ using Gemma-3 [55]. Top right is better. (right) CLIP directional similarity and LPIPS between $\mathbf{b}'$ and $\mathbf{b}$. Bottom-right is better. Our method pushes the Pareto front of edit accuracy-preservation, achieving higher edit accuracy while strongly preserving the input image.

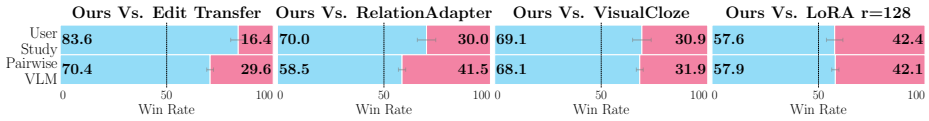

Fig. 6: Pairwise image comparisons. We compare LoRWeB to four baselines on overall edit quality preference via both a user study and using a VLM. LoRWeB produces edits that are favored by both. Error bars are the 68% Wilson score interval.

replacements, object insertion, object displacement and more. In Fig. 4 we show qualitative comparisons of LoRWeB against the baselines. Notably, existing approaches either struggle with maintaining the content of the original image, or fail on some of the tasks. Crucially, some baselines struggle to maintain subject identities, *e.g.* the cat in RelationAdapter or the woman in VisualCloze, or to accurately capture the details of the transformation, *e.g.* the precise crown in the cat example. LoRWeB shows greater adaptability and succeeds in a wider range of tasks. Additional results appear in App. B.3 and App. B.6.

4.2 Quantitative Evaluations

Automated Evaluation Metrics. For quantitative evaluations, we follow prior work [8, 19, 52] and evaluate performance across standard metrics such as LPIPS [70] between the source and generated image, and CLIP directional similarity between both analogy pairs. In addition, we build on recent image editing work [28], which demonstrates that VLMs often better correlate with human preference than CLIP-based methods, and implement a VLM-based assessment protocol. Specifically, we conduct two VLM-based experiments: In the first, we provide Gemma-3 [55] with $\{\mathbf{a}, \mathbf{a}', \mathbf{b}, \mathbf{b}'\}$, and ask the VLM to evaluate the quality of results on two criteria: consistency with the source image, and accuracy of the applied transformation relative to the reference transformation.

Table 1: Results for the ablation study of LoRWeB described in Sec. 4.3, for different hyperparameter and architecture choices.

Model	Pres. ↑ (VLM)	Acc. ↑ (VLM)	LPIPS ↓	CLIP Dir. ↑	Pairwise VLM (%) ↑ LoRA $r=128$	ET	VC	RA
LoRWeB (full, $r=4, N=32$)	7.87	5.94	0.31	0.21	57.9	70.4	68.1	58.5
+ $r=16$	8.13	4.92	0.20	0.11	51.8	63.9	62.4	49.6
+ $r=16, N=8$	7.82	5.49	0.29	0.19	59.9	73.1	67.0	56.7
+ $N=16$	7.74	5.95	0.31	0.23	60.4	70.5	68.5	56.6
+ Tanh activation	7.94	4.49	0.18	0.09	48.2	58.3	51.8	42.1
+ 2×2 Enc. Input	7.90	5.75	0.28	0.20	61.9	73.3	68.2	53.9
+ SigLip2	7.83	5.82	0.31	0.21	59.0	71.7	71.5	55.5
+ SigLip2 & 2×2 Enc. Input	7.85	5.71	0.29	0.20	59.5	73.8	66.8	58.3

We name these metrics *Preservation (VLM)* and *Edit Accuracy (VLM)*, respectively. As a second quality metric, we take a 2-alternative-forced-choice design (2AFC). We show Gemma-3 $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$, the $\mathbf{b}'$ result of our model, and the $\mathbf{b}'$ result generated by a baseline, and ask it to select the image that best applies the analogy. We report this metric as *Pairwise VLM*. The prompts given to the VLM and further details appear in App. A.4. The results are shown in Fig. 5 and Fig. 6. When considering preservation and editing accuracy tradeoffs (Fig. 5), our model pushes the Pareto front, achieving high edit accuracy while better maintaining the input's structure and appearance.

