Injecting Image Guidance into Text-Conditioned Diffusion Models at Inference

Agata Żywot [1] **Iason Skylitsis** [1] **Thijmen Nijdam** [1] **Zoe Tzifa-Kratira** [1] **Derck Prinzhorn** [1] **Konrad Szewczyk** [1] **Aritra Bhowmik** [1]

Abstract

Text-to-image diffusion models like Stable Diffusion generate high-quality images from text, but lack a way to inject visual guidance (e.g. sketches, styles) at inference without retraining. Existing methods either require computationally expensive fine-tuning or rely on style transfer techniques that risk semantic misalignment with textual prompts. We introduce Visual Concept Fusion (VCF), a training-free method that enables visual concept injection into Stable Diffusion by aligning CLIP image features with the text embedding space. VCF consists of three components: (1) a lightweight aligner that maps image tokens to the text embedding manifold using InfoNCE and cross-attention reconstruction losses, (2) a fusion strategy that preserves both textual and visual semantics, and (3) an optional Prompt-Noise Optimization (PNO) module for test-time refinement. Our experiments demonstrate that VCF successfully transfers visual attributes including style, composition, and color palette from reference images while maintaining prompt adherence. Quantitative results show a trade-off between text alignment (CLIP score) and visual correspondence (LPIPS), with VCF outperforming baselines in reference fidelity.

1. Introduction

Recent advancements in text-to-image diffusion models, such as Stable Diffusion (Rombach et al., 2021), have enabled the creation of highly realistic and diverse images conditioned on natural language prompts. The samples generated by these models frequently exhibit rich textures and meaningful semantics, indicating a strong ability to capture information at both low (edges, textures) and high (semantics, composition) levels. However, guiding the models to represent users' ideas faithfully often requires significant effort dedicated to precise prompt engineering (Liu & Chilton, 2023).

To reduce reliance on precise prompting, an emerging solution is to incorporate visual references alongside text, such as sketches, style references, or exemplary images. While this method of conditioning can allow for more accurate and human-friendly guidance of the generation process, existing methods typically require additional fine-tuning (Mou et al., 2023; Zhang et al., 2023; Ruiz et al., 2023). Such fine-tuning can be computationally expensive and necessitates access to additional datasets. Alternative approaches, such as style transfer (e.g., AdaIN (Huang & Belongie, 2017)), may risk semantic misalignment with the textual prompt. Furthermore, even models designed for joint conditioning on text and image can be prone to overlooking or inadequately integrating reference image cues. As shown in Figure 1, such models may preserve a reference style (*Starry Night*) but apply it inconsistently to the textual subject (e.g., *a photo of a cat*). Effectively integrating such visual cues often demands further costly fine-tuning. Conversely, naively introducing image features into standard text-conditioned pipelines—such as directly adding image tokens through a weighted sum—presents an extrapolation problem, typically yielding poor-quality outputs. This highlights a critical gap: either the model must be retrained extensively for joint conditioning, or visual cues must be integrated in a more sophisticated, non-naive manner. This raises the question: *Can we guide image generation using visual references at inference, without retraining the underlying diffusion model?*

Figure 1: Illustration of challenges in visual guidance. Left: Reference image and text prompt. Middle: Output from a model trained for joint image-text conditioning (Ruiz et al., 2023), struggling with full style integration. Right: Output from a standard text-to-image model (SD) with naively blended image features, resulting in a distorted image.

[1]University of Amsterdam, Netherlands.

UvA Foundation Models Course 2025, Project Paper Submission.

In this paper, we explore the feasibility of injecting visual cues into text-to-image diffusion models at inference time without finetuning the generative model. Based on intuition stemming from previous works on adapter models (Mou et al., 2023), we posit that diffusion models can be efficiently controlled by adjusting the conditioning signal based on reference image features. However, naive methods for blending textual and image features yield unsatisfactory results due to misalignment between the distribution of textual and image features.

Therefore, we propose Visual Concept Fusion (VCF), an efficient approach for enabling style transfer capabilities in text-to-image diffusion models without the need for finetuning the diffusion model. Our method can be decomposed into three major components:

- **Modality alignment:** We train a small feature aligner model to alleviate the distribution mismatch between image and textual features. The training requires only a small amount of image–caption data and does not involve the generative diffusion model.
- **Text–image fusion:** We experiment with three distinct fusion methods for blending image and text tokens: (1) Naive fusion, (2) Concatenation, and (3) Cross-attention fusion.
- **Prompt–Noise Optimisation (PNO):** An optional test-time optimisation loop designed to further enhance semantic alignment. It refines both the conditioning signal and the initial noise input to the diffusion process, aiming to maximise the similarity between the generated image and a target visual reference in CLIP's embedding space.

In our work, we demonstrate that the images generated using VCF exhibit similarities in style, composition or colour palette with the reference images, while capturing the contents of the textual prompts. Moreover, we show empirically the impact that the choice of major components of our method (e.g. the aligner, PNO) has on the faithfulness and the quality of the generated samples.

2. Related Work

Deep generative image modelling. The generation of novel images has been a long-studied area of computer vision and deep learning research. Early approaches include Variational Autoencoders (VAEs) (Kingma & Welling, 2013), which learn an easy-to-sample latent space representation mapped to the image space with a trained decoder, and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which pit a generator against a discriminator during the training phase to produce increasingly realistic samples. While GANs in particular have been proven capable of achieving remarkable image quality (Karras et al., 2019; 2020), both of these models suffer from training instability and the potential for mode collapse.

