DiT-Edit: An Image Editing Framework for Diffusion Transformers

Antonio Mari (377119), Emanuele Nevali (358702), Matteo Santelmo (376844), Luca Sbicego (377536)
CS-503 Final Project Report

***Abstract*—Diffusion Transformers (DiTs) have recently redefined the state of the art in text-to-image synthesis, but their capacity for direct, training-free image editing remains unexplored. Drawing inspiration from existing editing pipelines, we propose a unifying framework built around six core design axes—noise initialization, QKV-level patching, layer and timestep selection, background consistency, and (optionally) text prompting. Instantiating this framework yields DiT-Edit, the first method to perform training-free image composition with DiTs: it seamlessly merges a user-provided foreground into an arbitrary background. Extensive experiments demonstrate that DiT-Edit outperforms existing baselines both qualitatively and via a large-scale user study on many domains. Remarkably, our framework can also be extended to support mixed text- and image-based edits, opening new avenues for multimodal image manipulation without additional training.**

I. Introduction

Diffusion Transformers (DiTs) [1] have emerged as the leading architecture for text-to-image (T2I) generation, offering high-quality and semantically rich outputs that surpass earlier U-Net-based approaches. While DiTs excel at synthesizing images from textual prompts, users often want the option to modify generated images based on new text inputs or reference images.

Recent research has explored both training-free and training-based methods to address these challenges; however, many of these techniques were developed for earlier U-Net-based models[2; 3], and their applicability to modern DiT architectures remains uncertain. Recently, a few methods emerged for text-based image editing in DiTs [4; 5], delivering high quality results even without additional training. However, to the best of our knowledge, no equivalent techniques exist for *image-based image editing* (also referred to as *image composition*) within the DiT framework. In our work, we aim to close this gap. Our contributions can be summarized as follows:

- We identify the key design choices in image-editing methods and introduce a unified framework that brings together diverse existing approaches. This framework serves as a foundation for developing novel text- and image-based editing methods.
- We present **DiT-EDIT**, the first DiT-based image-composition method that requires no additional training. A user study demonstrates clear and tangible improvements over prior methods.

II. Background

Deterministic Diffusion with DiTs: Deterministic diffusion models—such as flow matching [6]—are naturally expressed as ordinary differential equations (ODEs) that transport samples from a simple prior to the target data distribution. Concretely, we initialize $x_0 \sim \mathcal{N}(0, I)$ and follow the probability flow ODE

$$\mathrm{d}x_t = \Big(f(x_t, t) - \tfrac{1}{2}\, g^2(t)\, \nabla_{x_t} \log p(x_t) \Big)\, \mathrm{d}t, \quad t \in [0, T],$$

replacing the true score function $\nabla_{x_t} \log p(x_t)$ by a neural estimate s_θ. This process is deterministic and reversible: integrating forward transforms noise into data, while integrating backward recovers the noise.

Rectified Flow [7] is one prominent example of such invertible diffusion that often requires fewer denoising and inversion steps than denoising diffusion implicit models (DDIM) [8]. In our work, we build on Flux-dev [9], a recent diffusion transformer implemented within the rectified-flow framework. Flux-dev achieves state-of-the-art image synthesis and can generate high quality samples in just 28 denoising steps.

Image Composition Task: Given a background and foreground image pair (denoted as `bg` and `fg`), the goal is to transfer a selected element from `fg` to a specific location in `bg` while preserving the content of `bg`. This operation should account for differences in size, perspective, or style, adapting the element to the context of `bg`. The only training-free solutions proposed are TF-ICON [10], which also introduce a public evaluation benchmark, and PrimeComposer [11]. More training based approaches exist, requiring heavier computation and sometimes even pre-training models from scratch [12; 13; 14; 15; 16]. However, all of these methods predominantly focus on older, U-Net based architectures.

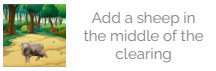
Text-based editing: Contrary to image-based editing, users seeks to edit an image by textually describing ("prompting") the required changes, without any visual inputs. The most notable training-free approaches for DiT are KV-Edit [4], StableFlow [5] and RF-edit [17], while older approaches include Prompt-to-Prompt [18] and MasaCtrl [19]. Given the prevalence of existing research and solutions for prompt-based editing, we also take relevant existing text-based solutions into account.

III. Method

A. Framework for Generative Image Editing

We analyze existing image editing methods that use diffusion models and identify the key design choices underlying their formulations. Outlining six design choices, we compose a framework that provides practical guidelines to practitioners to design novel techniques. Table I summarizes this framework and classifies three existing solutions and

Table I: Framework for generative image editing. Compared our method against image-editing baseline TF-ICON as well as text-based editing methods KV-Edit and Stableflow across 5 main design axes.

Design Axis	**Illustration**	**DiT-EDIT** (ours)	**TF-ICON** [10]	**KV-Edit** [4]	**StableFlow** [5]
Noise Initialization: Setting up the starting noise x_T.		Composed Noise	Copy & Paste, Inversion, transition noise	Inversion, XOR noise	Inversion w/ latent nudging
QKV Patching: How to patch the QKV of the `fg` and `bg`		Patch foreground Q, K, V	Patch foreground Q, K	Regenerate foreground Q, K, V	Replace latents
Affected Layers: In which layers to perform QKV patching.		All layers	All layers	All layers	Vital (most impactful layers only)
Affected Timesteps: At which diffusion timesteps to perform QKV patching.		Patching before step τ_α	Patching before step τ_α	All timesteps	All timesteps
Background Consistency: How to preserve the background		`bg` latents before step τ_β	`bg` latents before step τ_β	Cache Q, K, V of background	None
Text-based: uses prompt to perform text-based editing		Image with support for prompting	Image only	Prompting only	Prompting only

our proposed method *DiT-EDIT* along each axis. The key design choices are as follows.