User Study. Beyond automated metrics, we also conduct a two-alternative forced choice user study. We show each user a reference pair $(\mathbf{a}, \mathbf{a}')$, an input image $\mathbf{b}$, and two results (one from our model and one of a random baseline), in a randomized order, filtering out cases where no method succeeded in editing. Users are asked to select their preferred editing result. In total, we collected responses from 33 users covering 45 image pairs. The results (Fig. 6) align with the automated metrics, showing that users favor our approach over all baselines. In in App. A.4 we additionally evaluate the alignment between the scores of the VLM and the preferences of humans.

All in all, our experiments demonstrate that our approach can meaningfully improve on the existing state of the art, and better generalize to unseen tasks.

4.3 Ablations

We next study the importance of different components of LoRWeB.

Capacity Effect. We compare LoRWeB across modified capacities in both basis sizes N and ranks r. Specifically, we compare our original variation ($\{N=32, r=4\}$), with $\{N=8, r=16\}$, $\{N=16, r=4\}$ and $\{N=32, r=16\}$. We use the same evaluation setup as in Sec. 4.2. Results are reported in Tab. 1. Reducing the basis size while maintaining the capacity ($r=16, N=8$) leads to a slight drop in

performance, as does simply reducing capacity ($r = 4, N = 16$). This highlights the importance of a large basis for generalization. Similarly, a naïve increase in rank can hamper editability, which we hypothesize to be a consequence of the data, leading to increased overfitting. We provide additional capacity results for LoRWeB and a single LoRA with a higher capacity in App. B.2. Here, again, naïve parameter addition does not strictly correlate with better performance.

Similarity Normalizing Function. The normalization function choice in Eq. (4) can also affect the learned basis. For example, the used softmax is bound to $[0, 1]$, hence it cannot result in negative coefficients for any LoRA. An alternative approach is to use Tanh, which is instead bound to $[-1, 1]$. In practice, we find it to drastically underperform. We propose that this may be due to Tanh allowing the model to compose mixed LoRAs with much greater norms, possibly taking the model too far out of domain. Another alternative can be a differential activation function, as proposed by several recent work [41, 65]. However, we leave further investigation of activations to future work.

Layout of Encoder Input. In our approach, we elected to separately encode each of the conditioning analogy images using CLIP, and concatenate their representations. Our intuition is that CLIP requires resizing the image to 224×224, which can severely constrain the level of detail in each quadrant of the 2×2 grid that we provide Flux as a context. Moreover, concatenated features could allow the model to better understand which encoding represents each conditioning image (*i.e.* $\mathbf{a}$, $\mathbf{a}'$ and $\mathbf{b}$), allowing it to better reason over the analogy. We verify this experimentally by comparing to a version that provides CLIP with just the context image (the 2×2 grid). As seen in Tab. 1, this diminishes performance, mainly decreasing the editing-accuracy metrics.

Alternative Image Encoders. Although our approach uses CLIP [44] as the encoder backbone, we validate our robustness to an alternative, common choice: SigLIP2 [58]. The results in Tab. 1 indicate that this change does not significantly alter our performances. We leave further tuning of encoders to future work.

Importance of Prompts and Reference Images. We follow existing baselines and use prompts to augment the model's understanding. Since our goal is analogy based editing, and not simply text-based modification, we verify that indeed the output of our model depends not only on the text prompt, but on the analogy pair itself. Specifically, we conduct two complementing experiments. First, we examine how the same input image, $\mathbf{b}$, reacts to different reference pairs $\{\mathbf{a}, \mathbf{a}'\}$ under the same editing prompt. As can be seen in Fig. 7, the reference pair dictates the details of the analogy task, and particularly the visual details that are not captured by the prompts. For example, in the first row it copies the poster banner from the analogy image and adapts its specific style, and in the second row it matches the design and colors of the given crown. In comparison,

Fig. 7: Effect of different reference analogy pairs. LoRWeB directly leverages the analogy pair to understand the details of the proposed task, applying an edit that is beyond just text-based editing based on the given prompt. For example, when the prompt is "Give this create a crown of crystals", the analogy context passes information on the amount and color of the crystals.