More recently, Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have emerged as a powerful class of image generative models, demonstrating state-of-the-art performance. At their core are two processes — a fixed forward (diffusion) process that gradually adds Gaussian noise to an input sample over a sequence of T steps, and a learned reverse (denoising) process that reconstructs a sample from the target data distribution by gradually removing noise, starting from pure Gaussian noise. A significant improvement in making diffusion models more efficient, particularly when working with high-resolution data, was a class of models known as Latent Diffusion Models (LDMs) (Rombach et al., 2021). Instead of operating in the high-dimensional pixel space, these models perform diffusion and denoising in a lower-dimensional latent space, drastically reducing computational requirements.

Stable Diffusion (Rombach et al., 2021) is a prominent example of an LDM trained for the task of text-to-image generation. It uses CLIP (Radford et al., 2021) text embeddings as conditioning within the denoising model by injecting them via cross-attention mechanisms. This provided a significant breakthrough in highly realistic image synthesis; however, the conditioning signal is limited to text and introducing other conditioning modalities, such as reference images, poses a difficult challenge due to the features lying in misaligned data distributions.

Fine-tuning and adapter-based conditioning. A prominent line of work aiming to solve this problem involves augmenting or fine-tuning pre-trained diffusion models to accept additional image-based conditioning. DreamBooth (Ruiz et al., 2023) enables the personalisation of models by fine-tuning them on a small set of subject images. Similarly, textual inversion techniques (Gal et al., 2022) learn a distribution of new pseudo-words to represent specific visual styles or objects, which can be further combined for more precise personalisation.

Another family of approaches includes T2I-Adapter (Mou et al., 2023) and ControlNet (Zhang et al., 2023), which utilise lightweight, trainable modules that inject additional conditioning (e.g., based on visual cues from reference depth maps or sketches) into the frozen backbone of a pre-trained diffusion model. While enabling precise model steering based on various types of visual cues, these methods require training the adapter modules on large datasets of paired image–condition data. Although the core diffusion backbone remains frozen, the training process still demands computationally expensive image sampling at every training step. Our work diverges from these approaches by explicitly avoiding any training that would involve the denoising

model directly.

Training-free guidance. Training-free diffusion guidance methods aim to steer the generation process at inference time, leveraging the knowledge already present within a pre-trained model. While prompt engineering (Oppenlaender, 2023) can be used to steer generation, it is often complex and time-consuming to achieve results that faithfully reflect the user's intent. As one of the first approaches enabling training-free injection of a visual reference, SDEdit (Meng et al., 2022) and its application on models such as Stable Diffusion demonstrated that when a noisy version of a source image is denoised with a diffusion model, the result retains aspects of the source image while adhering to the original conditioning. However, this method is mostly limited to tasks in which the composition of the target image should resemble the reference image and, thus, does not work well for style transfer and similar problems.

Moreover, several techniques focus on manipulating the sampling process of pre-trained diffusion models. Plug-and-Play Diffusion Features (Tumanyan et al., 2022) allow for generation control by inverting the reference image using DDIM inversion (Song et al., 2022) into the initial noise, which is then denoised using a text-conditioned pre-trained model. Similarly, Add-It (Tewel et al., 2024) enables efficient object insertion into reference images by injecting additional information—provided by an external segmentation model (Ravi et al., 2024)—into the attention mechanism of the denoising model. However, both of those methods share the same problem as SDEdit. In contrast, our method is capable of transferring also the high-level concepts such as the art-style or content from the reference image.

3. Method

We propose Visual Concept Fusion (VCF), a novel pipeline that integrates image guidance into text-conditioned diffusion models. As shown in Figure 2, VCF comprises three key components: (1) an Image Aligner that maps image tokens into the text embedding space for modality alignment; (2) a Text–Image Fusion block that merges aligned image and text features; and (3) an optional Prompt–Noise Optimisation (PNO) module that optimises the generation process at inference.

3.1. Image-to-Text Alignment

Stable Diffusion v2 (SDv2) conditions its denoising network on *pre-projection* tokens from the CLIP text encoder. We denote these tokens by $T \in \mathbb{R}^{n \times d_{\text{text}}}$, drawn from the distribution $p_{\text{text}}(T)$. Pre-projection tokens are preferred because they preserve richer linguistic detail than the final projected text vector—a single $1 \times d_{\text{proj}}$ embedding—used in CLIP's final contrastive loss during training.

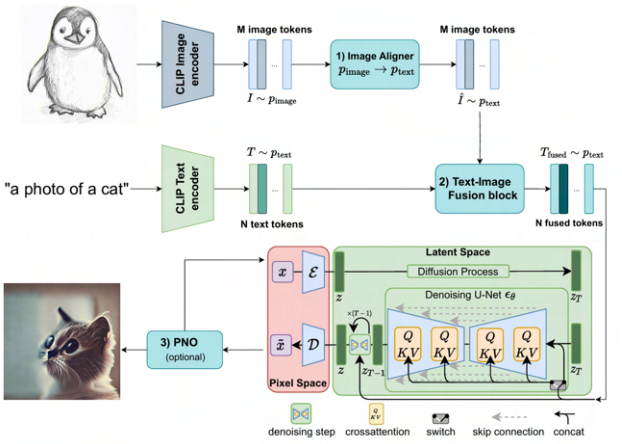

Figure 2: VCF pipeline overview. The pipeline integrates image guidance into text-conditioned diffusion models via three key components: (1) the Image Aligner maps image tokens to the text embedding space; (2) the Text–Image Fusion module combines aligned image and text tokens into fused representations; and (3) PNO (optional) refines the fused conditioning and initial noise to enhance visual alignment in the final output.