1. Noise Initialization. Diffusion denoising processes generate images from random gaussian starting noise. To edit an arbitrary image $x_T \sim \mathcal{D}$, we need first to invert it into its corresponding starting noise $x_0 \sim \mathcal{N}(0, I)$. This way, we can get the starting noise representation for the foreground and background images `fg` and `bg`. Editing methods use these starting noises individually or they compose it in some way to start the denoising process. This choice is crucial because all this noise implicitly contains all the information about the image that the model is going to generate.

2. QKV-Patching. During denoising, most existing methods patch queries, keys and/or values of transformer blocks from some reference generations to the one used to perform image editing. This operation forces adherence to the reference, imposing semantic and structure. TF-ICON implements an operation called "attention composition", can be proved equivalent to patching queries and keys. KV-Edit and Stableflow patch only keys and values of `bg` in all locations.

3. Affected Layers. while most methods apply QKV-patching on all layers, such operation can be applied to selected transformer layers instead. For DiTs, recent work [5] showed that specific layers play an vital role on the final output, opting to inject attention features only in these layers to construct their text-based editing solution.

4. Affected Timesteps. The final composed image is obtained after multiple forward passes through the DiT/U-net and each pass represents one diffusion timestep. QKV-patching can be applied to only some steps in the generation process, defining a schedule. Intuitively, injecting attention features early leads to a larger impact in the generation, since the earlier timesteps contribute to set-up low-frequency features, while latter timesteps refine the generation, acting on details and on providing higher quality high-frequency features.

However, injecting in final timesteps might hinder the blending of the foreground and background, with unnatural results that resemble just overlapping `fg` and `bg`. Thus, setting up the right QKV-patching schedule regulates the freedom the model has in deviating from the `fg` image.

5. Background Consistency. If the model can freely diffuse from the composed noise it will inevitably lead to deviations in the background, even in the case of flow-based models. Empirical observations [4] show that inversion only achieves similarity rather than perfect consistency in content. Thus, a background preservation method needs to be introduced. For instance, TF-ICON imposes latents of `bg` after each denoising step, while KV-Edit caches keys and values of the inversion process and reintroduces those, copying them during denoising. The latter strategy is more costly due to prohibitive memory consumption.

Thus, popular editing methods set an additional threshold of timestep τ_β, after which to impose the latents of the background diffusion stream to improve background preservation [10; 20].

6. Prompting: Prompt-based editing methods leverage prompt to edit images. This is normally straightforward, so that while generating the edited image the input prompt expresses the desired changes. Moreover, as [10] show, the choice of prompt might be important for image-composition as well. They introduce an "exceptional prompt" which allows them to achieve better inversion. However, while they work on U-net models, we discuss the use of prompt for Flux in our experiments.

B. DiT-EDIT

To design DiT-EDIT, we drew inspiration from existing copy-paste and KV-patching methods (Table I). Algorithm 1 and figure 1 summarizes our pipeline.

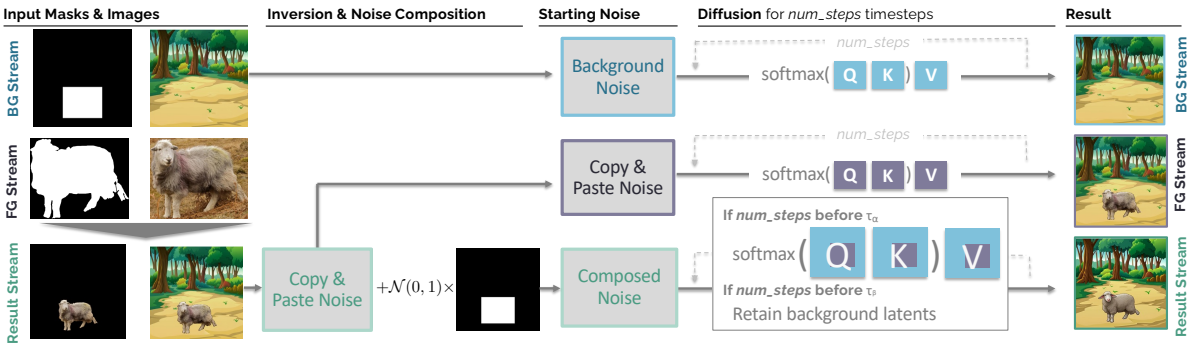


Figure 1: Flowchart detailing DiT-EDIT.

