UNIFUSION: VISION-LANGUAGE MODEL AS UNIFIED ENCODER IN IMAGE GENERATION AND EDITING

Yu-Teng Li   Manuel Brack   Sudeep Katakol   Hareesh Ravi   Ajinkya Kale
Adobe
yutengl@adobe.com

ABSTRACT

Most generative models rely on separate encoders for text and images (e.g., large language models and VAE latents), which complicates high-fidelity editing and limits cross-modal knowledge transfer due to heterogeneous embedding spaces. We present UNIFUSION, a framework for diffusion generative models conditioned solely on a frozen vision-language model (VLM) that serves as a unified multimodal encoder. UNIFUSION consists of two key components. First, Layerwise Attention Pooling (LAP) aggregates representations across VLM layers to capture both high-level semantics and fine-grained details for both image and text. Second, we introduce VLM-Enabled Rewriting Injection with Flexible Inference (VERIFI), which conditions the diffusion transformer (DiT) on rewritten text tokens produced in-model by the conditioning VLM, improving distribution alignment between tasks, while leveraging the VLM's reasoning.

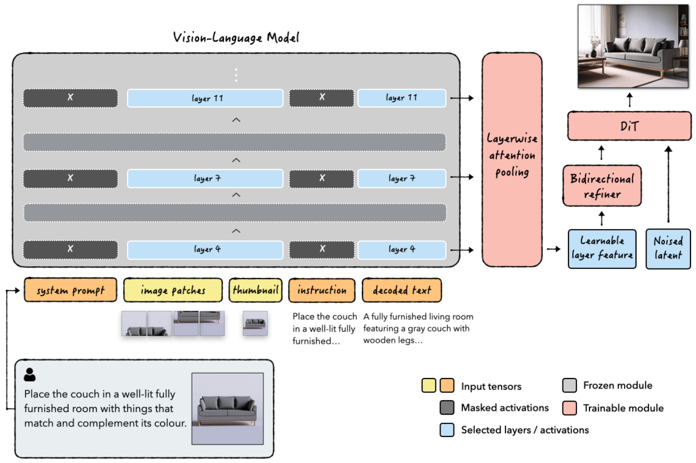
To the best of our knowledge, UNIFUSION is the first architecture to perform competitive image editing using only VLM-based input conditioning, without auxiliary signals from a VAE or CLIP. With an 8B VLM and an 8B DiT, UNIFUSION surpasses Flux.1 [dev] and BAGEL on DPG-Bench using a smaller training set, and compares favorably to Flux.1 Kontext [dev] and Qwen-Image-Edit on editing without post-training. Moreover, the unified-encoder framework with LAP yields emergent behaviors, including zero-shot multi-reference generation despite training only on single-reference pairs, and capability transfer where editing training improves text-to-image quality both quantitatively and qualitatively.

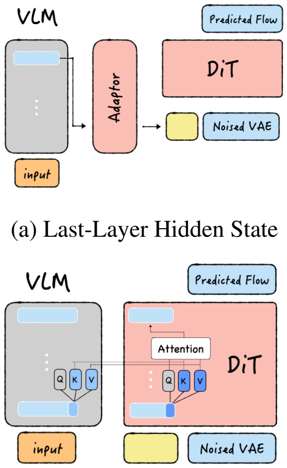
1 INTRODUCTION

Current image-generation models typically condition on separate representation spaces for text and image inputs. Most commonly, T5 embeddings (Chung et al., 2024) for text and variational autoencoder (VAE) latents for images (Seedream et al., 2025; Labs et al., 2025). However, these encoders operate at fundamentally different levels of abstraction: T5 captures high-level semantic meaning presented via text prompts while VAEs preserve low-level pixel-level detail from images. This mismatch is particularly evident in editing tasks, where models struggle to balance content preservation with instruction adherence, often producing either unnatural copy-paste artifacts or excessive modifications (Wang et al., 2025). We argue that separate encoding spaces force the DiT to expend capacity aligning heterogeneous features rather than synthesizing images, and that a unified semantic space can alleviate this burden. VLMs naturally offer such a shared representation for both text and images, but prior works conditioning on VLM features report failure to preserve the fine-grained visual details required for high-fidelity editing (Bellagente et al., 2023; Team, 2025).

We propose UNIFUSION, a framework for building image generation and editing models with unified text and image encoding. A frozen VLM serves as a unified encoder for both input modalities, eliminating the need for separate conditioning spaces. The framework comprises two key components: (1) Layerwise Attention Pooling (LAP), which aggregates information across multiple VLM layers to capture both fine-grained visual details and high-level semantic abstractions, and (2) VLM-Enabled Rewriting Injection with Flexible Inference (VERIFI), which only exposes the

Project page: http://thekevinli.github.io/unifusion/


Vision-Language Model
X
layer 11
X
layer 11
X
layer 7
X
layer 7
X
layer 4
X
layer 4
Layerwise attention pooling
DiT
Bidirectional refiner
Learnable layer feature
Noised latent
system prompt
image patches
thumbnail
instruction
decoded text
Place the couch in a well-lit fully furnished…
A fully furnished living room featuring a gray couch with wooden legs…
Place the couch in a well-lit fully furnished room with things that match and complement its colour.
Input tensors
Frozen module
Masked activations
Trainable module
Selected layers / activations



(a) Last-Layer Hidden State

(b) Layerwise KV Fusion

Figure 1: UNIFUSION architecture and inference paradigm. We extract multimodal representations from multiple layers of a frozen LLM and aggregate with a learnable layerwise attention pooling (LAP) module. A subsequent refiner counteracts the VLM's position bias due to causal attention. VERIFI rewrites the original input in-context. The rewritten tokens used for DiT conditioning leverage the VLM's reasoning capabilities to contextualize the target scene into a unified representation.

Figure 2: Options for VLM conditioning. Blue blocks within VLM and DiT denote selected layers, red for trainable modules, and gray for frozen modules.

DiT a rewritten target prompt based on the original user input. VERIFI reduces distribution shift between different input prompt formats (e.g. instructions in editing vs descriptions in generation) and incorporates VLM's reasoning capabilities and world knowledge into the representations.

We demonstrate the effectiveness of the UNIFUSION by training a single model that achieves competitive performance on both text-to-image generation and editing compared to strong contemporaries, without requiring any supervised fine-tuning or reinforcement learning. Notably, the resulting model exhibits remarkable zero-shot generalization capabilities: it handles multi-reference image inputs despite being trained only on single-reference editing data, and can perform image-to-image variations when exclusively trained on text-conditional generation. We further observe cross-task positive transfer, where training on editing tasks improves the model's text-to-image prompt adherence and aesthetic quality.

Our contributions can be summarized as follows: We propose (1) UNIFUSION, a framework for image generation and editing with unified multimodal encoder via a frozen VLM, comprising two key components: Layerwise Attention Pooling (LAP) for multi-layer feature aggregation and VLM-Enabled Rewriting Injection with Flexible Inference (VERIFI) for distribution alignment. (2) We conduct extensive ablations across prominent conditioning strategies, demonstrating that LAP outperforms conventional last-layer extraction and alternative fusion schemes while maintaining architectural flexibility. (3) We train and validate a single model that achieves competitive performance on both text-to-image generation and editing tasks using **only VLM input features**, eliminating the need for separate VAE-based image reference conditioning (See Fig. 12 and Fig. 13 for high-resolution text-to-image and editing samples). (4) We demonstrate **zero-shot generalization** capabilities, including multi-reference composition despite training only on single-reference data (See Fig. 10), and image-to-image variation despite training exclusively on text-conditional generation.

2 ARCHITECTURE SELECTION

In this section, we first formally introduce potential paradigms for VLM-conditioned image generation. We perform direct comparisons in which LAP outperforms all other methods, and demonstrate the preservation of fine-grained image details through LAP.

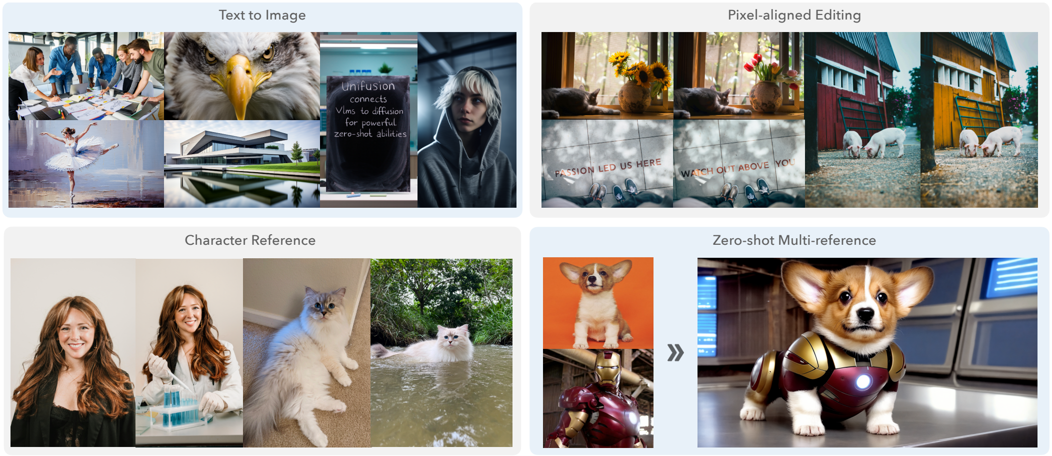

Text to Image
UniFusion connects Vlms to diffusion for powerful zero-shot abilities
Pixel-aligned Editing
PASSION LED US HERE
WATCH OUT ABOVE YOU
Character Reference
Zero-shot Multi-reference
»

Figure 3: UNIFUSION demonstrates great details in generation, high-fidelity editing with only VLM semantic features, and zero-shot capabilities enabled by the unified embedding space.

2.1 VLM Conditioning Candidates

We consider four different architectures for a VLM-conditioned unified encoder as depicted in Fig. 1 and 2. For all approaches, we extract features from a frozen VLM to condition a DiT.

Last-Layer Hidden State Encoding. An intuitive approach is to extract representations from the last hidden layer of the VLM as a drop-in replacement for text conditioning in existing architectures (Fig. 1a). Recently, multiple papers have similarly used the last hidden layer of a strong LLMs (Team, 2025; Xie et al., 2025; Chen et al., 2025a). Other variants of this approach have been proposed that use the penultimate layer instead of the last one (Qin et al., 2025).

Layerwise Key-Value Fusion. One of the first proposed methods utilizing information from multiple layers is layer-wise Key-Value Fusion (Fig. 1b). Liu et al. (2024) proposed to match the number of layers and hidden dimension of the image generator to the encoder. In each attention layer, we then concatenate the Keys and Values of DiT with the respective Keys and Values of the encoder.