To inject visual guidance, we likewise extract pre-projection tokens from the CLIP image encoder, yielding $I \in \mathbb{R}^{m \times d_{\text{image}}}$ with distribution $p_{\text{image}}(I)$. Although the text and image branches are trained jointly, their alignment is enforced only *after* the linear projection layers used for the contrastive loss. Consequently, the two pre-projection spaces are not yet aligned, so $p_{\text{text}} \neq p_{\text{image}}$. Injecting I directly into a text-conditioned SDv2 model therefore creates a modality mismatch, which we quantify via the KL divergence

$$\Delta_{\text{KL}} = \text{KL}\big(p_\theta(x_0 \mid I) \,\|\, p_\theta(x_0 \mid T)\big),$$

where x_0 denotes the final denoised sample. A large Δ_{KL} leads to unstable denoising and images that are neither faithful to the reference nor well aligned with the prompt.

Aligner architecture. To mitigate this mismatch, we introduce a lightweight aligner f_ϕ: a two-layer MLP with LayerNorm and ReLU activations. It is the only component in the VCF pipeline that is trained from scratch; the underlying SD model remains frozen. The aligner maps image tokens to an aligned representation $\hat{I} = f_\phi(I) \in \mathbb{R}^{m \times d_{\text{text}}}$.

Global alignment objective. We encourage the distribution of the aligned tokens $\hat{I}$ to match that of the text tokens T via an InfoNCE loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\big(\cos(\mu_{\hat{I}}, \mu_T)/\tau\big)}{\sum_j \exp\big(\cos(\mu_{\hat{I}}, \mu_{T_j})/\tau\big)},$$

where $\mu_{\hat{I}}$ and μ_T are mean embeddings of the image and text tokens, respectively, and τ is a learnable temperature.

Local alignment objective. To preserve token-level structure, we add a cross-attention reconstruction loss. Text tokens are reconstructed from the aligned image tokens:

$$T' = \mathrm{Attn}(Q = \hat{I},\ K = T,\ V = T), \quad \mathcal{L}_{\text{attn}} = \|T' - T\|_2^2.$$

Joint training. The aligner parameters ϕ are learned with the combined loss:

$$\mathcal{L}_{\text{align}} = \lambda\, \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{attn}}.$$

We set $\lambda = 0.2$. Minimising $\mathcal{L}_{\text{align}}$ realigns the image-derived tokens with the text-embedding manifold, thereby reducing Δ_{KL} and enabling SD to utilise reference images without sacrificing prompt fidelity.

3.2. Text–Image Fusion

After aligning the image tokens $\hat{I}$ to the text embedding space, we fuse them with the original text tokens T so that both modalities can guide the diffusion process. We consider three fusion strategies.

Naive (mean) fusion. The simplest strategy injects the *same* image-derived signal into every text token. Given $\hat{I} \in \mathbb{R}^{m \times d_{\text{text}}}$ and $T \in \mathbb{R}^{n \times d_{\text{text}}}$ with $m \neq n$, we first average the image tokens,

$$\hat{I}_{\text{global}} \;=\; \frac{1}{m} \sum_{j=1}^{m} \hat{I}_j \in \mathbb{R}^{d_{\text{text}}},$$

and linearly blend this vector with each text token:

$$T_i^{\text{fused}} \;=\; (1-\alpha)\, T_i + \alpha\, \hat{I}_{\text{global}}, \qquad i = 1, \ldots, n,$$

where $\alpha \in [0, 1]$ controls the influence of the image signal. Although straightforward, this uniform perturbation often suppresses linguistic nuances in T, leading to noisy and semantically inconsistent outputs; we therefore retain it only as a baseline and refer to it as *naive fusion*.

Concatenation fusion (VCF). Our primary method simply concatenates the aligned image tokens to the end of the text sequence, $[T; \hat{I}]$, and feeds the combined tokens to Stable Diffusion unchanged. This preserves the individual semantics of each modality and, empirically, yields the best balance between prompt fidelity and reference adherence.

Cross-attention fusion. A third variant allows the text tokens to attend to the image tokens, producing a cross-attended representation that is re-scaled and blended back into the text at every denoising step. While this approach alleviates some artifacts of naive fusion, it does not match the performance of concatenation fusion in our experiments. Implementation details and qualitative examples appear in Appendix B.

3.3. Prompt–Noise Optimisation

The final optional component in our VCF pipeline is Prompt–Noise Optimisation (PNO), a test-time procedure that refines both the conditioning tokens T_{final} and the initial diffusion noise x_T before commencing the reverse sampling process. Our approach is inspired by the original Prompt–Noise Optimisation work (Peng et al., 2024), which aimed to mitigate undesirable toxicity in generated images by optimising prompt embeddings and the noise trajectory. We adapt this framework by modifying the optimisation objective: instead of minimising a toxicity score, our PNO seeks to maximise the similarity (i.e., minimise the negative similarity) between the eventually generated image x_0 and a user-provided visual reference x_{guide}.