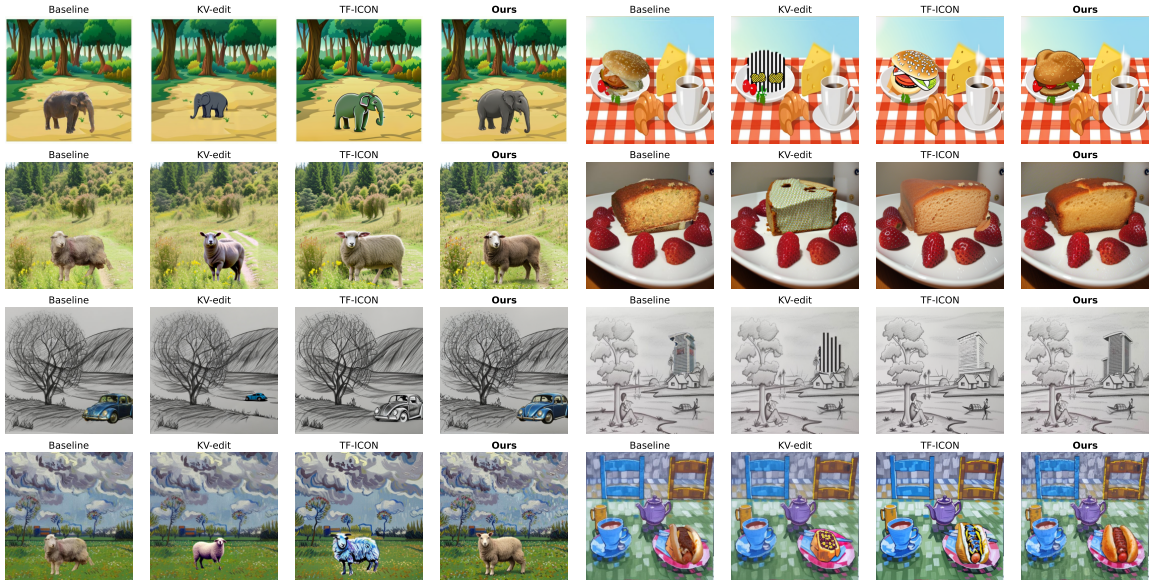


Figure 2: Qualitative Evaluation of our method compared to baselines for 8 samples across four domains, in order: Real-Cartoon, Real-Real, Real-Sketch and Real-Painting

Copy-paste image DiT-EDIT begins by forming a simple "copy-paste" composite, denoted as `cp`, which later serves as both a baseline and a structural guide. First, we extract the foreground subject from `fg` using its binary mask $M_{\texttt{fg}}$ (1 inside the subject region, 0 elsewhere) to discard all background pixels. Next, we specify the target insertion area on `bg` via a rectangular bounding box $M_{\texttt{bg}}$ (1 inside the box, 0 outside). We resize the segemented `fg` (referred to as `res_fg` in Algorithm 1) to match the dimensions of $M_{\texttt{bg}}$ (allowing for scaling or cropping). Finally, we overlay `res_fg` onto `bg` at the location defined by $M_{\texttt{bg}}$. The result is our initial copy-paste image `cp`.

Noise initialization We encode both `bg` and `cp` into latents via the DiT VAE, then invert each through T diffusion steps—using the same reversed time-step schedule and latent nudging from [5]—to recover their initial noise $x_0^{\texttt{bg}}$ and $x_0^{\texttt{cp}}$.

We then compose the initial noise:

$$x_0^{\texttt{comp}} = (1 - M_{\texttt{fg}}) \odot x_0^{\texttt{bg}} + M_{\texttt{fg}} \odot \left((1-\alpha)\, x_0^{\texttt{cp}} + \alpha\, z\right)$$

where $z \sim \mathcal{N}(0, I)$ introduces independent random noise (with same shape as `bg`), and $\alpha \in [0, 1]$ controls the strength of this noise blend to aid seamless subject–background integration.

Denoising with QKV-Patching. Building on KV-Edit's KV-patching and TF-ICON's attention composition (equivalent to QK-patching), we apply full QKV-patching across all transformer layers. To perform this patching, we simply copy queries, keys and values of `bg`and `cp`, then use again the segmentation mask to compose them

$$\texttt{qkv} \leftarrow (1 - M_{\texttt{fg}}) \odot \texttt{qkv\_bg} + M_{\texttt{fg}} \odot \texttt{qkv\_cp}$$

and replace the composite stream's QKV tensors accordingly.

The hyperparameter $\tau_\alpha \in [0, 1]$ determines how long this patching is applied: lower values give the model more freedom but will result in lower subject fidelity, whereas higher values enforce stronger adherence to the copied structure, at the expense of smooth blending.

Background consistency To improve background consistency, we overwrite the `comp`'s background latents with those from `bg` stream. Here, $\tau_\beta \in [0, 1]$ controls how long

Algorithm 1 DiT-EDIT

Input: The background image $\texttt{bg}$, the foreground image $\texttt{fg}$, the target bounding box $M_{\texttt{bg}}$, the segmentation mask $M_{\texttt{fg}}$, thresholds $0 \leq \tau_A, \tau_B \leq 1$, $\texttt{prompt}$ for composition, num of steps T, $0 \leq \alpha \leq 1$ for noise initialization.
Output: The composition result $\texttt{comp}$.

1. Create Copy-paste image
$\texttt{res\_fg} \leftarrow \mathsf{resize}(\texttt{fg} \odot M_{\texttt{fg}}, \mathsf{size\_of} = M_{\texttt{bg}})$.
$\texttt{cp} \leftarrow \mathsf{superimpose}(\texttt{res\_fg}, \texttt{bg}, \mathsf{where} = M_{\texttt{bg}})$

2. Noise Initialization
$x_T^{\texttt{bg}} \leftarrow \mathsf{VAE\_encode}(\texttt{bg})$, $x_T^{\texttt{cp}} \leftarrow \mathsf{VAE\_encode}(\texttt{cp})$
$x_0^{\texttt{bg}} \leftarrow \mathsf{invert}(x_T^{\texttt{bg}}, T)$, $x_0^{\texttt{cp}} \leftarrow \mathsf{invert}(x_T^{\texttt{cp}}, T)$
$z \sim \mathcal{N}(0, I)$ (same shape as $\texttt{bg}$)
$x_0^{comp} = (1 - M_{\texttt{fg}}) \odot x_0^{\texttt{cp}} + M_{\texttt{fg}} \odot ((1-\alpha)x_0^{cp} + \alpha z$