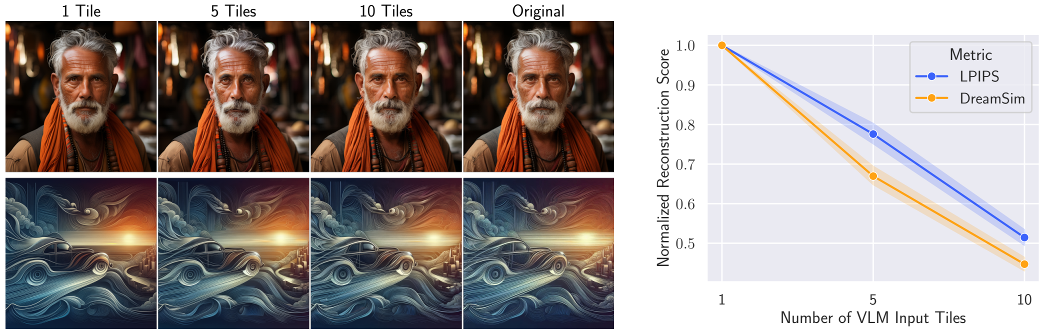
Hidden State Injection (HSI). We also consider an improvement over the previous approach that eliminates the need for Key-Value matching. Instead, we inject the representation from corresponding layers directly in the DiT through numerical addition of the residual stream after each block.

Layerwise Attention Pooling (LAP). We propose aggregating tokens from multiple layers using a learnable pooling module (Fig. 1). LAP consists of 2 transformer blocks that attend to the same token across layers, followed by a linear layer pooling the representations into one feature. Concretely speaking, we extract multiple VLM layers, then stack the tensor of shape ($\texttt{bs, sl, n, h}_E$) as X^E ($\texttt{bs}\star\texttt{sl, n, h}_E$) before proceeding to transformer blocks. Eventually the tensor unstacks and is fed into a linear layer to pool $\texttt{n}$ layers into one.

Benefits and Shortcomings of each Approach. The main limitation of Last-Layer encodings is that information is restricted to one layer. Multiple prior works have established that transformer layers at different depths encode varying levels of features (Lindsey et al., 2025; Durrani et al., 2020; 2023; Hennigen et al., 2020). We argue that intermediate layers also carry different levels of *semantic* abstraction and fine-grained details that are necessary for a unified multimodal encoder. While Layerwise Key-Value Fusion and Hidden State Injection extract features from multiple layers, they force tight coupling between the encoder VLM and the generative model, losing flexibility in the architectural design of the DiT.

2.2 Experimental Evaluation

Experimental Setup. We utilize a 5B-parameter, standard latent DiT architecture with full self-attention and 2x2 patchification. In line with previous work (Liu et al., 2024; Ma et al., 2024), we use frozen Llama3.1-8B (Grattafiori et al., 2024) for text-to-image tasks, and subsequently apply



(a) Qualitative Examples

(b) Quantitative Reconstruction Quality

Figure 5: Image reconstruction quality when the input image is patchified into 1, 5, and 10 tiles before being fed to the VLM. LAP-extracted VLM features are capable of preserving input image details without additional feature injection. Fine-grained images require more input tiles for to be captured accurately.

our findings to multimodal tasks using InternVL2.5-8B (Chen et al., 2025b). We train on a global batch size of 1024 for 200k steps, unless stated otherwise.

Text-to-Image Prompt Following. In Fig. 4, we provide GenAI Bench performance of all architectures. Overall, LAP stands out as the best option in terms of prompt adherence. The Llama 3.1 LAP version outperforms Last Hidden layer with an even larger improvement over Key-Value fusion. While LAP emerged as the best candidate, no Llama-3 conditioned checkpoint reaches the performance of the T5 baseline. These results align with independent observations (Ma et al., 2024). We observed that T5-conditioned models seem to converge faster than Llama ones. After 400k training steps the gap in VQA score between the T5 and Llama-3 LAP model closes to 0.7 percentage (from 4.1 percent at 200k).

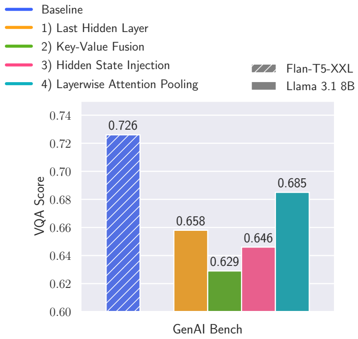


Figure 4: Different conditioning candidates after 200k steps on text-to-image task (measured by VQA Score (Lin et al., 2024) on GenAI-Bench (Li et al., 2024)).

Based on our Llama-3.1 ablations, we decided to move forward with LAP as the architecture for UniFusion. It outperforms HSI and last hidden layer approaches, while maintaining high flexibility with no inherent requirements on layer count or hidden dimension. We explore how to best utilize our LAP setting in Sec. 3, which ends up clearly outperforming T5 baselines on text-to-image generation and simultaneously supports further use cases.

Image Information Preservation. With the benefits of LAP over other architectures established on text-to-image tasks, we shift our focus to image inputs. A unified encoder should be able to preserve fine-grained visual details to obtain precise edits, but previous work reported that VLM-based features specifically fall short of that hurdle (Bellagente et al., 2023; Team, 2025) when conditioning only on a single layer. In addition to multi-layer conditioning, the representation capacity of the extracted features also plays an important role. The number of image tokens at a given hidden dimension is often significantly lower than that of comparable VAEs, for example. Naturally, in such a setting, adding VAE-encoded image input tokens improves the preservation of fine-grained details.

By comparing different number of image tiles used in VLM image encoding, we observed a scaling trend. In Fig. 5, the preservation of small features scales with the number of VLM tiles. At 10 tiles, any reconstruction errors become largely imperceptible. Even fine-grained structures, such as hairs or complex patterns, are preserved well. Thus, we conclude that VLM features are sufficient for image encoding, given a sufficient number of tiles or image tokens, and conditioning on early-layer features. See A.2 for more image reconstruction samples.

Representation Injection. We also considered learning dedicated LAPs for injection in different layers of the DiT. However, in a compute controlled comparison we found general input conditioning to strongly outperform any deep injection approaches. We provide further details in App. A.1.1.

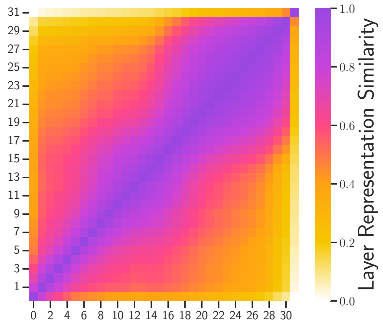


Figure 6: Adjacent VLM layers produce highly similar features. Consequently, extracting features from every layer may yield redundant information.

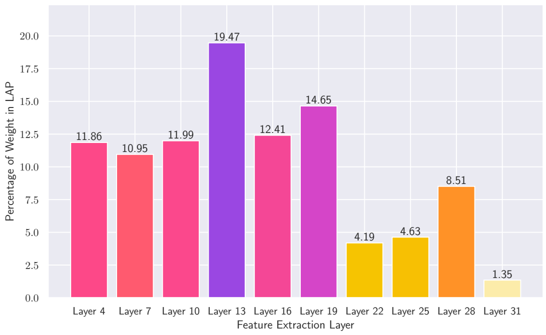


Figure 7: A pooling layer visualization within LAP using only every 3rd layer of the VLM encoder learns more uniform layer weights.

3 UniFusion Design

We have established Layerwise Attention Pooling (LAP) as the most promising conditioning strategy for a unified encoder in Sec. 2. In this section, we go into detailed design choices of UniFusion.

Layer Selection. While the previous results have shown clear benefits of aggregating representations from multiple layers, not all layers will be equally relevant. Utilizing all layers may cause high memory overhead and incentivizing DiT to overfit on a small subset of layers.

More concretely, we discovered that LAP often pools outputs from a contiguous set of layers to form the final representation of a token. When plotting the Query-Key norms for individual tokens (Fig. 18 in Appendix), we found highly clustered activation patterns. For example, the word "drinking" shows clusters for the 1st-4th or 21st-24th layer. Activations of adjacent layers in the transformer are highly similar, as shown in Fig. 6 and in other works (Krause et al., 2025; Lawson et al., 2025; Liu et al., 2023). We argue that while the image generator still benefits from extracted representations from all depths of a VLM encoder, considering every layer adds unnecessary redundancy and suboptimal parameter utilization.

Based on these insights, our final LAP architecture takes in every third layer of the input encoder. A model trained with this refined configuration exhibits more uniform weights as seen in Fig. 7. In the final UniFusion architecture, we perform a VLM layer dropout experiment as demonstrated in Fig. 8. We observe that image generation does not strongly rely on the first and last layers. When zeroing out respective weights during pooling, the overall image composition remains unchanged. In contrast, dropping information from middle layers results in massive deviations in the output.

Position Bias. During our analysis in Sec. 2.2, we identified cases where Llama-conditioned models randomly fail to accurately capture key subjects from the text prompt. This issue can be attributed to bias introduced by *causal attention* masking in the encoder transformer. Since a given token will only be attended to by the ones following it, information about a subject mentioned late in the prompt will be insufficiently represented. Similar to Ma et al. (2024), we combat this bias by adding a bidirectional refiner before LAP.

VLM-Enabled Rewriting Injection with Flexible Inference (Verifi). We propose VLM-Enabled Rewriting Injection with Flexible Inference (Verifi), a conditioning strategy which lets us effectively fold the VLM's world knowledge and reasoning into our unified representation space.

We formulate the training and inference of UniFusion so that the `decoded text` tokens are the only text tokens used to condition the DiT. These `decoded text` tokens form a long, detailed description of the intended image (the `target prompt`) and are obtained by asking the VLM—via a dedicated `system prompt`—to describe the scene in detail. Note that we also inject all image tokens (patch and thumbnail) in multimodal settings, for better visual content preservation.