we observe that some of the baselines are insensitive to the analogy pair, instead relying almost entirely on the prompt. Here, again, it can be seen some of the baselines opt to change the identity in the image (*e.g.* the dog of RelationAdapter and VisualCloze). For the second experiment, we examine how the same input triplet $\{\mathbf{a}, \mathbf{a}', \mathbf{b}\}$ react to reducing the detail level in the prompt. Here, we use an LLM [18] to omit details from prompts in our evaluation set (instruction appears in App. A.4). When comparing the pairwise VLM preference rate for such detail-reduced prompts, our results are preferred over RelationAdapter, Edit-Transfer and VisualCloze 59.4%, 70.4% and 66.8%. Qualitative examples can be seen in Fig. 8. These results show that LoRWeB can better rely on the analogy images, when compared to prior work which relies heavily on the text prompt itself. As both experiments demonstrate, our approach has learned to perform analogy-based editing, and to a greater degree than the existing baselines. We experiment with unaligned prompts and input images in App. B.5.

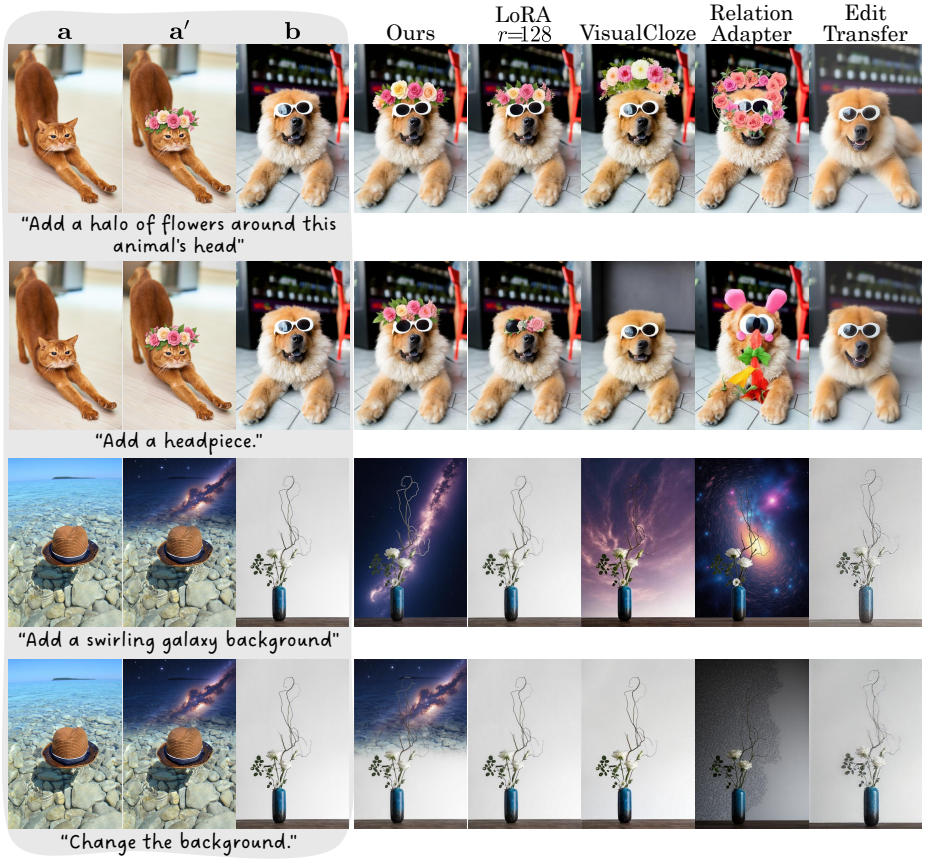

Fig. 8: Effect of less detailed prompts for the same input. While we follow prior approaches in using textual prompts to understand the context of the task, the details of the applied transformation are achieved by understanding the reference pair of images $\mathbf{a}$ and $\mathbf{a}'$. For example, when asked to "Add a halo of flowers around this animal's head", most approaches add a similar-looking flower crown to the dogs head. However, when the prompt changes to only "Add a headpiece", only LoRWeB understands from $\mathbf{a} : \mathbf{a}'$ the headpiece is a flower crown that should be added on top of the dog's head.