3. Image Composition
for $t = 0, \ldots, T-1$ **do**
  $(x_{t+1}^{\texttt{bg}}, \texttt{qkv\_bg}) \leftarrow \mathsf{DiTdenoise\_cache}(x_t^{\texttt{bg}}, t)$
  $(x_{t+1}^{\texttt{cp}}, \texttt{qkv\_cp}) \leftarrow \mathsf{DiTdenoise\_cache}(x_t^{\texttt{cp}}, t)$
  $\texttt{qkv} \leftarrow (1 - M_{\texttt{fg}}) \odot \texttt{qkv\_bg} + M_{\texttt{fg}} \odot \texttt{qkv\_cp}$
  if $t \leq \tau_A \times T$ **then**
    $x_{t-1}^{\texttt{comp}} \leftarrow \mathsf{DiTdenoise}(x_t^{\texttt{comp}}, t, \texttt{prompt}, \texttt{qkv})$
  else
    $x_{t-1}^{\texttt{comp}} \leftarrow \mathsf{DiTdenoise}(x_t^{\texttt{comp}}, t, \texttt{prompt})$
  end if
  if $t \leq \tau_B \times T$ **then**
    $x_{t+1}^{\texttt{comp}} \leftarrow x_{t+1}^{\texttt{comp}} \odot M^{\texttt{bg}} \; + \; x_{t+1}^{\texttt{bg}} \odot (1 - M^{\texttt{bg}})$
  end if
end for
$\texttt{comp} \leftarrow \mathsf{VAE\_decode}(x_T^{\texttt{comp}})$
return $\texttt{comp}$

this background enforcement is applied (typically $\tau_\beta \geq \tau_\alpha$). We choose this strategy because it is far more efficient than full KV-edit caching.

IV. Experiments

In this section, we discuss our experimental setup and report qualitative and quantitative evaluation.

We use the only available benchmarking dataset [10] for image-composition which contains 332 samples, spanning four visual styles: photorealism, pencil sketching, oil painting and cartoon animation. Each sample is composed of a background image corresponding to the style ($\texttt{bg}$), a real-style image containing the subject ($\texttt{fg}$) along with its segmentation mask and bounding box for location.

For all experiments we use Flux-dev and our default hyperparameters are $T = 50$, $\tau_\alpha = 0.4$, $\tau_\beta = 0.8$, $\alpha = 0.05$.

A. Qualitative Evaluation

Figure 2 shows eight examples across all evaluation domains for our method and all baselines.

Our method consistently achieves visually pleasing and semantically consistent results, especially when compared to our main image-composition baseline, TF-ICON, which sometimes fails at style transfer. Please refer to the appendix and ablation study for more examples of results.

Interestingly, our method also seems to natively support prompt-based editing in addition to considering the $\texttt{fg}$ image. We explore this phenomenon in Figure 3, where we add, alter or superimpose objects while generating the composed image for 3 examples on the Real-Real domain. DiT-EDIT shows great promise to be a first multi-modal editing algorithm.

B. Quantitative Evaluation

To evaluate our results quantitatively, we leverage a range of metrics: HPSv2 score [21] to measure the quality of the inserted $\texttt{fg}$ subject, for semantic consistency we calculate the DINOv2, CLIP-FG and LPIPS similarity metrics between the inserted and original $\texttt{fg}$ subject. Similarly, we measure the CLIP-TXT similarity to the given prompt describing the image. For background consistency we measure deviations in the original background by calculating the mean squared error (MSE) of the background.

The results averaged over categories are presented in Table II. Reliable quantitative evaluation seems challenging for the task at hand. While our method performs on par or better on all metrics, no difference is statistically significant. The quantitative evaluation does not seem to reflect the large qualitative improvement and we hypothesize that some of these scores are not representative of human judgment. This is exemplified by the scores attained by the Copy-Paste baseline, whose output looks unnatural yet often reaches higher scores.

C. User Study

To address the shortcoming of the quantitative evaluation, we conduct a user study, collecting human preferences on outputs from our method and the three considered baselines on the benchmark data. Specifically, we show users the input images, combined prompt and bounding box, followed by outputs from the models, and let them choose their single preferred output. Each session collects between 4 and 12 human preferences for images from all four categories. The users are instructed to consider subject fidelity, background consistency and overall look of the image. The results are reported for 97 sessions for a total of 1021 preferences.

Figure 4 shows the results of the study, grouped by domain. Our method outperforms all baselines in photorealistic and real-to-cartoon domains, while struggling more in domains requiring a stronger style transfer, especially real-to-sketch. This confirms the results of the qualitative evaluation. On the most important application domain, photorealism, our method outperforms existing solutions.

D. Ablations

Here we discuss the main insights of our ablation studies. Full examples and complementary results as reported in Appendix A. Note that we focus on qualitative evaluation, given the earlier findings.