Verifi offers several advantages. At inference time, we do not have to rely on an external VLM for prompt rewriting, a critical part of most modern text-to-image systems (Betker et al., 2023; Liu et al., 2024; Esser et al., 2024; Brack et al., 2025). Since rewriting happens in-model during the

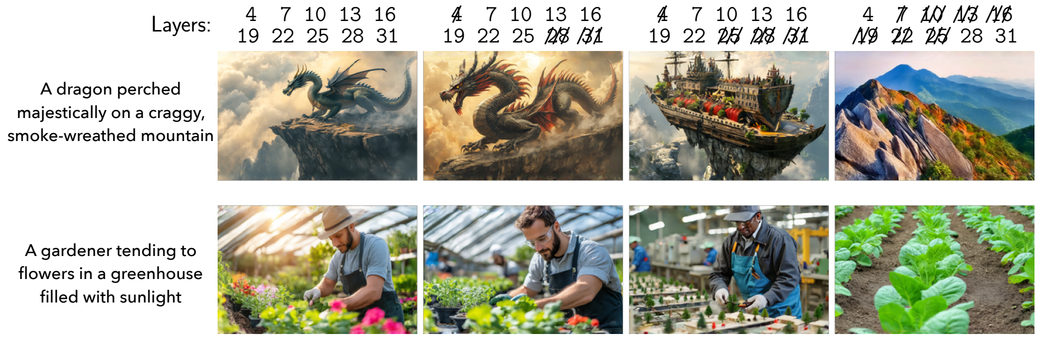


Figure 8: Analysis of layer selection's impact. We drop crossed-out layers in LAP aggregation. Middle layers are crucial for the overall scene composition, while the first and last only capture rudimentary scenes.

same forward pass used for feature extraction, we avoid the extra cost of re-encoding rewritten text. Additionally, the original user instruction is retained implicitly in the representations corresponding to `decoded text` tokens, since these tokens attend to the full input context inside the VLM. Finally, conditioning only on `decoded text` allows us to avoid injecting duplicate information to the DiT, mitigating position bias in causal attention models (Sec. 3).

We depict qualitative results and benchmark evaluations for self-rewriting in Fig. 19. We find that VERIFI significantly improves prompt following performance, reliably mitigating catastrophic failure cases that can occur otherwise. In the examples in Fig. 19, VERIFI correctly places the parrot on the buildings, adds the mouse to the generated image, and depicts ancient buildings. Further, we observed that VERIFI also improves performance on already long, detailed prompts. Moreover, VERIFI also enables zero-shot reasoning over complex inputs, which we explore more in Sec. 5.

4 FINAL UNIFUSION MODEL

Finally, we integrate all learnings from previous sections into a scaled-up model. We increase the DiT parameters to 8B and the training set to 830 million samples. The final UNIFUSION conditions on InternVL3-8B (Zhu et al., 2025), a more powerful successor than InternVL2.5 in earlier ablations.

UNIFUSION Design & Training. We design our final UNIFUSION model to extract features from every third layer of the VLM and aggregate them into a single representation via our LAP module. The LAP contains two transformer blocks aggregating the representation of any token across *layers*. This sequence is then pooled into one dense representation with a simple fully connected layer. The LAP is followed by a refiner of two bidirectional transformer blocks, mitigating position bias across the input *sequence*. We inject the extracted representation only in the DiT's input sequence, which operates on a VAE latent space with a compression factor of 16.

Note that we only encode input images through the VLM and do not concatenate any VAE tokens to the DiT input. UNIFUSION leverages self-rewrite of user inputs (i.e. VERIFI), with only image tiles and rewritten tokens being passed into the DiT. We train the model with up to 10 image tiles.

As described in Sec. 2, we train the base model on a mixture of text-to-image, image reconstruction, and joined text-and-image samples. Subsequently, we continue training with instruction data for image editing and reference workflows. We found roughly 10k steps of instruction training to be sufficient to support this task. For all stages of training, we use no web-scraped data and only rely on images with permissive licenses for generative image training. For this study, we do not perform any further post-training, which we leave for future work.

Qualitative Evaluation. We evaluate a single UNIFUSION model for both editing and generation workflows. Interestingly, we observed that training on editing task *improves* text-to-image generation capabilities (see Sec. 5). In Fig. 16 and Fig. 17, we observe that UNIFUSION exhibits competitive quality against models with much more parameters and training data, even without any post-training - both in terms of photorealism and aesthetics in text-to-image and editing fidelity.

Quantitative Evaluation. We evaluated UNIFUSION on DPG-Bench for text-to-image performance, while relying on outsourced human preference A/B test for Editing quality evaluation.

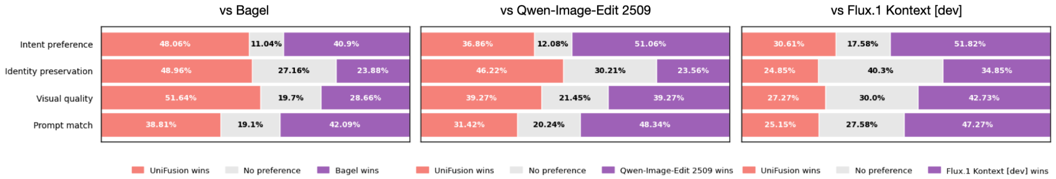


Figure 9: Human preference on Editing quality. UNIFUSION beats Bagel in all aspects, while compares favorably with Qwen-Image-Edit-2509 on visual quality and identity preservation.

Throughout building evaluation pipelines, we discovered several flaws in popular automated evaluation setups, such as GenEval and the original DPG-Bench, which significantly impacts the reliability of these metrics as is. More details in App. A.3.

In Tab. 1, we report UNIFUSION's performance in comparison to other models on a *refined* DPG-Bench, which is based on Gemma3-27B evaluator with CoT reasoning. We report both the best-out-of-4 score and the average. Despite no post-training and a limited training dataset, UNIFUSION remains competitive with significantly larger, heavily post-trained models.

When directly comparing the scores of Qwen-Image and UNIFUSION, we see a larger gap between the average and best-out-4 scores for UNIFUSION. The difference in Macro Avg. scores is 0.142 and 0.184 for Qwen-Image and UNIFUSION, respectively. We believe this to be a direct result of the post-training for Qwen-Image.

How does UNIFUSION perform on editing when conditioned only on VLM-encoded image inputs? We conducted a human A/B study with 200 annotators over a diverse set of edit types (see details in A.4). In Fig.9, we observe that UNIFUSION is preferred over BAGEL in aggregate preference, identity preservation, and visual quality. Compared to Qwen-Image-Edit 2509, UNIFUSION improves identity preservation and is roughly tied on visual quality. UNIFUSION also remains competitive with Flux.1 Kontext [dev], despite using an 8B DiT and a smaller training set.

We note that part of the remaining gap can be attributed to missing or underrepresented training data types (e.g., portrait retouching), which disproportionately affects prompt-match in those categories; we nevertheless retain these cases for a fair evaluation.

5 EMERGENT CAPABILITIES

We observed UNIFUSION to exhibit many valuable zero-shot capabilities without explicitly being trained for them. This behavior is a direct benefit of a unified VLM encoder architecture. Any of the capabilities learned from the VLM's extensive training regime are retained and transferred to image generation tasks. Additionally, the unified space of contextualized text and image eliminates large distribution shifts between tasks.

Reasoning & Complex Prompts. VERIFI allows the models to explicitly leverage the world knowledge and reasoning capabilities of the VLM encoder. For example, we can decompose hypothetical scenarios to perform image editing requiring multi-hop reasoning and world knowledge. We showcase some examples in Fig. 25. The VLM correctly reasons about hypothetical effects of the mass of different animals or impacts of temperature changes over time and decomposes those into an edit instruction. The DiT is then able to perform an edit satisfying the original user intent. We observe similar capability in decomposing convoluted instructions and logical puzzles. Details in Fig. 25 and Fig. 26.

Generalization to Unseen Modalities. Throughout all experiments, we observed models to generalize well to inputs that were never observed during DiT training. For example, a model solely trained on text-to-image generation can still capture the semantics of image inputs as seen in Fig. 15a. While the reconstruction is not pixel-perfect, the generated image still accurately captures almost all important aspects of the input image. This behavior can be attributed VLMs having aligned representation spaces between text and image, enabling zero-shot transfer to new modalities.

Similarly, we find that models only trained on text *or* image sequences, but never simultaneously (i.e. editing pairs), can still be used for image editing. Fig. 15b shows some zero-shot editing samples

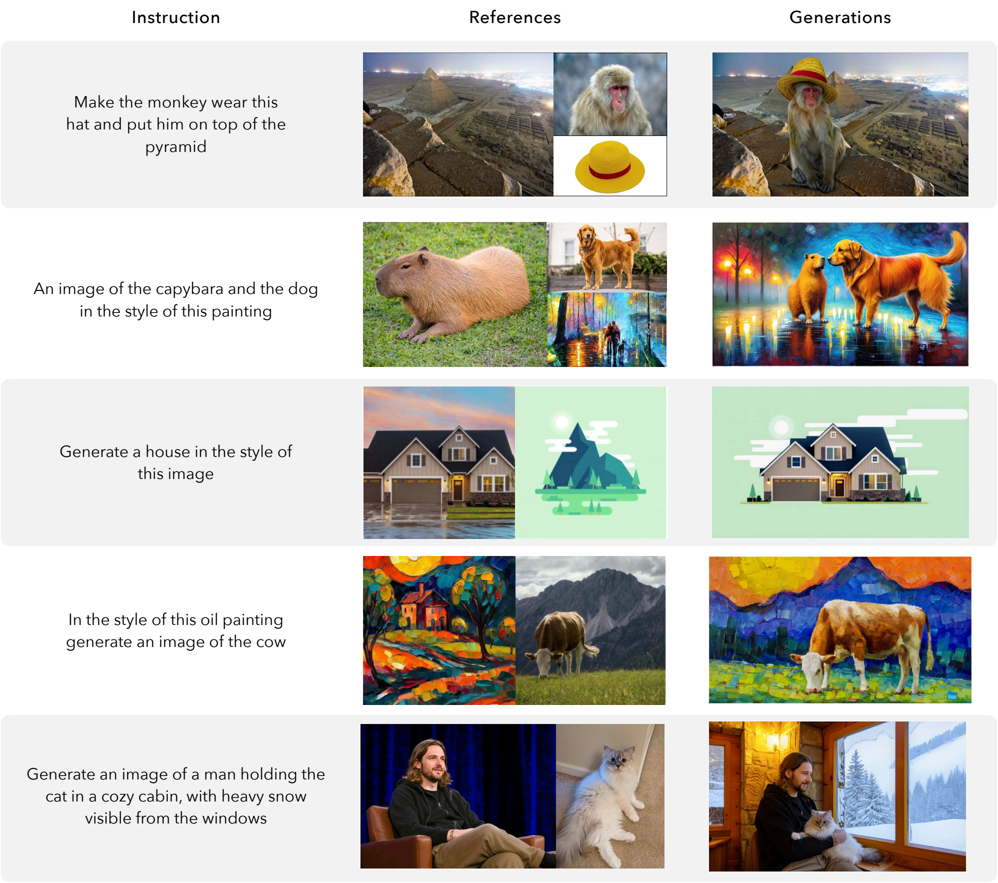

Instruction
References
Generations
Make the monkey wear this hat and put him on top of the pyramid
An image of the capybara and the dog in the style of this painting
Generate a house in the style of this image
In the style of this oil painting generate an image of the cow
Generate an image of a man holding the cat in a cozy cabin, with heavy snow visible from the windows

Figure 10: Zero-shot multi-reference generations from UNIFUSION when only trained on single-ref samples.