This optimisation leverages the CLIP model's embedding space. Specifically, we jointly optimise T_{final} and x_T to improve the cosine similarity between the CLIP embedding of the generated image x_0 and that of the reference image x_{guide}, while applying a regularisation term to the initial noise x_T:

$$\begin{aligned}\mathcal{L}_{\text{PNO}} = \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(x_T) \\ - \cos\left(\text{CLIP}(f(x_T, T_{\text{final}})),\ \text{CLIP}(x_{\text{guide}})\right)\end{aligned}$$

Here, $f(x_T, T_{\text{final}})$ represents the diffusion model's generation process that yields x_0 from x_T and T_{final}. x_{guide} is the reference image capturing the desired visual concept. $\mathcal{L}_{\text{reg}}$ is a noise trajectory regularisation loss designed to prevent degenerate solutions and maintain a plausible noise structure for x_T. λ_{reg} is a weighting factor balancing the two terms (set to 0.1 by default). In our experiments, this optimisation is performed for a small number of gradient steps (10–50) prior to initiating the full DDIM sampling.

It is important to note the role of x_T optimisation in this context. While the original PNO paper (Peng et al., 2024) discussed the concept of optimising the entire noise trajectory, which controls detailed image features, we operate within the framework of a deterministic DDIM sampler. For DDIM, the entire denoising trajectory—and consequently the final generated image x_0—is uniquely determined by the initial noise x_T (given fixed conditioning T_{final} and model parameters). Therefore, in our PNO implementation, optimising the *noise trajectory* effectively translates to optimising this initial noise x_T. Modifying x_T allows us to steer the generation towards better alignment with x_{guide} without compromising image quality, as significant deviations from a standard Gaussian distribution for intermediate noise steps could degrade generation quality.

3.4. Temporal Conditioning Ablation and Attention Map Visualization

To further understand the behavior of our VCF pipeline, we conducted two additional experimental analyses. First, we explored the temporal aspects of visual conditioning by implementing a conditioning schedule that applies visual guidance only during specific portions of the diffusion sampling process (e.g., first half, second half, or middle portion). This analysis, conducted using the naive fusion approach, reveals how different timesteps contribute to various aspects of visual transfer.

Second, we performed attention map visualization to understand how different fusion weights affect the model's attention patterns during generation. Detailed results and analysis for both the conditioning schedule experiments and attention map visualizations are provided in Appendices C and D, respectively.

4. Results

We evaluate the effectiveness of our VCF pipeline on the task of guided image generation, where both a reference image and a textual prompt jointly influence the output. We first describe the experimental setup and evaluation metrics, followed by an qualitative and quantitative analysis of the results.

4.1. Experimental Setup

All experiments are conducted using the publicly available Stable Diffusion v2 model[1] (768-ema-pruned variant), with DDIM sampling over 50 steps at a resolution of 768×768 pixels. Our aligner is trained on a 10% subset of the COCO Captions dataset[2], consisting of approximately 60,000 randomly selected image–caption pairs. We use an 80/10/10 split for training, validation, and testing, respectively. The training objective combines InfoNCE with a cross-attention reconstruction loss, as described in section 3. Training the aligner is computationally lightweight and completes in under two hours on a single A100 GPU.

Dataset. COCO Captions (Chen et al., 2015) is a large-scale image–caption dataset comprising over 120,000 images, each annotated with five human-written descriptions. The captions exhibit a high degree of linguistic diversity, often including compositional and stylistic elements, making the dataset well suited for learning rich text–image alignments. During training, we randomly sample one of the five captions for each image in every epoch to encourage robustness to paraphrasing.

Hyperparameters. We adopt standard diffusion settings and introduce additional parameters for the aligner and Prompt–Noise Optimisation (PNO). The InfoNCE loss uses a learnable temperature parameter τ, and we balance it with the cross-attention reconstruction loss using a fixed weight of $\lambda_{\text{align}} = 0.2$. We use fusion strength $\alpha = 0.3$, and apply PNO as an optional test-time refinement. Full hyperparameter details, grouped by component, are provided in Table 1.

Table 1: Hyperparameters used in all experiments, grouped by component.

Diffusion Parameters	
Base model	Stable Diffusion v2.1 (768-ema-pruned)
Image resolution	768×768
Sampling method	DDIM
DDIM steps	50
Aligner Parameters	
Training dataset	COCO Captions (10%)
Loss function	InfoNCE + Cross-Attention Reconstruction
Loss weighting λ_{InfoNCE}	0.2
InfoNCE temperature τ	Learnable
Training epochs	10
PNO Parameters	
PNO steps	10–50
Learning rate (PNO)	1×10^{-2}
Noise regularisation λ_{reg}	0.1
Gradient clipping	1.0

4.2. Evaluation Metrics

To evaluate the quality of generated images, we consider two main criteria: alignment with the input text prompt, and correspondence to the visual reference. The following metrics are used:

CLIP Score (Text Alignment). We quantify semantic alignment between the generated image and the text prompt using the CLIP similarity score. Specifically, we compute the cosine similarity between their embeddings in the CLIP space:

$$\text{CLIP}(x, t) = \frac{f_{\text{CLIP}}(x) \cdot f_{\text{CLIP}}(t)}{\|f_{\text{CLIP}}(x)\| \, \|f_{\text{CLIP}}(t)\|}$$

where $f_{\text{CLIP}}(\cdot)$ denotes the CLIP encoder applied to images and text, respectively. Higher values indicate stronger alignment.