5 Discussion

We introduced LoRWeB, a modular framework for visual analogy completion that learns a basis of LoRA adapters and dynamically composes them using a shared encoder conditioned on the input analogy. Our approach addresses the limitations of single-adapter fine-tuning or multi-adapter optimization at inference time by enabling flexible, layer-specific adaptations to diverse and unseen transformations. Through extensive comparisons, we showed how LoRWeB outperforms and generalizes better than competing naive LoRA-based methods across various visual analogy tasks. However, this generalization is not with-

out limitations. For example, LoRWeB may still struggle with tasks that are significantly different from the training corpus (See App. B.7 for examples).

Additionally, a strong assumption in image analogies is the access to a reference image pair $\{\mathbf{a}, \mathbf{a}'\}$ where the noticeable change depicts the transformation alone, while keeping the rest of the image details exactly the same (*e.g.* depicting the same person in different poses in the same environment). In practical circumstances, finding such reference pairs in the wild, which convey the exact needed transformation, can be difficult. Nevertheless, we show in App. B.4 that similarity isn't strictly required, and LoRWeB can work on non-identical input pairs with textual guidance, within limit. Indeed, an interesting future work could explore better decomposing the components of a depicted transformation (*e.g.* the pose and the background), and allow for interactively choosing which components are used when applying the transformation to $\mathbf{b}$.

While our focus here is on visual analogy completion, a similar LoRA-basis approach could be broadly applicable, possibly replacing LoRAs in other tasks where generalization is needed. We hope to explore this direction in future work.

Acknowledgments

This research was partially supported by the Israel Science Foundation (grant no. 2318/22) and the Planning and Budgeting Committee of the Israeli Council for Higher Education. The authors are grateful to Matan Kleiner and Yoad Tewel for their insightful discussions and input throughout this work.

References

1. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer. In: ACM SIGGRAPH 2024 conference papers. pp. 1–12 (2024)
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
3. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in Neural Information Processing Systems **35**, 25005–25017 (2022)
4. Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: FLUX. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cai, S., Chan, E.R., Zhang, Y., Guibas, L., Wu, J., Wetzstein, G.: Diffusion self-distillation for zero-shot customized image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18434–18443 (2025)

7. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023)
8. Chen, L., Mao, Q., Gu, Y., Shou, M.Z.: Edit transfer: Learning image editing via vision in-context relations. arXiv preprint arXiv:2503.13327 (2025)
9. Cheng, M., Gong, J., Eldardiry, H.: Sci-LoRA: Mixture of scientific LoRAs for cross-domain lay paraphrasing. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 18524–18541 (2025)
10. Deutch, G., Gal, R., Garibi, D., Patashnik, O., Cohen-Or, D.: Turboedit: Text-based image editing using few-step diffusion models. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
11. Dravid, A., Gandelsman, Y., Wang, K.C., Abdal, R., Wetzstein, G., Efros, A., Aberman, K.: Interpreting the weight space of customized diffusion models. Advances in Neural Information Processing Systems **37**, 137334–137371 (2024)
12. Feng, W., Hao, C., Zhang, Y., Han, Y., Wang, H.: Mixture-of-LoRAs: An efficient multitask tuning method for large language models. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11371–11380 (2024)
13. Fischer, M., Li, Z., Nguyen-Phuoc, T., Bozic, A., Dong, Z., Marshall, C., Ritschel, T.: Nerf analogies: Example-based visual attribute transfer for nerfs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4640–4650 (2024)
14. Fišer, J., Jamriška, O., Lukáč, M., Shechtman, E., Asente, P., Lu, J., Sỳkora, D.: Stylit: illumination-guided example-based stylization of 3d renderings. ACM Transactions on Graphics (TOG) **35**(4), 1–11 (2016)
15. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023)
16. Gal, R., Lichter, O., Richardson, E., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: LCM-lookahead for encoder-based text-to-image personalization. In: European Conference on Computer Vision. pp. 322–340. Springer (2024)
17. Gong, Y., Song, Y., Li, Y., Li, C., Zhang, Y.: Relationadapter: Learning and transferring visual relation with diffusion transformers. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
18. Google: Introducing gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3/` (2025)
19. Gu, Z., Yang, S., Liao, J., Huo, J., Gao, Y.: Analogist: Out-of-the-box visual in-context learning with image diffusion model. ACM Transactions on Graphics (TOG) **43**(4), 1–15 (2024)
20. Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., Graßhof, S., Brandt, S.S., Michaeli, T.: Discovering interpretable directions in the semantic latent space of diffusion models. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9. IEEE (2024)
21. He, M., Liao, J., Chen, D., Yuan, L., Sander, P.V.: Progressive color transfer with dense semantic correspondences. ACM Transactions on Graphics (TOG) **38**(2), 1–18 (2019)
22. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)