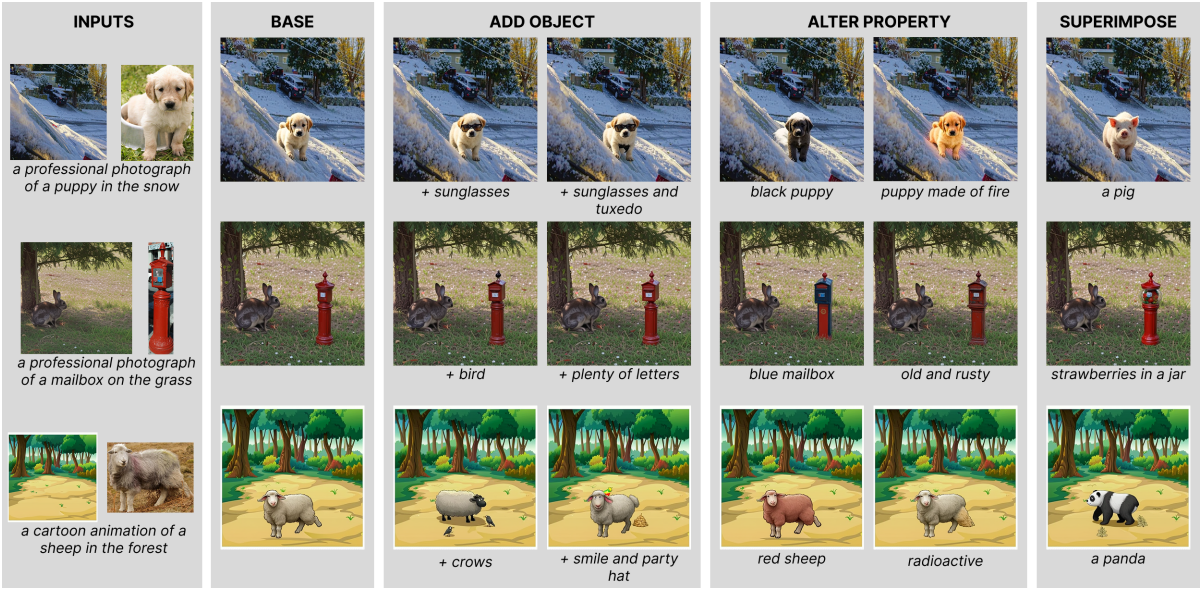


Figure 3: Generations for regular composition with DiT-Edit ("Base") and with additional prompt-based edits: adding objects, altering properties or superimposing a different subject.

Table II: Macro-Averaged Model Performance Across All Categories (Mean ± SE). Significant results are **bolded**.

Model	HPSv2-FG $\times 10^2$ ↑	DINOv2 Sim ↑	CLIP-Txt Sim ↑	CLIP-FG Sim ↑	LPIPS ↑	BG MSE $\times 10^3$ ↓
KV-EDIT [10]	30.0 ± 0.9	0.171 ± 0.018	0.263 ± 0.007	0.614 ± 0.011	0.873 ± 0.011	2.0 ± 0.0
Copy-Paste	32.9 ± 0.9	0.296 ± 0.018	0.288 ± 0.007	0.675 ± 0.013	0.866 ± 0.012	**1.0 ± 0.0**
TF-ICON [10]	33.5 ± 0.9	0.269 ± 0.020	0.289 ± 0.006	0.661 ± 0.012	0.867 ± 0.012	10.0 ± 1.0
DiT-EDIT (Ours)	33.6 ± 0.9	0.284 ± 0.019	0.289 ± 0.006	0.663 ± 0.012	0.863 ± 0.012	4.0 ± 0.0

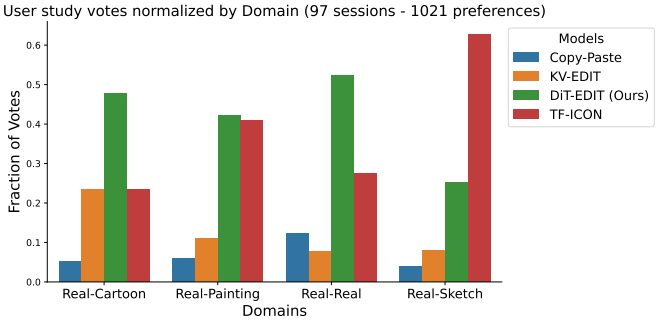


Figure 4: Results of user study as proportion of votes per model and domain, normalized on each domain.

Role of QKVs in DiTs: Our ablation studies on whether to inject Queries, Key and/or Values lead to inconsistent and inconclusive results that do not align with previous literature developed for older models. Notably, in their study on appearance transfer across images, [22] showed that in U-net based models queries tend to capture the structure of objects, whereas keys encode the semantic information and values carry the actual appearance details. Our experiments indicate that this might not necessarily be the case for DiTs. We leave the investigation of these effects to future work.

Importance of Prompt: The text-prompt (supplied by the benchmark data) plays an important role in the quality of the composed images, especially in the Real-Real domain. This result suggests that the interplay between image and text modalities plays is crucial in blending semantic and structural information in a reliable way. Interestingly, for additional prompt-based edits work better when patching only in specific "vital" layers, in line with the Stableflow method [5].

V. Conclusion and Limitations

In this work, we present DiT-EDIT, the first image-composition method native to DiT models. Our methods outperforms existing solutions under qualitative review and a large-scale user study.

Our analysis shows that our method still struggles with style transfer on the Real-to-Sketch domain. Further, quantitative comparisons remained inconclusive due to a shortcoming in commonly employed metrics for this task. Similarly, we only reported results for the FLUX-dev as the underlying DiT model.

In designing our solution, we also distilled six key design aspects of image-editing methods and composed them in a framework, opening up the design of new approaches.

Interestingly, preliminary experiments suggest that the very same method can also be employed as a text- and image-based editing method, unifying the two paradigms of editing solutions. This emerging aspect needs requires deeper exploration, which we leave for future work.

VI. Individual contributions

A.M. conceptualized the project idea. A.M., E.N., M.S., and L.S. contributed equally to the development of the codebase, brainstorming and implementation of research ideas, interpretation of results and writing the report.