Category	Bagel		Flux.1 [dev]		Qwen-Image		UNIFUSION (Ours)	
	Avg.	Top-4	Avg.	Top-4	Avg.	Top-4	Avg.	Top-4
Macro Avg	.715	.901	.693	.899	**.802**	**.943**	.731	.915
Micro Avg	.786	.873	.753	.851	**.841**	**.914**	.787	.880
Entity	.795	.947	.761	.933	**.848**	**.968**	.791	.957
Attribute	.724	.910	.689	.901	**.787**	**.938**	.737	.926
Relation	.643	.875	.613	.854	**.740**	**.929**	.678	.887
Global	.639	.864	.641	.897	**.714**	.888	.688	**.898**
Other	.685	.852	.710	.889	**.875**	**.960**	.703	.863
Model Size	14B MoT		12B		20B		8B	

Table 1: UNIFUSION achieves competitive performance against much larger models trained on more data. Scores on modified DPG-Bench. We report average and best generation across four seeds at 1024px resolution. Macro Average is taken as the mean over scores per subcategory, whereas Micro averages scores across all prompts. **Bold** and underlined denote best and second-best score, respectively.

where an image is modified by the respective textual scene descriptions. Specifically, when we condition the DiT with image tokens for the first 10-20% of steps, and text tokens for the remaining 80-90%, the output image preserves most of the input image's content and semantics, even though the model was *never* trained with any textual image editing data.

Cross-Task Knowledge Transfer. In Section 4, we observed that continued training on image editing tasks improves the model's text-to-image quality. In Tab. 2, we see a strong improvement on DPG-Bench of over 2 percentage points in Micro Avg. We further conducted a human user study comparing checkpoints before and after brief finetuning on editing data, as shown in Fig. 11. Annotators strongly prefer images generated by the UNIFUSION-Edit checkpoint across all aspects.

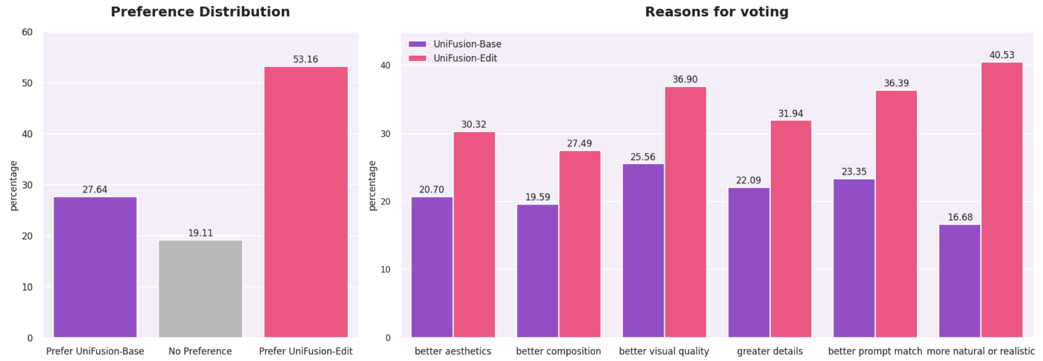


Figure 11: UNIFUSION-Edit leads UNIFUSION-Base by a significant margin in text-to-image A/B test with 180 annotators, 616 prompts across diverse concepts, with 2 seeds each (3 votes per image pair).

We hypothesize that this behavior mostly results from the unified encoder space for text and image. Since the representation space always supported multimodal inputs, the shift of model weights from text-to-image towards editing tasks becomes less disruptive. Instead, it increases concept coverage and refines the model's representations from multiple modalities' angles. Since UNIFUSION eliminates the need of VAE-encoded reference inputs, the DiT does not need to adjust its embedding behavior.

Category	UNIFUSION-Base		UNIFUSION-Edit	
	Avg.	Top-4	Avg.	Top-4
Macro Avg	.699	.906	**.731**	**.915**
Micro Avg	.760	.863	**.787**	**.880**
Entity	.758	.950	**.791**	**.957**
Attribute	.711	.923	**.737**	**.926**
Relation	.636	.885	**.678**	**.887**
Global	.670	.893	**.688**	**.898**
Other	.656	.825	**.703**	**.863**

Table 2: Editing training massively improves UNIFUSION capabilities in text-to-image performance (DPG-Bench).

Zero-shot Multi-reference Capabilities. We observed strong zero-shot abilities for image reference tasks. The editing data in Sec. 4, contained only examples with a single reference image. Moreover, all training samples fix input and output images to the same aspect ratio. Still, UNIFUSION is capable of composing scenes from multiple reference images of different aspect ratios, and applying unprompted shifts in perspective when needed, as seen in Fig. 10.

6 CONCLUSION & LIMITATIONS

We introduced UNIFUSION, a flexible framework that uses a single VLM as Unified Multimodal Encoder for generative image models. We proposed a novel Layerwise Attention Pooling (LAP) module which aggregates features from multiple VLM layers, enabling high-fidelity editing and image reconstruction solely with semantic features. With the text-image aligned representations within VLMs, UNIFUSION gains significant emergent capabilities such as zero-shot multi-reference generation and cross-task capability transfer. In spirit of aligning input representations, we also proposed VERIFI to further reduce distribution shift between generation and editing tasks.

Lastly, to ensure that UNIFUSION generalises beyond one VLM family (InternVL in our case), we trained an additional model based on Gemma (see App. A.1.4) and observed similar success.

Limitations. We identified some issues related to rendering text in scenes, which also impact the respective scores in Tab. 1. In general, UNIFUSION is capable of generating and editing typography. However, we found InternVL to be subpar at spelling, which sometimes leads the DiT often misspells text in the final image, given erroneous spellings in the decoded text during VERIFI. We also noticed that while VLM image semantic features are sufficient to reconstruct images well (see A.2), leading to performant editing results rivaling Bagel and Qwen-Image-Edit, they still suffer at extremely fine details when compared to the best VAE recipe.

REFERENCES

Marco Bellagente, Manuel Brack, Hannah Teufel, Felix Friedrich, Björn Deiseroth, Constantin Eichenberg, Andrew Dai, Robert Baldock, Souradeep Nanda, Koen Oostermeijer, Andrés Felipe Cruz-Salinas, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. Multifusion: Fusing pre-trained models for multi-lingual, multimodal image generation. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ba8d1b46292c5e82cbfb3b3dc3b968af-Abstract-Conference.html.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, and OpenAI. Improving image generation with better captions, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Manuel Brack, Sudeep Katakol, Felix Friedrich, Patrick Schramowski, Hareesh Ravi, Kristian Kersting, and Ajinkya Kale. How to train your text-to-image model: Evaluating design choices for synthetic training captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. URL https://arxiv.org/abs/2506.16679.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025a. URL https://arxiv.org/abs/2502.20172.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://doi.org/10.1109/CVPR52688.2022.00135.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 25, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. doi: 10.48550/arXiv.2505.14683. URL https://arxiv.org/abs/2505.14683. Introduces the unified multimodal model BAGEL; v3 (2025-07-27).

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.395.

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. Discovering salient neurons in deep NLP models. *Journal of Machine Learning Research (JMLR)*, 24, 2023. URL http://jmlr.org/papers/v24/23-0074.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.15.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Felix Krause, Timy Phan, Vincent Tao Hu, and Björn Ommer. TREAD: token routing for efficient architecture-agnostic diffusion training. *CoRR*, abs/2501.04765, 2025. doi: 10.48550/ARXIV.2501.04765. URL https://doi.org/10.48550/arXiv.2501.04765.

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer saes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=XAjfjizaKs.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. URL https://arxiv.org/abs/2406.13743.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models, 2024. URL https://arxiv.org/abs/2409.10695.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. Deja vu: Contextual sparsity for efficient llms at inference time. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/liu23am.html.

Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=7b2DrIBGZz.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu Qiao, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. URL https://arxiv.org/abs/2503.21758.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. doi: 10.48550/arXiv.2509.20427. URL https://arxiv.org/abs/2509.20427. Technical report; v2 (2025-09-28).

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qwen-image technical report. *arXiv preprint*, 2025. URL https://qianwen-res.oss-cn-beijing.aliyuncs.com/Qwen-Image/Qwen_Image.pdf.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025. doi: 10.48550/arXiv.2506.05083. URL https://arxiv.org/abs/2506.05083.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.05977.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Appendix

A.1 Additional Results & Experimental Details

In this section, we provide additional experimental details and results.

A.1.1 Representation Injection

In Fig. 21 we compare different injection paradigms for LAP. In the first setting, we train a dedicated LAP for each DiT layer and inject the respective representation through hidden state injection. In the second setting, we only extract one LAP representation and input it to the DiT without injections in later layers.

In this direct comparison, the latter setting strongly outperforms the former. These results suggest that injecting conditioning into later layers of the DiT may be counterproductive. Considering the analysis in Fig. 20, there are two-factors contributing to this difference. For one, we found that injecting into later layers of the DiT can adversely affect the usefulness of representations in the residual stream. Additionally, this effect is compounded by taking compute and capacity away from more useful early layer injection.

A.1.2 Bidirectional Refiner

In Fig. 22, we measure the benefit of a bi-directional refiner. We compare a T5 baseline against two InternVL-2.5 8B models. The first uses a bi-directional refiner on penultimate layer features, and the second combines an LAP with a bi-directional refiner. Comparing these results to Fig. 4, we observe that the combination of InternVL instead of Llama and the addition of a bi-directional refiner now closes the gap to the T5 baseline on text-to-image capabilities. Nonetheless, layer-wise attention pooling still outperforms representation extraction from last layers. Consequently, both multi-layer feature extraction and bi-directional refinement are crucial when using decoder-only auto-regressive models for input encoding.

A.1.3 Continued Training vs Finetuning

In order to assess whether UniFusion encoding requires training from scratch or could benefit from continued training from a T5 model, we make a direct comparison.

With a total compute budget of 250k steps, we train two different models. One that was trained for 100k steps using T5 and changes to InternVL-2.5-8B for the remaining 150k steps. The second one is trained using InternVL-2.5-8B from scratch. As shown in Fig. 23, both models converge to the exact same performance and substantially outperform the T5 baseline.

Based on these results, we can draw two conclusions. First, given an existing T5-conditioned model, we can save compute by continuing late

A.1.4 Gemma-Based UniFusion

In addition to the InternVL-based models in the main body, we also trained a UniFusion version based on Gemma-3-12B-it Team et al. (2025). With Verifi the model achieves a strong VQA score of 84.4% on GenAI Bench Li et al. (2024). We provide qualitative examples in Fig. 24 for text-to-image generation and image reconstruction.