LPIPS (Reference Image Correspondence). The Learned Perceptual Image Patch Similarity (LPIPS)

[1] https://github.com/Stability-AI/stablediffusion

[2] https://huggingface.co/datasets/sentence-transformers/coco-captions

(Ghazanfari et al., 2023) metric measures perceptual similarity between the generated image $\hat{x}$ and the reference image x_{ref}. It is defined as:

$$\text{LPIPS}(x_{\text{ref}}, \hat{x}) = \sum_l w_l \, \|\phi_l(x_{\text{ref}}) - \phi_l(\hat{x})\|_2^2$$

where ϕ_l are features extracted from layer l of a pretrained VGG network (Simonyan & Zisserman, 2015), and w_l are learned weights. In our setup, we do not learn custom weights and instead fix $w_l = 1$ across all layers. Lower LPIPS scores indicate greater perceptual similarity to the reference image.

4.3. Qualitative Results

We present an overview of qualitative results in Figure 3, comparing three generation modes: (i) text-only generation using SDv2, (ii) naive fusion, and (iii) our proposed VCF pipeline. All outputs are conditioned on the same prompt—"A photo of a cat"—with only the reference image varying across samples to isolate its influence on the output. Additional examples are provided in Appendix A.

As expected, naive fusion does not reliably integrate information from the reference image. While the generated images depict cats, they often appear less realistic and exhibit elevated visual noise. In many instances, these outputs closely resemble those produced by the text-only baseline, indicating that naive fusion fails to meaningfully modulate generation based on the visual reference.

By contrast, generations produced by our VCF method exhibit a much stronger correspondence with the reference image. The transferred features span both high-level semantics (e.g., artistic style, presence of background objects) and low-level visual cues (e.g., colour distribution, shading, depth). For example, when a dog is used as the reference, the output often resembles a hybrid "cat–dog" entity that blends shape and colour characteristics from both the text prompt and the image. Moreover, the level of realism in the generated outputs tends to reflect the style of the reference: photorealistic inputs yield realistic generations, while stylised references—such as paintings or prints—result in outputs with matching stylistic attributes.

4.4. Quantitative Results

Table 2 reports performance across two metrics: CLIP score, which measures alignment with the text prompt, and LPIPS, which quantifies perceptual similarity between the generated image and the visual reference.

As expected, the text-only SDv2 model achieves the highest CLIP score, reflecting strong semantic adherence to the prompt. Naive fusion yields a slightly reduced CLIP score, likely due to the noisier and less coherent outputs. Our VCF method shows a further reduction in CLIP score, which is anticipated given its increased reliance on visual guidance. This trade-off is evident in cases such as the "cat–dog" hybrid or stylised cat generations shown in Figure 3, where visual fidelity to the reference image overrides strict prompt literalism.

Figure 3: Qualitative comparison of generation methods. Each row shows (left → right): the reference image, baseline text-only SDv2 output, naive fusion, and our proposed VCF.

In contrast, VCF achieves the highest LPIPS score, indicating the greatest perceptual similarity to the reference images. This result confirms that our method more effectively integrates visual features from the reference. Naive fusion, by comparison, obtains the lowest LPIPS score, consistent with its limited capacity to meaningfully condition on the reference and its tendency to revert toward the text-only baseline.

Table 2: Quantitative comparison of generation methods. CLIP indicates alignment with the text prompt (higher is better), and LPIPS measures perceptual similarity to the reference image (higher is better). Best results per metric are shown in bold.

Method	**CLIP** ↑	**LPIPS** ↓
SDv2 (text-only)	**0.29**	0.78
Naive fusion	0.28	0.77
VCF (Ours)	0.27	**0.76**

5. Ablations

To further assess the contributions of individual components in the VCF pipeline, we conduct a series of ablation experiments. All generations are conditioned on the same prompt—"A photo of a cat"—as in previous evaluations.

5.1. Effect of the Aligner Loss Function

The VCF aligner is trained using a combined objective comprising an InfoNCE loss and a cross-attention reconstruction loss. To understand the role of each term, we retrain the aligner under two ablated configurations: (i) InfoNCE-only, and (ii) cross-attention-only. Qualitative results are shown in Figure 4.

With the InfoNCE-only aligner, generated images display little or no visual resemblance to the reference image, although the overall image quality remains comparable to SDv2. This suggests that global distribution alignment alone is insufficient to guide the cross-attention mechanism in Stable Diffusion.

In contrast, using only the cross-attention loss produces outputs that closely follow the reference image, often at the expense of prompt fidelity. For instance, when given a dog as reference, the model generates an image of a dog—even though the prompt specifies a cat. Similarly, a reference depicting a girl in a floral setting yields an output of a girl surrounded by flowers.

Combining both losses achieves a more desirable balance. The InfoNCE term regularises the embedding space globally, while the cross-attention term injects local structure and fine-grained visual cues. This combination enables VCF to produce outputs that respect both the semantics of the prompt and the salient features of the reference.