23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
24. 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)
25. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. p. 327–340. SIGGRAPH '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383259.383295, https://doi.org/10.1145/383259.383295
26. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
27. Huang, C., Liu, Q., Lin, B.Y., Pang, T., Du, C., Lin, M.: LoraHub: Efficient cross-task generalization via dynamic loRA composition. In: First Conference on Language Modeling (2024)
28. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
29. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly DDPM noise space: Inversion and manipulations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12469–12478 (2024)
30. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
31. Li, Z.Y., Du, R., Yan, J., Zhuo, L., Li, Z., Gao, P., Ma, Z., Cheng, M.M.: Visualcloze: A universal image generation framework via visual in-context learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18969–18979 (October 2025)
32. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. ACM Transactions on Graphics (TOG) **36**(4), 1–15 (2017)
33. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
34. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
35. Liu, Y., Chen, X., Ma, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Unifying image processing as visual prompting question answering. In: International Conference on Machine Learning. pp. 30873–30891. PMLR (2024)
36. Lu, H., Chen, J., Yang, Z., Gnanha, A.T., Wang, F.L., Qing, L., Mao, X.: Pairedit: Learning semantic variations for exemplar-based image editing. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
37. Manor, H., Michaeli, T.: Zero-shot unsupervised and text-based audio editing using DDPM inversion. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 34603–34629. PMLR (21–27 Jul 2024)

38. Mao, F., Hao, A., Chen, J., Liu, D., Feng, X., Zhu, J., Wu, M., Chen, C., Wu, J., Chu, X.: Omni-effects: Unified and spatially-controllable visual effects generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7927–7935 (2026)
39. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
41. Misrahi, A., Chirkova, N., Louis, M., Nikoulina, V.: DiffLoRA: Differential low-rank adapters for large language models. arXiv preprint arXiv:2507.23588 (2025)
42. Ortiz, J.J.G., Guttag, J., Dalca, A.V.: Magnitude invariant parametrizations improve hypernetwork learning. In: The Twelfth International Conference on Learning Representations (2024)
43. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
46. Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. Advances in neural information processing systems **28** (2015)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.org/10.1109/CVPR52688.2022.01042
48. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
49. Ryu, S.: Cloneofsimo/lora: Low-rank adaptation for fast text-to-image diffusion fine-tuning (2023)
50. 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)
51. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
52. Song, X., Cui, J., Zhang, H., Shi, J., Chen, J., Zhang, C., Jiang, Y.G.: LoRA of change: Learning to generate LoRA for the editing instruction from a single before-after image pair. arXiv preprint arXiv:2411.19156 (2024)

53. Šubrtová, A., Lukáč, M., Čech, J., Futschik, D., Shechtman, E., Sỳkora, D.: Diffusion image analogies. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–10 (2023)
54. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
55. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
56. Tewel, Y., Gal, R., Samuel, D., Atzmon, Y., Wolf, L., Chechik, G.: Add-it: Training-free object insertion in images with pretrained diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)
57. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-free consistent text-to-image generation. ACM Transactions on Graphics (TOG) **43**(4), 1–18 (2024)
58. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
59. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
60. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)
61. Wang, Z., Jiang, Y., Lu, Y., He, P., Chen, W., Wang, Z., Zhou, M., et al.: In-context learning unlocked for diffusion models. Advances in Neural Information Processing Systems **36**, 8542–8562 (2023)
62. Wu, X., Huang, S., Wei, F.: Mixture of loRA experts. In: The Twelfth International Conference on Learning Representations (2024)
63. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)
64. Yang, Y., Peng, H., Shen, Y., Yang, Y., Hu, H., Qiu, L., Koike, H., et al.: Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. Advances in Neural Information Processing Systems **36**, 48723–48743 (2023)
65. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. In: The Thirteenth International Conference on Learning Representations (2025)
66. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)
67. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
68. Zhang, J.C., Xiong, Y.J.: Subject or style: Adaptive and training-free mixture of LoRAs. arXiv preprint arXiv:2508.02165 (2025)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)

70. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
71. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)