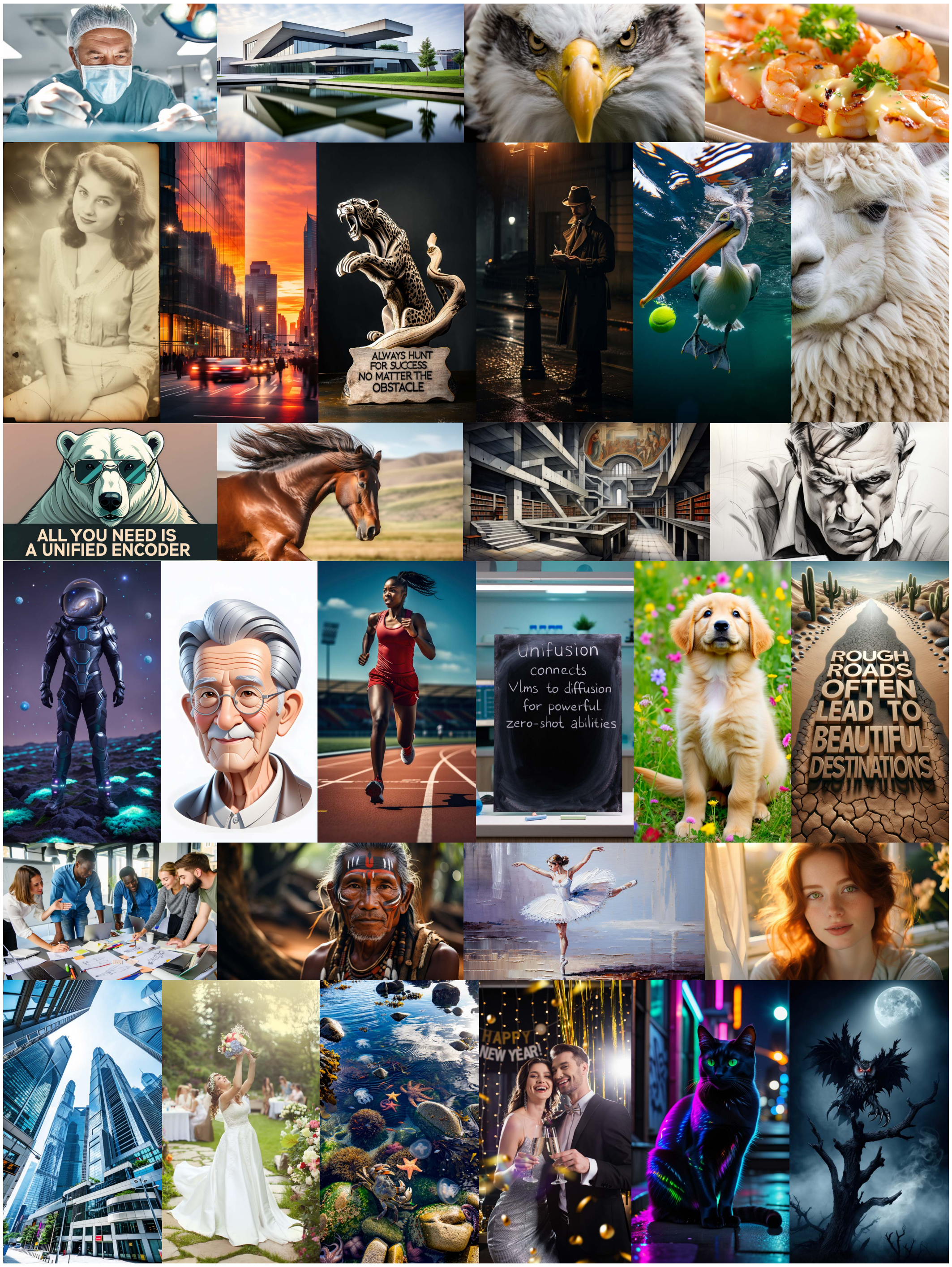

ALWAYS HUNT FOR SUCCESS NO MATTER THE OBSTACLE
ALL YOU NEED IS A UNIFIED ENCODER
Unifusion connects Vlms to diffusion for powerful zero-shot abilities
ROUGH ROADS OFTEN LEAD TO BEAUTIFUL DESTINATIONS
HAPPY NEW YEAR!

Figure 12: Diverse text-to-image generation with UNIFUSION(Zoom in for more details).

For image reconstruction using one tile (i.e., thumbnail) as input to the VLM, we observe slight variations. Based on our insights in Sec 2.2, we expect these artifacts can be resolved by increasing the tiling in the VLM inputs. Additionally, Gemma has a higher compression ratio for InternVL when using the same number of tiles. These results provide evidence that our UNIFUSION approach works reliably across different models and architectures.

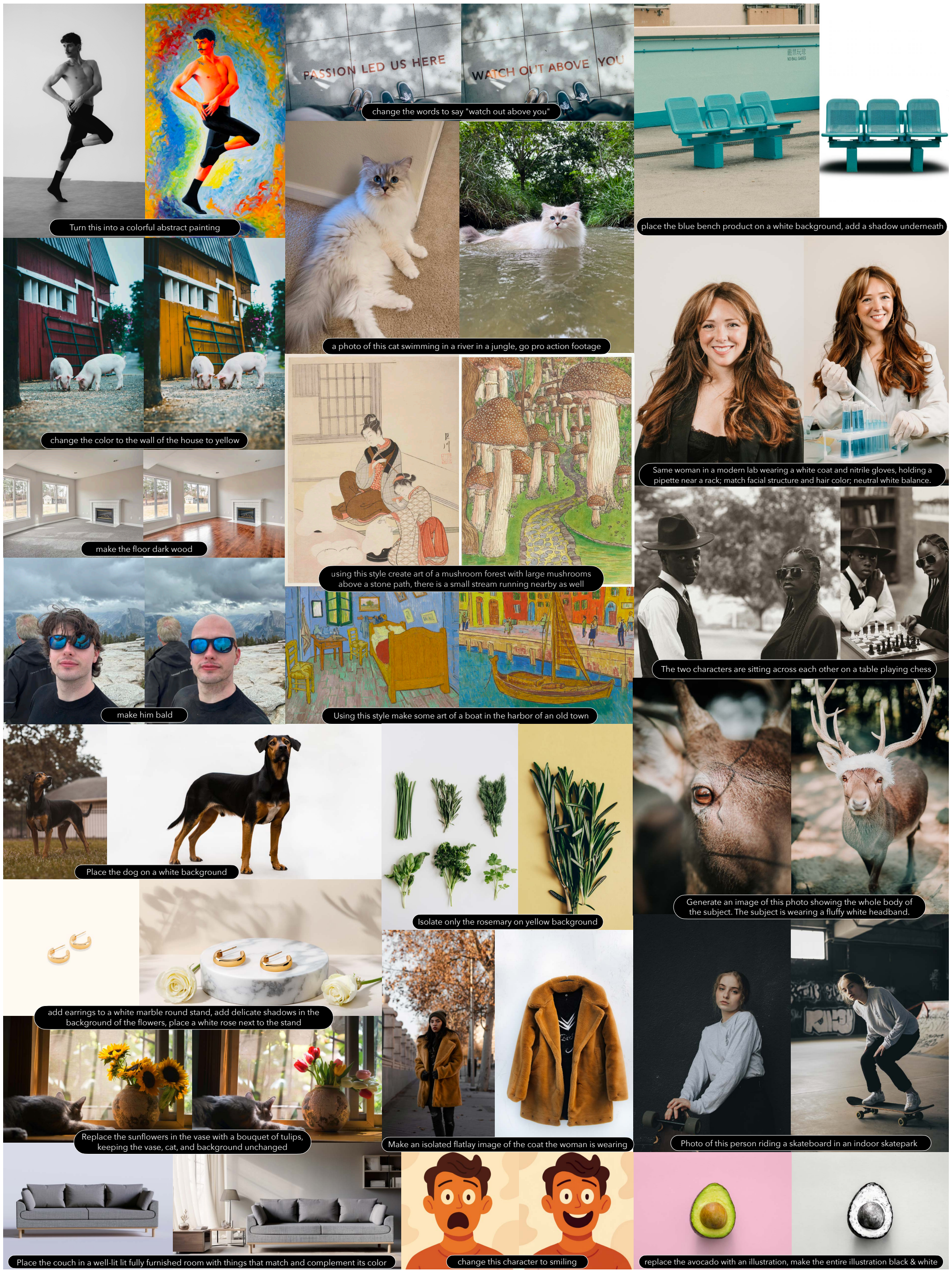

Turn this into a colorful abstract painting
PASSION LED US HERE
WATCH OUT ABOVE YOU
change the words to say "watch out above you"
place the blue bench product on a white background, add a shadow underneath
a photo of this cat swimming in a river in a jungle, go pro action footage
change the color to the wall of the house to yellow
Same woman in a modern lab wearing a white coat and nitrile gloves, holding a pipette near a rack; match facial structure and hair color; neutral white balance.
make the floor dark wood
using this style create art of a mushroom forest with large mushrooms above a stone path, there is a small stream running nearby as well
The two characters are sitting across each other on a table playing chess
make him bald
Using this style make some art of a boat in the harbor of an old town
Place the dog on a white background
Isolate only the rosemary on yellow background
Generate an image of this photo showing the whole body of the subject. The subject is wearing a fluffy white headband.
add earrings to a white marble round stand, add delicate shadows in the background of the flowers, place a white rose next to the stand
Replace the sunflowers in the vase with a bouquet of tulips, keeping the vase, cat, and background unchanged
Make an isolated flatlay image of the coat the woman is wearing
Photo of this person riding a skateboard in an indoor skatepark
Place the couch in a well-lit lit fully furnished room with things that match and complement its color
change this character to smiling
replace the avocado with an illustration, make the entire illustration black & white

Figure 13: Diverse textual image editing and image reference workflows with UNIFUSION. All images encoded by VLM features only, no VAE tokens involved. (Zoom in for more details).

A.2 SAMPLES OF IMAGE RECONSTRUCTION VIA VLM SEMANTIC FEATURES

Fig.14 shows 1024px image reconstruction from only VLM image features. With multi-layer conditioning, UNIFUSION is able to reconstruct images to fine details with just semantic features.