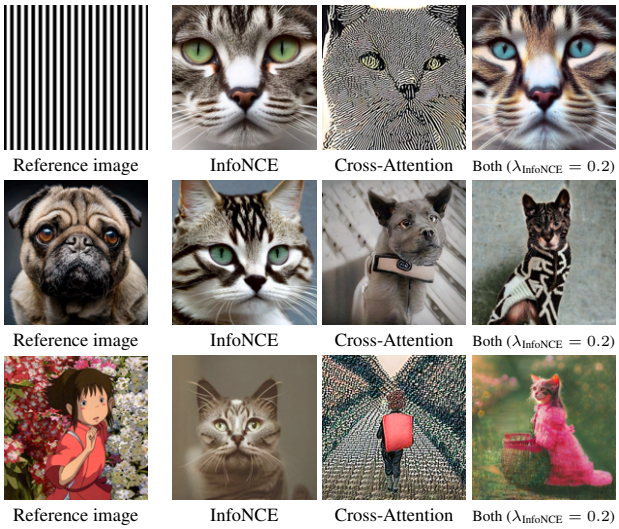

Figure 4: Ablation on aligner loss functions. Each row presents the reference image (left) followed by generations using an InfoNCE loss, a cross-attention reconstruction loss, and their combination with $\lambda_{\text{InfoNCE}} = 0.2$.

5.2. Effect of Prompt–Noise Optimisation (PNO)

We investigate the impact of the PNO module on the final image generation. PNO is applied at test time to refine both the conditioning signal and the initial noise, aiming to enhance alignment with the reference image.

Figure 5 illustrates the effect of applying PNO to the text-only SDv2 model. Even without fusion-based guidance, incorporating a reference image during the optimisation process leads to outputs that exhibit improved structure and visual similarity to the reference. Figure 6 shows the qualitative effect of PNO when applied to generations produced using the cross-attention fusion method, with an image guidance strength of $\alpha = 0.3$. The number of PNO steps is fixed at 50.

These results highlight the value of PNO as a refinement mechanism. In Figure 5, PNO improves both structural alignment and fidelity to the reference image, even in the absence of explicit fusion. When used in conjunction with cross-attention fusion, PNO helps suppress visual noise and artefacts introduced during fusion (e.g., Figure 6, top row), and can further steer the output toward reference-specific details (e.g., Figure 6, bottom row). For instance, PNO enhances colour fidelity by amplifying characteristic features such as the orange stripes on the cat.

Overall, these qualitative examples suggest that PNO consistently improves both perceptual alignment with the reference image and the visual quality of the generated output.

6. Discussion

Our experiments demonstrate that Visual Concept Fusion (VCF) provides an effective framework for integrating reference images into text-conditioned diffusion models. The results show that naive fusion fails to meaningfully steer generation, whereas VCF consistently produces outputs that reflect both the prompt and the reference. These outputs capture a range of visual attributes, including style, shape, and texture, and adapt to the realism or abstraction of the reference image. The ablations confirm that both components of our method—the aligner and the Prompt–Noise Optimisation—contribute to this improved control.

Limitations While promising, our work also has several limitations. First, there is no mechanism to control which visual features of the reference image are incorporated into the final output, which may result in unpredictable or overly

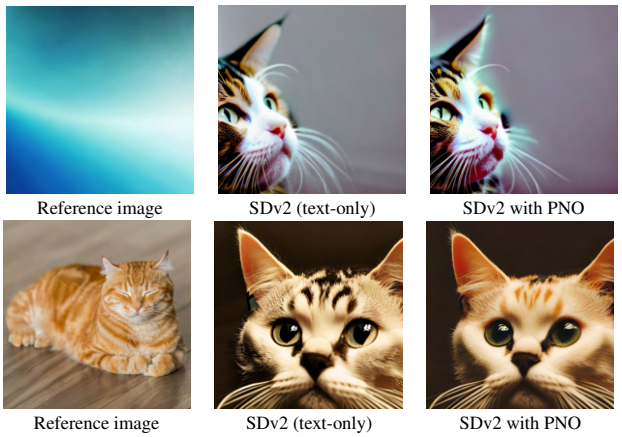

Figure 5: Effect of PNO on text-only SDv2. Each row shows: the reference image (left), generation using only the text prompt (middle), and the result after applying PNO (right). PNO improves alignment with reference image features.

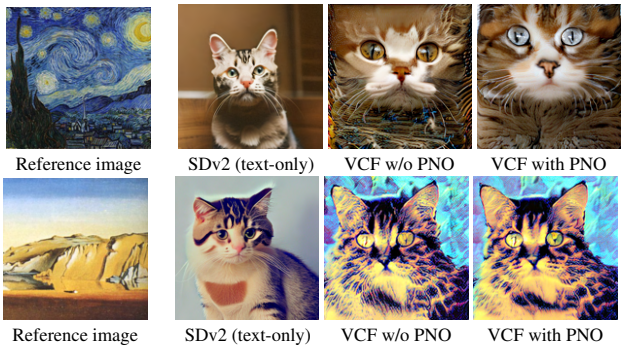

Figure 6: Qualitative results of Prompt-Noise Optimization (PNO). Each row shows a reference image, the text-only generation from Stable Diffusion v2, our VCF pipeline without PNO, and our VCF pipeline with PNO. PNO can reduce noise (top row) and improve adherence to reference image details like color patterns (bottom row, more orange stripes).

dominant influence. Second, our ablation studies on aligner training are limited: we only compare loss functions (InfoNCE, cross-attention, or both), using a single dataset (COCO) and one randomly sampled caption per image. Exploring different datasets (e.g., Flickr30K (Plummer et al., 2015)) or caption strategies may further improve alignment. Lastly, due to time constraints, we were unable to benchmark VCF against existing reference-guided baselines such as SDEdit (Meng et al., 2022), limiting direct comparison with prior work.