References

[1] W. Peebles and S. Xie, "Scalable diffusion models with transformers," 2023. [Online]. Available: https://arxiv.org/abs/2212.09748

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *CoRR*, vol. abs/2010.02502, 2020. [Online]. Available: https://arxiv.org/abs/2010.02502

[4] T. Zhu, S. Zhang, J. Shao, and Y. Tang, "Kv-edit: Training-free image editing for precise background preservation," *arXiv preprint arXiv:2502.17363*, 2025.

[5] O. Avrahami, O. Patashnik, O. Fried, E. Nemchinov, K. Aberman, D. Lischinski, and D. Cohen-Or, "Stable flow: Vital layers for training-free image editing," 2025. [Online]. Available: https://arxiv.org/abs/2411.14430

[6] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2403.03206

[7] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," 2022. [Online]. Available: https://arxiv.org/abs/2209.03003

[8] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022. [Online]. Available: https://arxiv.org/abs/2010.02502

[9] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

[10] S. Lu, Y. Liu, and A. W.-K. Kong, "Tf-icon: Diffusion-based training-free cross-domain image composition," 2023. [Online]. Available: https://arxiv.org/abs/2307.12493

[11] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "Primecomposer: Faster progressively combined diffusion for image composition with attention steering," 2024. [Online]. Available: https://arxiv.org/abs/2403.05053

[12] X. Chen, Y. Feng, M. Chen, Y. Wang, S. Zhang, Y. Liu, Y. Shen, and H. Zhao, "Zero-shot image editing with reference imitation," 2024. [Online]. Available: https://arxiv.org/abs/2406.07547

[13] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 381–18 391.

[14] L. Lu, J. Li, B. Zhang, and L. Niu, "Dreamcom: Finetuning text-guided inpainting model for image composition," 2024. [Online]. Available: https://arxiv.org/abs/2309.15508

[15] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu, "Controlcom: Controllable image composition using diffusion model," 2023. [Online]. Available: https://arxiv.org/abs/2308.10040

[16] D. Winter, M. Cohen, S. Fruchter, Y. Pritch, A. Rav-Acha, and Y. Hoshen, "Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion," 2024. [Online]. Available: https://arxiv.org/abs/2403.18818

[17] J. Wang, J. Pu, Z. Qi, J. Guo, Y. Ma, N. Huang, Y. Chen, X. Li, and Y. Shan, "Taming rectified flow for inversion and editing," 2024. [Online]. Available: https://arxiv.org/abs/2411.04746

[18] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022. [Online]. Available: https://arxiv.org/abs/2208.01626

[19] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," 2023. [Online]. Available: https://arxiv.org/abs/2304.08465

[20] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 208–18 218.

[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. [Online]. Available: https://arxiv.org/abs/1801.03924

[22] Y. Alaluf, D. Garibi, O. Patashnik, H. Averbuch-Elor, and D. Cohen-Or, "Cross-image attention for zero-shot appearance transfer," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

APPENDIX

The following images report generations obtained with DiT-Edit on samples from all the domains of the benchmark. In particular, for each ablated hyperparameter we fix all the others with the following default values: $\tau_\alpha = 0.4$, $\tau_\beta = 0.8$, α-noise $= 0.05$, num_steps $= 50$, with Q, K and V are injected at all layers. The copy-paste baseline is reported as reference.

QKV-patch ablation: We conduct an ablation to investigate the importance and the role of injecting the queries keys and values of the background and the foreground image during the denoising process of our output. For the sake of this purpose we fixed all the other hyperparameters,

- we picked τ_α to be 1 so that the injection is stronger during the denoising;
- we set $\alpha = 0$ so to avoid additional randomness that might change the final results.

. Figure 5, presents some examples of results. For each benchmark example in each category we run 4 different denoising processes and we perform one of the alternatives:

1) QKV-patching
2) QK-patching (ablating the effect of Values)
3) QV-patching (ablating the effect of Keys)
4) KV-patching (ablating the effect of Queries)

Results are suprising and in constrast with the previous literature [10; 22]: the impact of each component seems to be negligible. Future work might try to focus to explore why this is the case.

Guidance: classifier free guidance stands as an additional hyperparameter that dictates adherence to the prompt. For the sake of completeness, we show that using a high value for guidance might make the foreground deviate from the original image (see fig. 10, last row). Our default value 3 seems a solid choice, not interfering neither with subject nor background.

τ_α and τ_β: We start by investigating the role of τ_α and τ_β independently. As expected, when fixing all parameters but τ_α we observe (fig. 6) that lower values make the generative process explore more diverse trajectories, often resulting in a more pleasing style transfer and blending with the background, while lacking more accurate feature of the specific subject. Similarly, fig. 7 show the effect of τ_β on background preservation. Despite achieving a better consistency, we can also observe that higher values of this parameter seem to be inducing the emergence of artifacts particularly when background is rich of textures.

Time-steps: To identify the best trade-off value for the number of steps of both inversion and denoising process, we iterate generation with values ranging from 1 to 100 8, establishing 50 to be our reference value as it represents a good trade-off between quality of generations and computation time.

Prompting: Despite imposing structural and semantic features through QKV patching, we investigate the impact of having access to the composed prompt. This is also motivated by the success of recent text-based editing methods [5; 4]. Figure 11 shows 2 examples for each domain comparing the generation with and without access to the prompt. Some images show completely destroyed generations while others only demonstrate minor degradation of the output. This result suggests that the interplay between image and text modalities plays an important role in blending semantic and structural information in a reliable way.