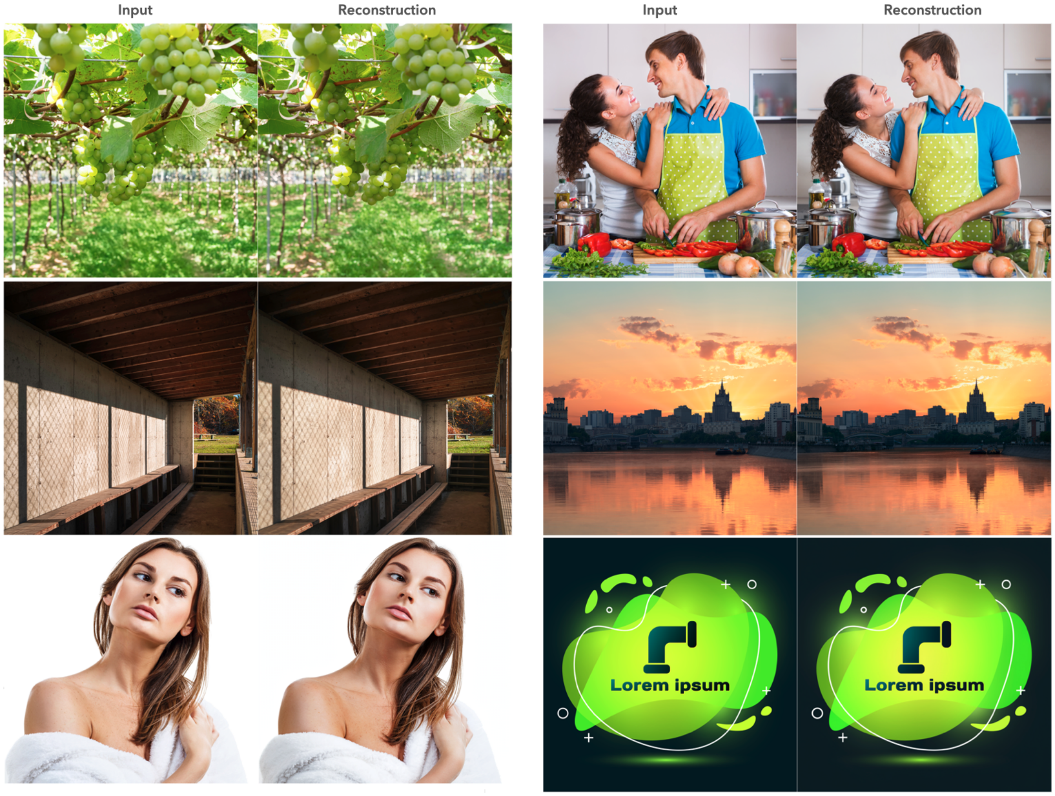
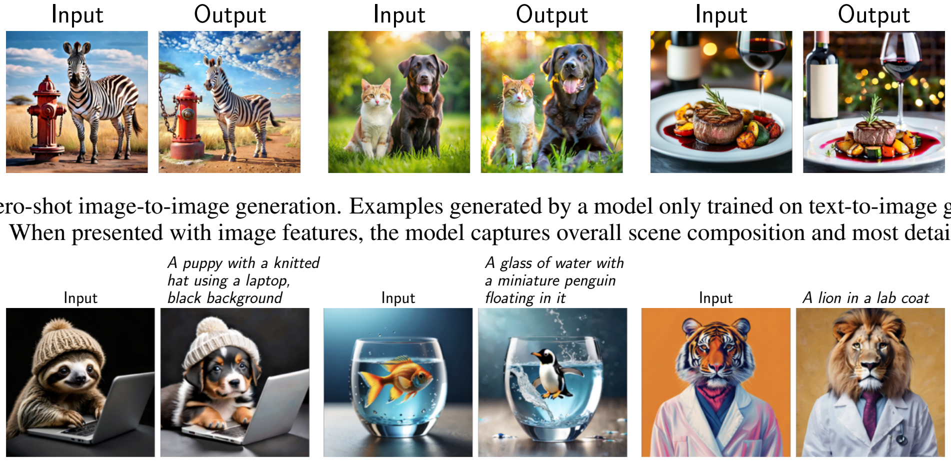


Figure 14: Image reconstruction with 10 image tiles on 1024px input image *solely with VLM image features*.


Input
Output
Input
Output
Input
Output
Input
A puppy with a knitted hat using a laptop, black background
Input
A glass of water with a miniature penguin floating in it
Input
A lion in a lab coat

(a) Zero-shot image-to-image generation. Examples generated by a model only trained on text-to-image generation. When presented with image features, the model captures overall scene composition and most details.

(b) Zero-shot image editing. Examples generated by a model never trained on image editing.

Figure 15: Conditioning image generation on LAP extracted VLM features enables zero-shot generalisation to unseen tasks and modalities.

A.3 On the Reliability of Image Generation Benchmarks

A.3.1 Evaluating Popular Benchmarks

In general, the limited reliability of these benchmarks and respective metrics can be broken down into three categories.

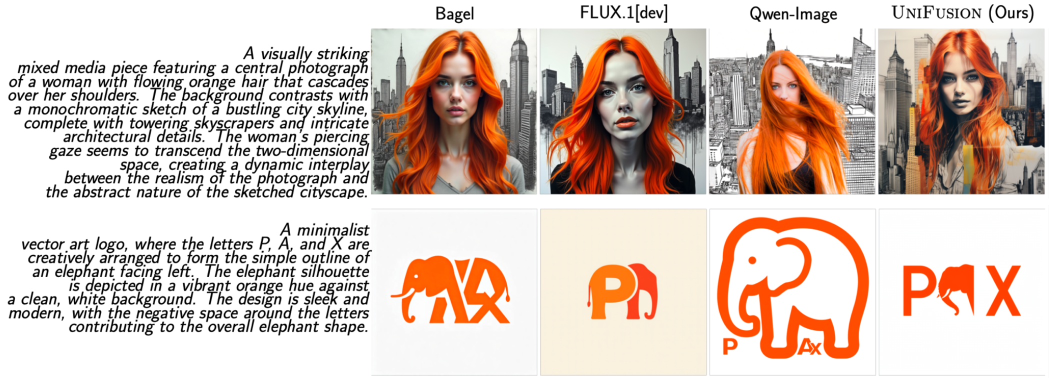

Bagel
FLUX.1[dev]
Qwen-Image
UNIFUSION (Ours)
A visually striking mixed media piece featuring a central photograph of a woman with flowing orange hair that cascades over her shoulders. The background contrasts with a monochromatic sketch of a bustling city skyline, complete with towering skyscrapers and intricate architectural details. The woman's piercing gaze seems to transcend the two-dimensional space, creating a dynamic interplay between the realism of the photograph and the abstract nature of the sketched cityscape.
A minimalist vector art logo, where the letters P, A, and X are creatively arranged to form the simple outline of an elephant facing left. The elephant silhouette is depicted in a vibrant orange hue against a clean, white background. The design is sleek and modern, with the negative space around the letters contributing to the overall elephant shape.

Figure 16: Qualitative comparison on long form text-to-image generation prompts comparing UNIFUSION to Bagel Deng et al. (2025), Flux.1 [dev] Labs (2024) and Qwen-Image Team (2025).

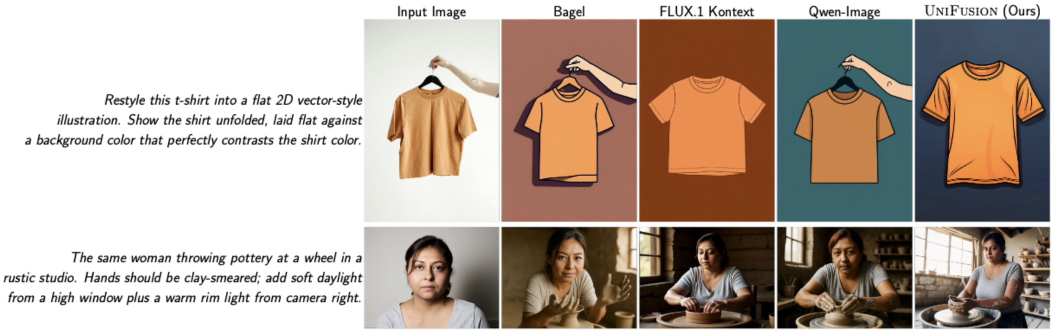

Input Image
Bagel
FLUX.1 Kontext
Qwen-Image
UNIFUSION (Ours)
Restyle this t-shirt into a flat 2D vector-style illustration. Show the shirt unfolded, laid flat against a background color that perfectly contrasts the shirt color.
The same woman throwing pottery at a wheel in a rustic studio. Hands should be clay-smeared; add soft daylight from a high window plus a warm rim light from camera right.

Figure 17: Qualitative comparison on editing tasks comparing UNIFUSION to Bagel Deng et al. (2025), Flux.1 Kontext Labs et al. (2025) and Qwen-Image Team (2025) (zoom in for details.)

Automated evaluation error. The majority of benchmarks rely on separate models to evaluate generated images. We observed the error rate of these models to far exceed reasonable metrics. For example, GenEval Ghosh et al. (2023) relies on a pre-trained object detection model Chen et al. (2019); Cheng et al. (2022) and CLIP Radford et al. (2021) for attribute matching. For a benchmark to remain useful, everybody should follow a pre-determined setting, making scores comparable. Unfortunately, these models become outdated quickly and have high failure rates for the designated tasks. We show examples of incorrect GenEval assessments in Fig. 27a where either the initial object detection, object count, or attribute binding fails. For some model evaluations, we found incorrectly flagged generation fails of this setup to exceed 70%. Given that current models tend to achieve good performance on these benchmarks, the error of the metric itself tends to exceed the difference between the compared models. Thus, discerning any perceived improvements from measurement noise becomes impossible.

We found question-answering-based settings like the one proposed by DPG-Bench Hu et al. (2024) to suffer from similar issues. We depict some examples in Fig. 27b. Specifically, the VLM proposed by DPG-Bench hallucinates incorrect answers at an alarming rate. As shown, these failures even occur for well-composed images, with no major artifacts and the subject in question clearly visible in the image. While some of these problems can be attenuated by using more capable models and a more comprehensive evaluation setting (See App. A.3.2), the underlying problems remain.

Ill-formulated tasks. Since comprehensive benchmarks are time-consuming to build, they often rely on LLMs to construct instructions or evaluation targets. However, this has led to an increasing number of evaluation objects that are impossible to evaluate objectively.