Future research could address these limitations by introducing finer control over transferred features, extending training regimes, and evaluating VCF in broader comparative settings.

References

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Ghazanfari, S., Garg, S., Krishnamurthy, P., Khorrami, F., and Araujo, A. R-lpips: An adversarially robust perceptual similarity metric. *arXiv preprint arXiv:2307.15157*, 2023.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. URL https://arxiv.org/abs/1703.06868.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan, 2020. URL https://arxiv.org/abs/1912.04958.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Liu, V. and Chilton, L. B. Design guidelines for prompt engineering text-to-image generative models, 2023. URL https://arxiv.org/abs/2109.06977.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL https://arxiv.org/abs/2108.01073.

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out

more controllable ability for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.08453.

Oppenlaender, J. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour amp; Information Technology*, 43(15):3763–3776, November 2023. ISSN 1362-3001. doi: 10.1080/0144929x.2023.2286532. URL http://dx.doi.org/10.1080/0144929X.2023.2286532.

Peng, J., Tang, Z., Liu, G., Fleming, C., and Hong, M. Safeguarding text-to-image generation via inference-time prompt-noise optimization, 2024. URL https://arxiv.org/abs/2412.03876.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Tewel, Y., Gal, R., Samuel, D., Atzmon, Y., Wolf, L., and Chechik, G. Add-it: Training-free object insertion in images with pretrained diffusion models, 2024. URL https://arxiv.org/abs/2411.07232.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. URL https://arxiv.org/abs/2211.12572.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023. URL https://api.semanticscholar.org/CorpusID:256827727.

APPENDIX

A. Additional Qualitative Examples of Main Results

An interesting observation is that image guidance becomes particularly crucial when the text prompt is somewhat vague or abstract. This is exemplified clearly in Figure 8, where the default Stable Diffusion model (SDv2)—conditioned solely on text—struggles to generate coherent and meaningful characters from the prompt "A charming character emerging from the scene". However, introducing reference image conditioning significantly improves the quality, detail, and coherence of the generated characters, making them visually captivating and semantically meaningful. Additionally, the continued poor performance of naive fusion further emphasizes the complexity of effectively integrating visual and textual modalities. This highlights the challenging nature of the problem and demonstrates the effectiveness of our proposed fusion method, which significantly improves visual coherence and semantic alignment.