Alpha Noise: As shown in fig. 9, the additional noise we add with the coefficient α does not improve the subject-background blending, rather resulting in a smoother and blurry subject that is not desirable.

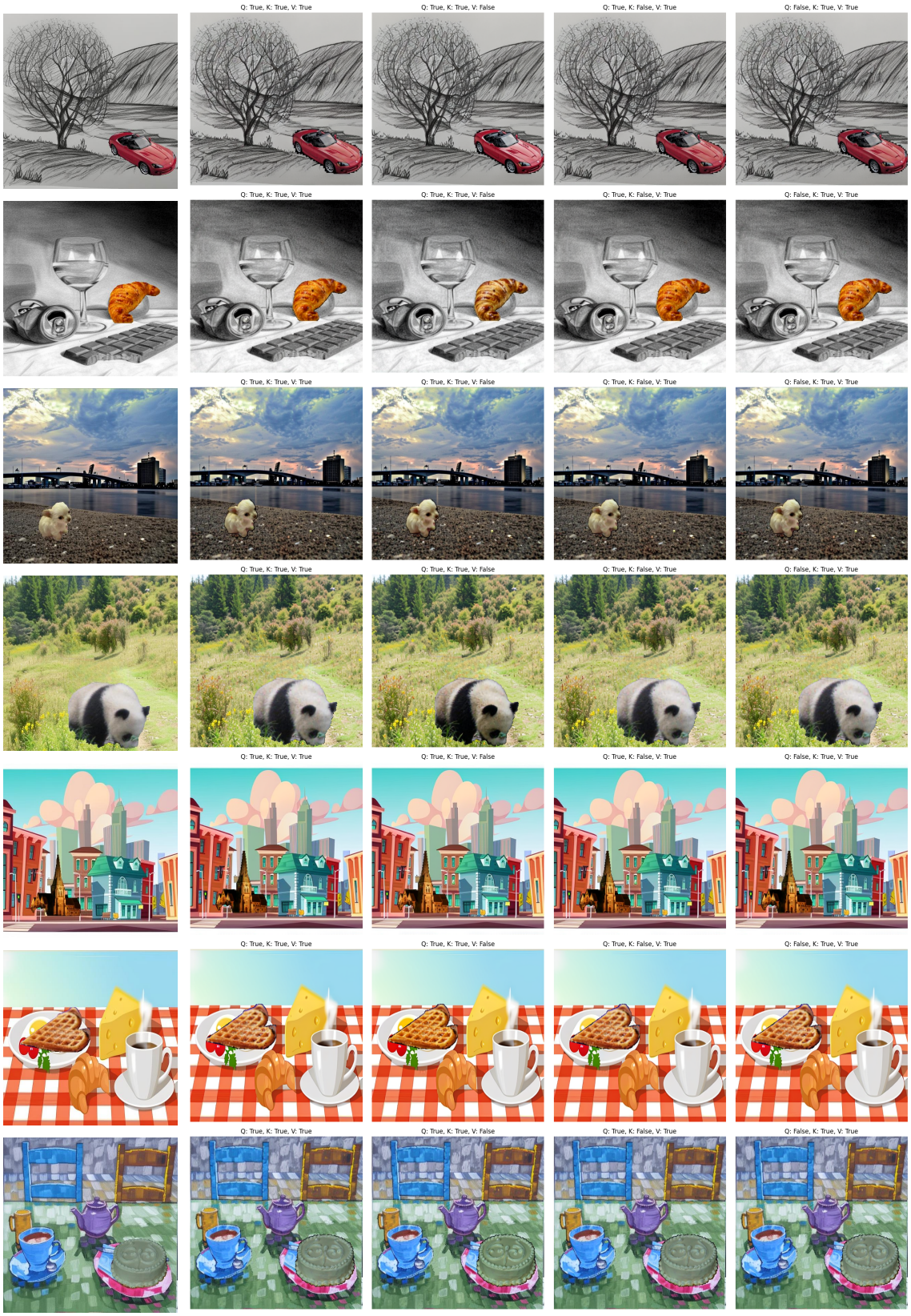


Figure 5: The first column is the baseline, so the image whose noise is used to initialize the output stream, then for each sample we show the model's output in 4 different settings: while injecting all at the same time queries, keys and values, injecting only queries and keys, injecting only queries and values, and finally injecting only keys and values. All these generations are ran with $\tau_\alpha = 1$ and with no random noise (e.g. $\alpha = 0$)

Figure 6: Generations obtained varying τ_α, demonstrating the trade-off between accuracy of subject transfer (high values) and blending and style transfer (lower values).

Figure 7: Generations with varying values of τ_β, demonstrating increased background consistency with prolonged latent injection.