For example, DPG-Bench contains a multitude of questions that are subjective to some extent, cannot be grounded in a single image, or are otherwise questionable. We provide some examples in Tab. 3. In general, there is a large number of questions attempting to ascertain the nationality of

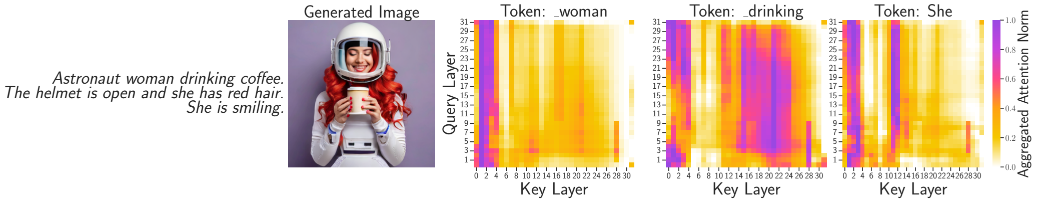
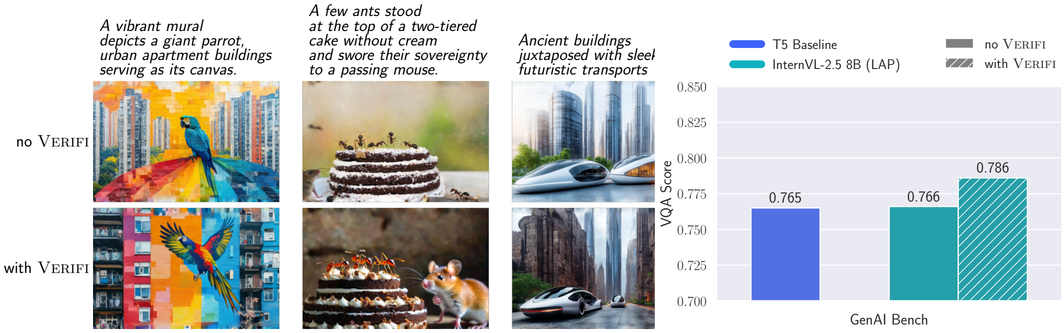


Figure 18: Qualitative example of local clusters in LAP Key-Value norm. On many tokens, the model utilizes implicit clusters of adjacent layers. Values are averaged over tokens if a word has more than 1 token.



(a) Qualitative Examples

(b) GenAI Bench Evaluation

Figure 19: VERIFI improves prompt following. Comparison of InternVL-2.5-8B conditioned DiT with and without VERIFI. Especially, complex prompts involving multiple subjects are generated more accurately.

people, which is impossible to assess without context. Further, since a lot of the underlying text prompts are heavily embellished with subjective adjectives. Given the collection methodology of DPG-Bench, this likely stems from synthetically written prompts. Crucially, the GPT-written questions often pick up on these adjectives. However, assessing if a painting does 'radiate' or if a squirrel is 'rebellious' is highly subjective and should not be central to an objective benchmark. Lastly, some questions like the historical significance of a photograph are next to impossible to assess from an image alone, without providing further context.

Questionable capability prioritization. Naturally, generative image tasks have to satisfy multiple–often orthogonal—constraints. However, we found that current benchmarks and metrics

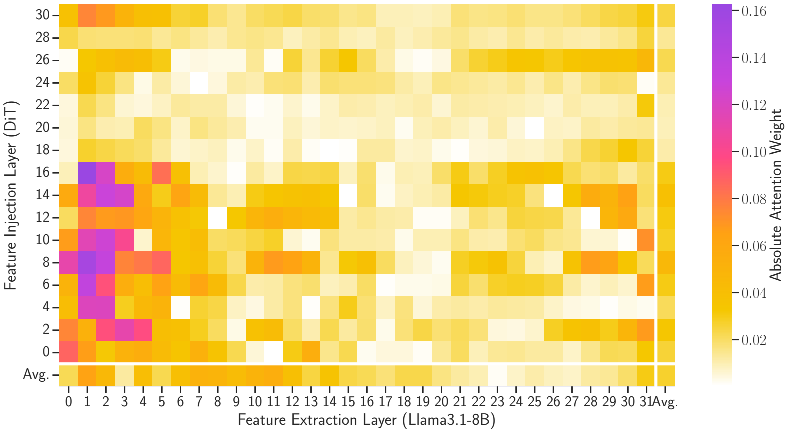
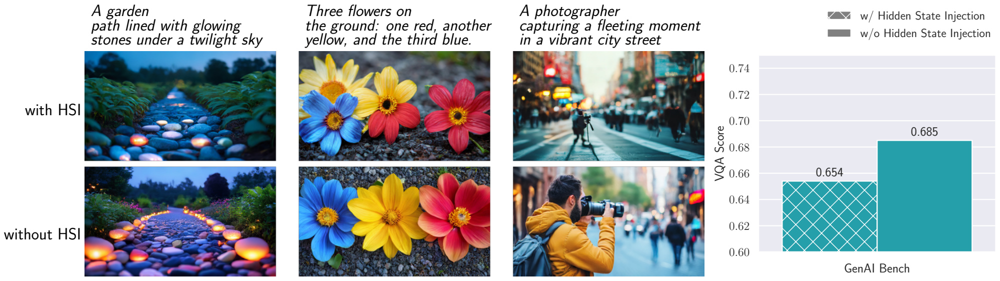


Figure 20: Weight visualization of LAP modules' pooling layers in Representation Injection setup (Sec 2.2). Each value denotes the magnitude of weights assigned to each VLM layer at a given LAP module's pooling layer (smaller y-coordinate denotes layers closer to DiT input). On average, early VLM layers contribute more than later ones, while layer injection at later DiT blocks has lower weights.



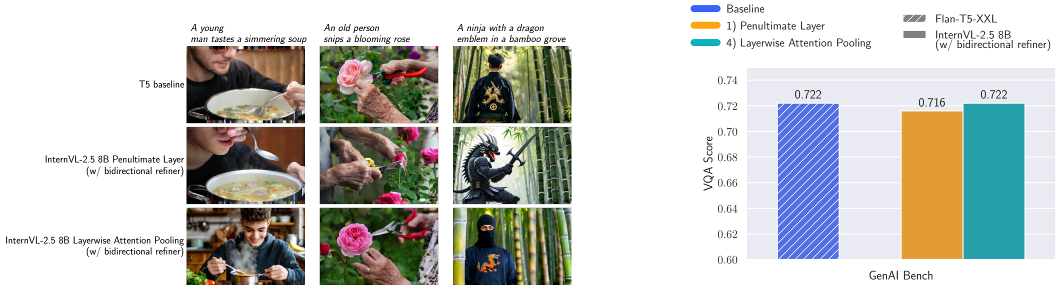
(a) Qualitative Examples (b) GenAI Bench Evaluation

Figure 21: Injecting aggregated representations at different DiT depth does not improve performance. Comparison of InternVL-2.5-8B with LAP feature extraction. First version injects dedicated representations per DiT layer (with HSI), second version pools one representation for DiT conditioning (without HSI). Comparison at 200k training steps.

Prompt-ID	Question
diffusiondb3	Is there an Indian woman?
diffusiondb3	Is there a Chinese man?
partiprompts162	Is the car tableau dreamlike?
midjourney21	Is the sculpture majestic?
partiprompts122	Does the scene feel expansive?
partiprompts159	Are the hues uplifting?
partiprompts83	Is the cup lovestruck?
partiprompts126	Does the squirrel have a rebellious punk rock vibe?
71	Are the printers humming with activity?
COCOval2014000000513096	Is the man in the suit explaining the significance of the exhibit?
posescript2	Does the individual exhibit bodily awareness?
countbench16	Are the plates likely originating from London in the year 1752?
countbench17	Do the photographs have historical significance?

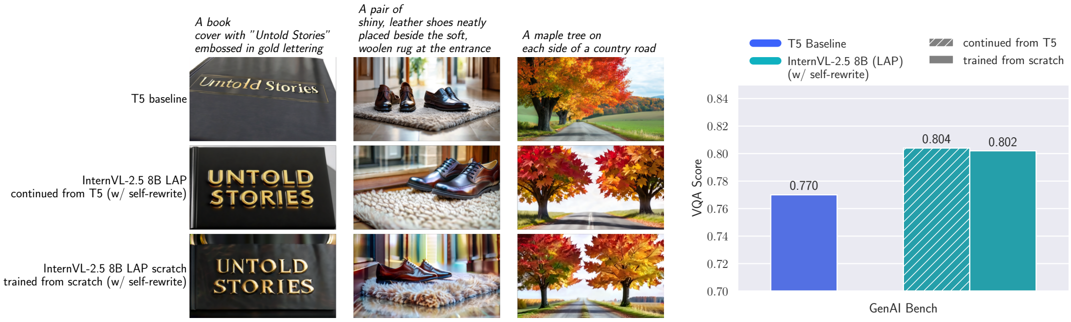
Table 3: Examples from questions in DPG-Bench that are hard to assess objectively from generated images.

tend to heavily prioritize very literal prompt adherence. Take, for example, the image of the toilet and mouse in Fig. 27a. One could argue that this scene composition satisfies some aesthetic aspects by placing the toilet only partially visible in the background. In general, we found all evaluation settings to judge incomplete objects or out-of-focus backgrounds as violating prompt adherence. However, both might be intended behavior for aesthetic quality and composition, as well as accurate depth of field for photographic image styles. Even human-preference metrics like ImageReward Xu



(a) Qualitative Examples (b) GenAI Bench Evaluation

Figure 22: Evaluation of bi-directional refiner impact. InternVL2.5-8B model with refiner closes the performance gap to the T5 baseline (cf. Fig. 4. Nonetheless, layer-wise attention pooling still outperforms representation extraction from the last layers. Comparison at 250k training steps.



(a) Qualitative Examples

(b) GenAI Bench Evaluation

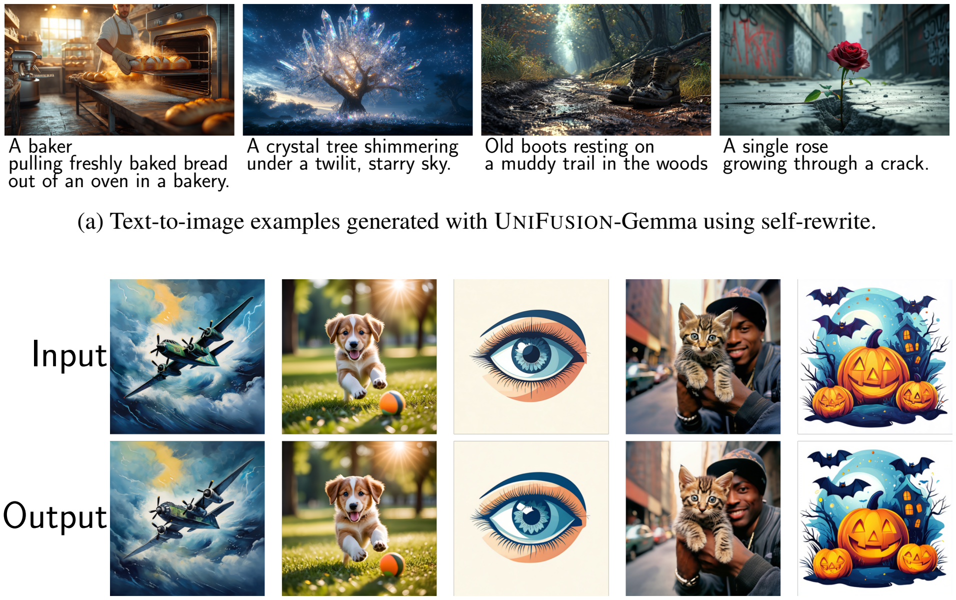
Figure 23: Evaluation of training UNIFUSION conditioning from scratch vs. continuing from T5. Both approaches produce models with the same capabilities. Comparison at 250k training steps. Continued checkpoint switches from T5 to InternVL2.5-8B at 100k steps.

et al. (2023), tend to prioritize very literal prompt following over other aspects. However, from in-house user studies, we found that this implicit waiting for strict prompt adherence over other quality aspects does not necessarily correlate with actual human preference.

A.3.2 REFINED DPG-BENCH

For our analysis in Sec. 4, we made the following adjustments to DPG-Bench.

Upgrade Question-Answering Model. We changed the VLM used for question answering to Gemma-3-27b-it Team et al. (2025). We specifically chose the strongest model from the Gemma family, since we were evaluating models conditioned on InternVL and QwenVL models. Consequently, to remove unintended evaluation bias, we opted for the strongest open-weight VLM outside of these model families.


A baker pulling freshly baked bread out of an oven in a bakery.
A crystal tree shimmering under a twilit, starry sky.
Old boots resting on a muddy trail in the woods
A single rose growing through a crack.
Input
Output

(a) Text-to-image examples generated with UNIFUSION-Gemma using self-rewrite.

(b) Image reconstruction with UNIFUSION-Gemma at 1 input tile. Similar to the experiments in Sec. 2.2, we observe slight variations when using only one tile. We expect these artifacts to resolve themselves at increased input resolution.

Figure 24: Text-to-image and image reconstruction examples of the UNIFUSION-Gemma model.

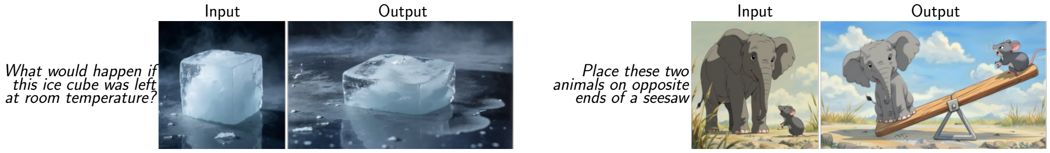
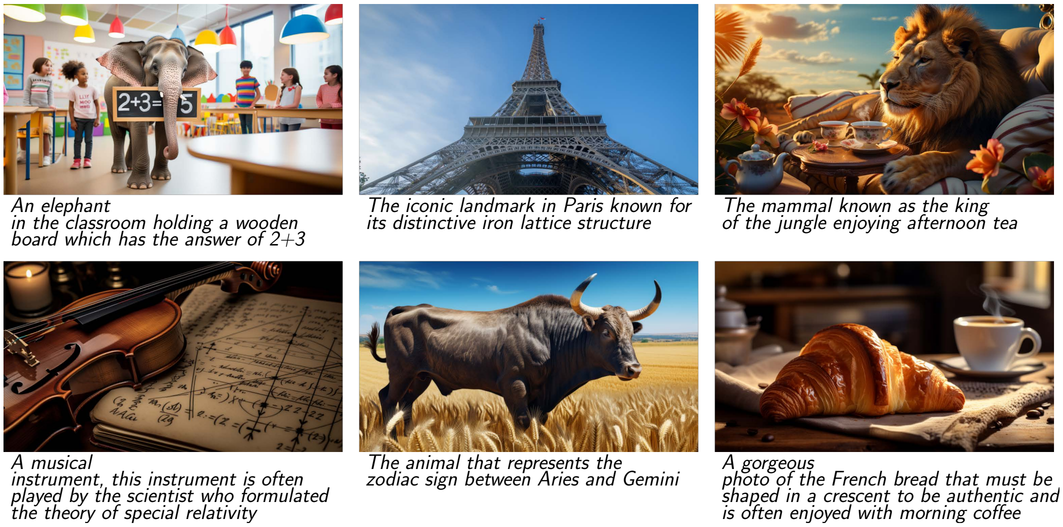

Input
Output
What would happen if this ice cube was left at room temperature?
Input
Output
Place these two animals on opposite ends of a seesaw

Figure 25: The unified VLM encoder enables advanced visual reasoning for textual image editing.


An elephant in the classroom holding a wooden board which has the answer of 2+3

The iconic landmark in Paris known for its distinctive iron lattice structure

The mammal known as the king of the jungle enjoying afternoon tea

A musical instrument, this instrument is often played by the scientist who formulated the theory of special relativity

The animal that represents the zodiac sign between Aries and Gemini

A gorgeous photo of the French bread that must be shaped in a crescent to be authentic and is often enjoyed with morning coffee

Figure 26: Visual reasoning on ambiguous logic puzzles. VERIFI allows UNIFUSION to leverage the world knowledge of the VLM encoder.

Instead of prompting the model to directly generate a 'yes/no' answer, we extended the inference time to compute for each question. To that end, we tasked the model to perform extensive **chain-of-thought** generation for all image aspects relevant to the question, before generating a 'yes/no' answer.

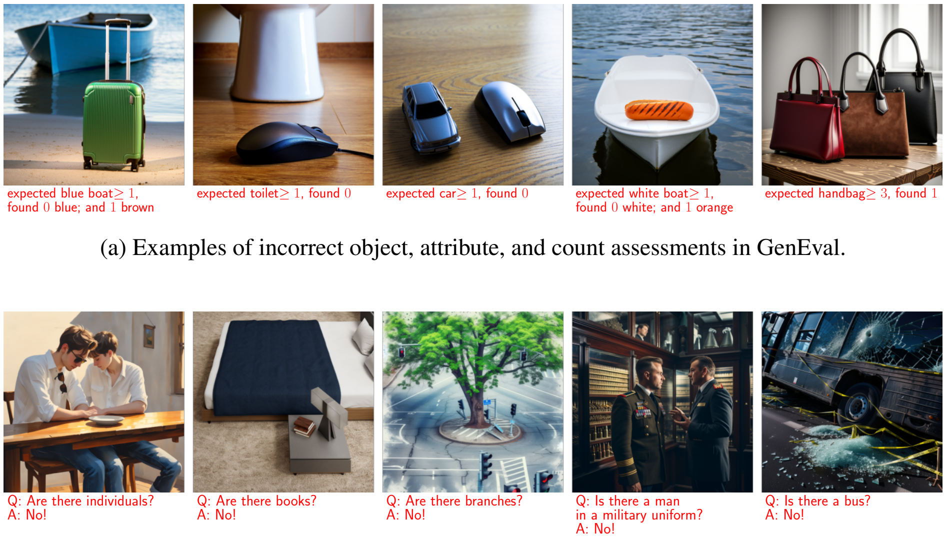
Fix score aggregation. The official DPG evaluation script provided in the author's GitHub does not aggregate scores correctly. While the overall score is calculated across all images per prompt, the subcategories only use the score of the last image. This implementation bug has also been pointed out by other users [0] but remains unfixed at the time of writing. Consequently, we re-implemented score aggregation to ensure correct results.

Improved presentation. Since subcategories in DPG-Bench are heavily skewed towards entities, we not only report the overall mean (Micro Avg), but also the mean of category-wise performance (Macro Avg). In line with classic Computer Vision literature, we also report the best-out-of-n performance in addition to the mean over multiple seeds. This score more accurately reflects the experience of most users, since many image generation platforms and local setups will provide multiple seeds to pick from.

A.4 EVAL PROMPT DISTRIBUTION

In our AB test, we outsourced to 200 annotators to evaluate 448 prompts of diverse editing types as shown in Fig 28. Each comparison requires at least 3 votes to be considered a valid sample in the win-rate calculation. Miscellaneous section in Fig 28 encompasses categories below 2% such as Add text, Change perspective, Colorization, Change time of day, Remove distortion, Cinematic lighting, Remove shadows, Replace materials... along with more than 20 other tasks.

[0] https://github.com/TencentQQGYLab/ELLA/issues/60


expected blue boat≥ 1, found 0 blue; and 1 brown
expected toilet≥ 1, found 0
expected car≥ 1, found 0
expected white boat≥ 1, found 0 white; and 1 orange
expected handbag≥ 3, found 1
Q: Are there individuals? A: No!
Q: Are there books? A: No!
Q: Are there branches? A: No!
Q: Is there a man in a military uniform? A: No!
Q: Is there a bus? A: No!

(a) Examples of incorrect object, attribute, and count assessments in GenEval.

(b) Examples of Question-Answering failures in DPG-Bench assessments.

Figure 27: Current generative image benchmarks incorrectly score simple examples, including the presence of clear subjects in the image.

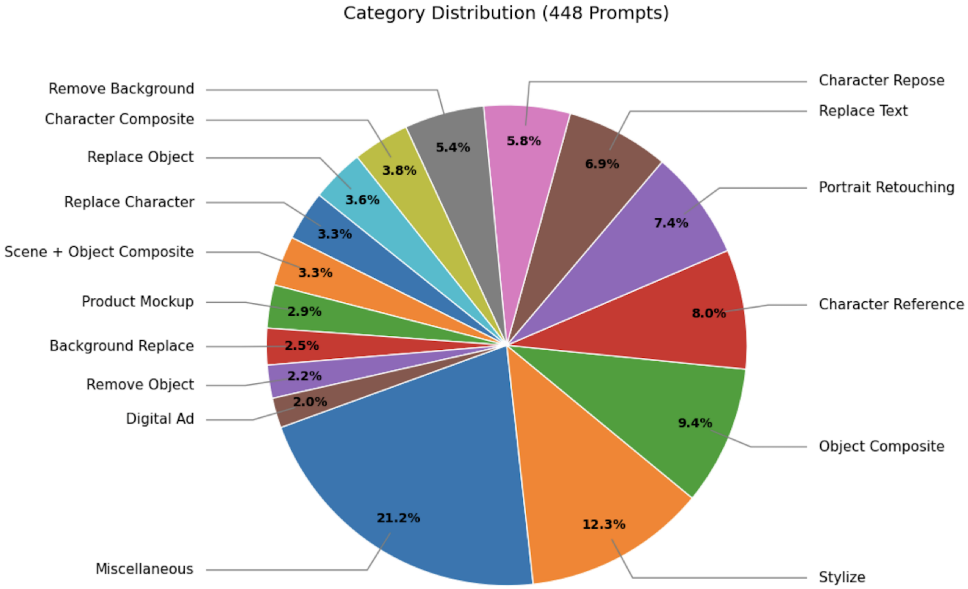


Figure 28: Distribution of edit-types in the A/B test set.