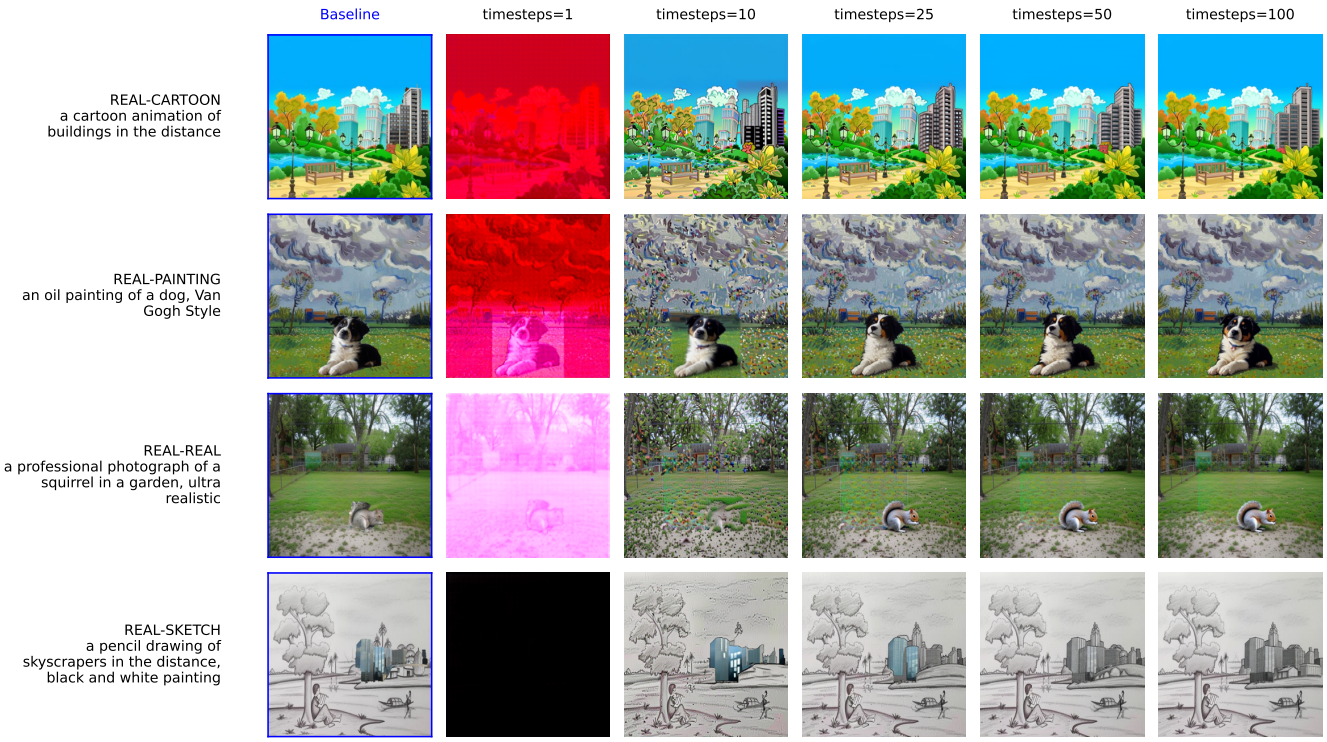


Figure 8: Outputs of composition with different numbers of total timesteps. The same number of timesteps is used both for inversion and denoising. Despite 100 steps resulted in better looking outputs, we fix our reference value to 50 as a good trade-off between quality of results and computation time.

Figure 9: Ablation on α-noise, i.e. the scale of the random noise added to the initial noise in the bounding box area

Figure 10: Ablation on guidance scale.

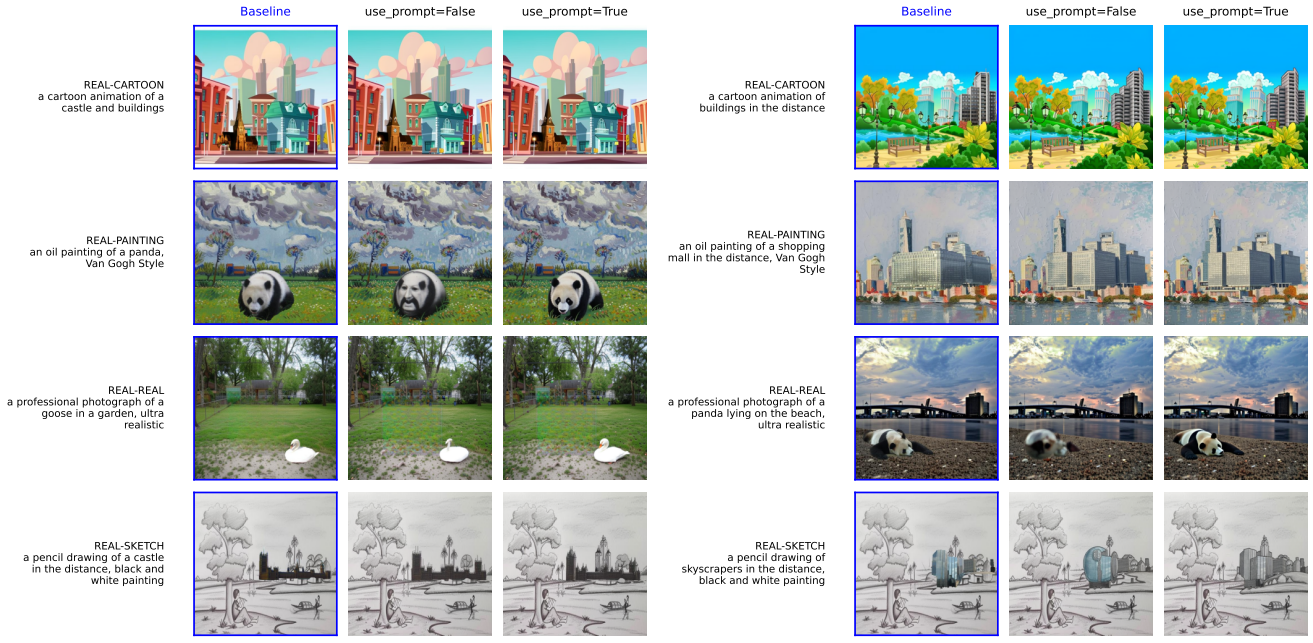


Figure 11: Impact of attending on a prompt describing the composed image during the denoising process.