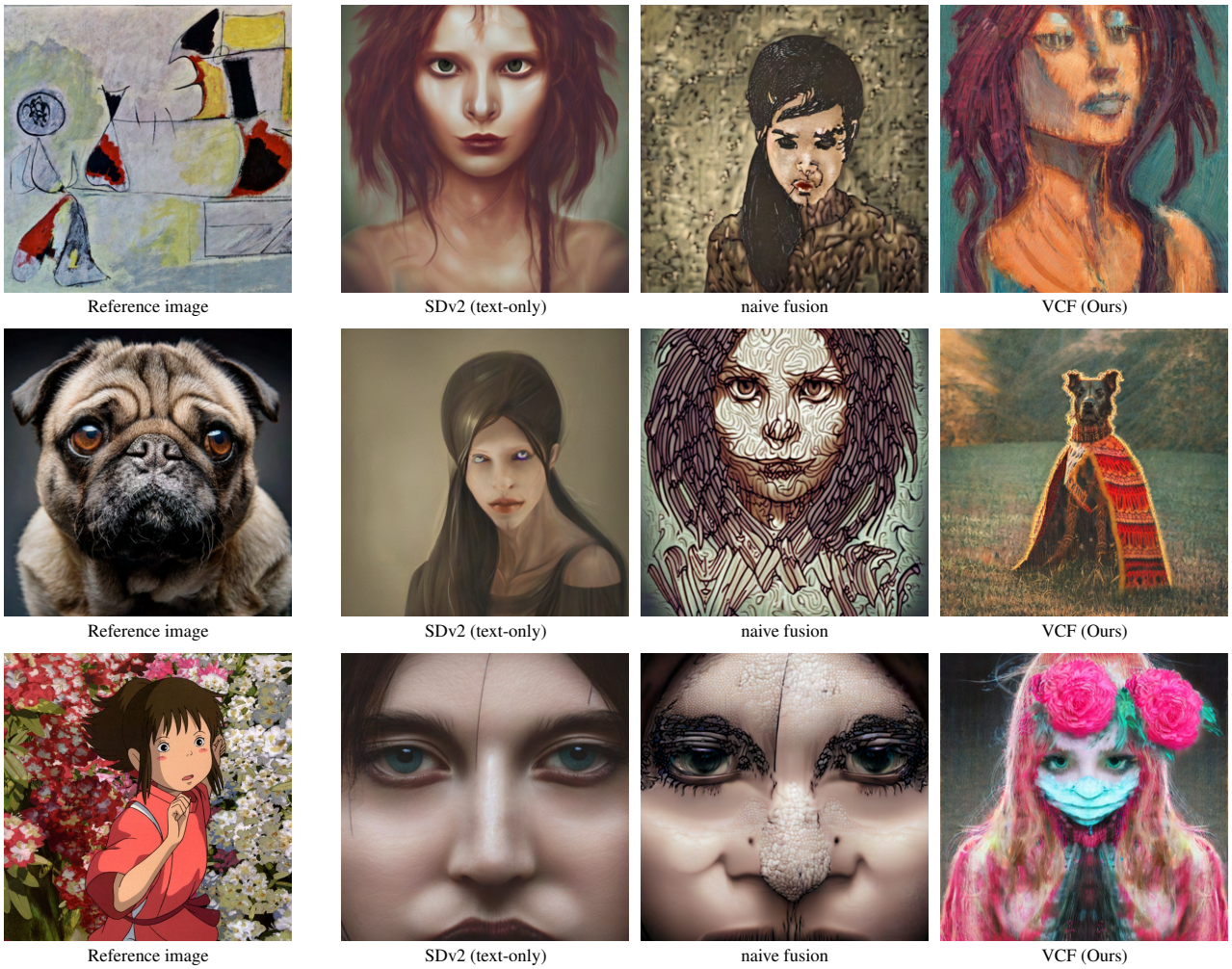

Figure 7: Additional qualitative examples of the main results using the prompt "A beautiful portrait of a mysterious character". Each row shows (left → right): the reference image, baseline text-only SDv2 output, naive fusion, and our proposed VCF.

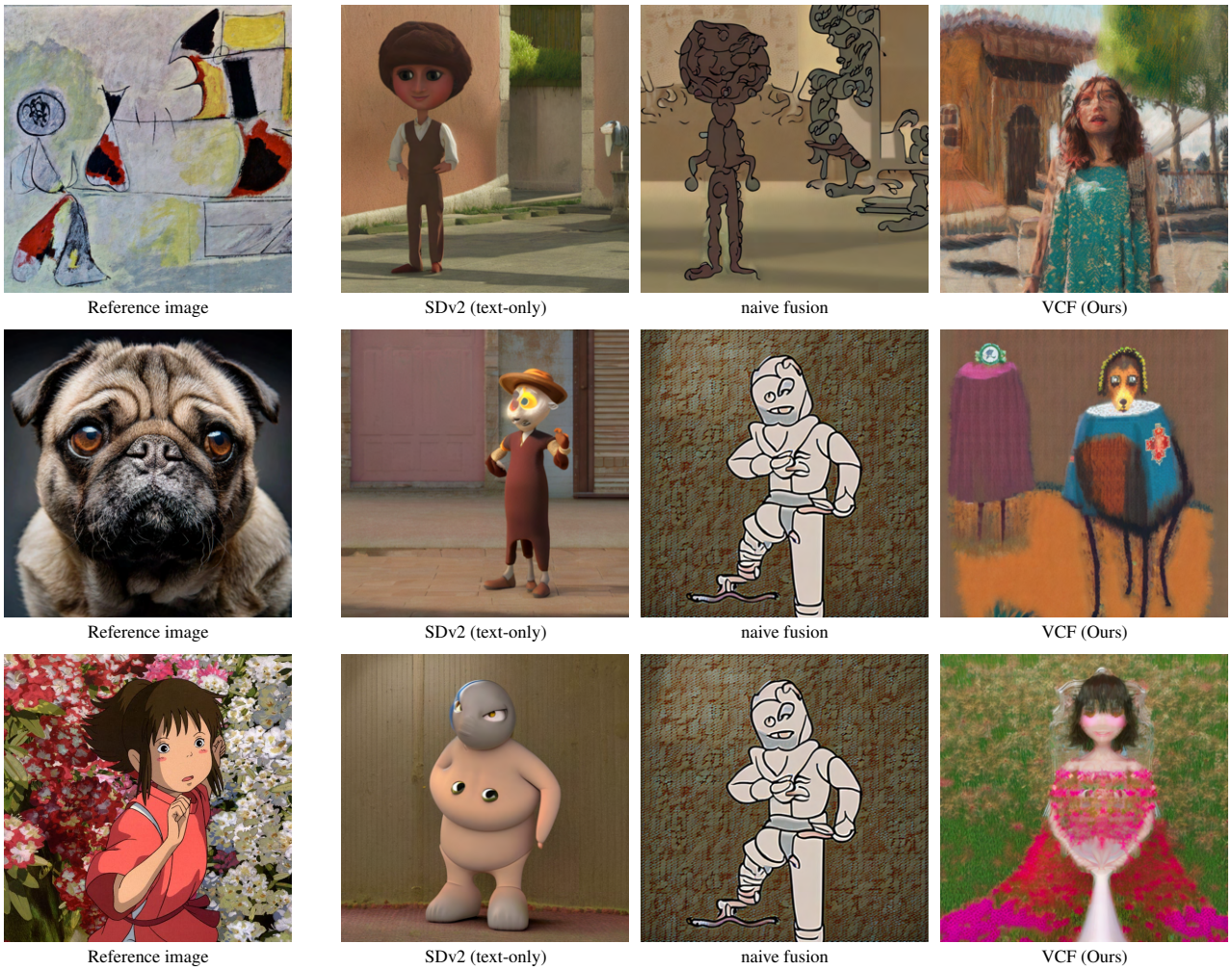

Figure 8: Additional qualitative examples of the main results using the prompt ”A charming character emerging from the scene”. Each row shows (left $\rightarrow$ right): the reference image, baseline text-only SDv2 output, naive fusion, and our proposed VCF.

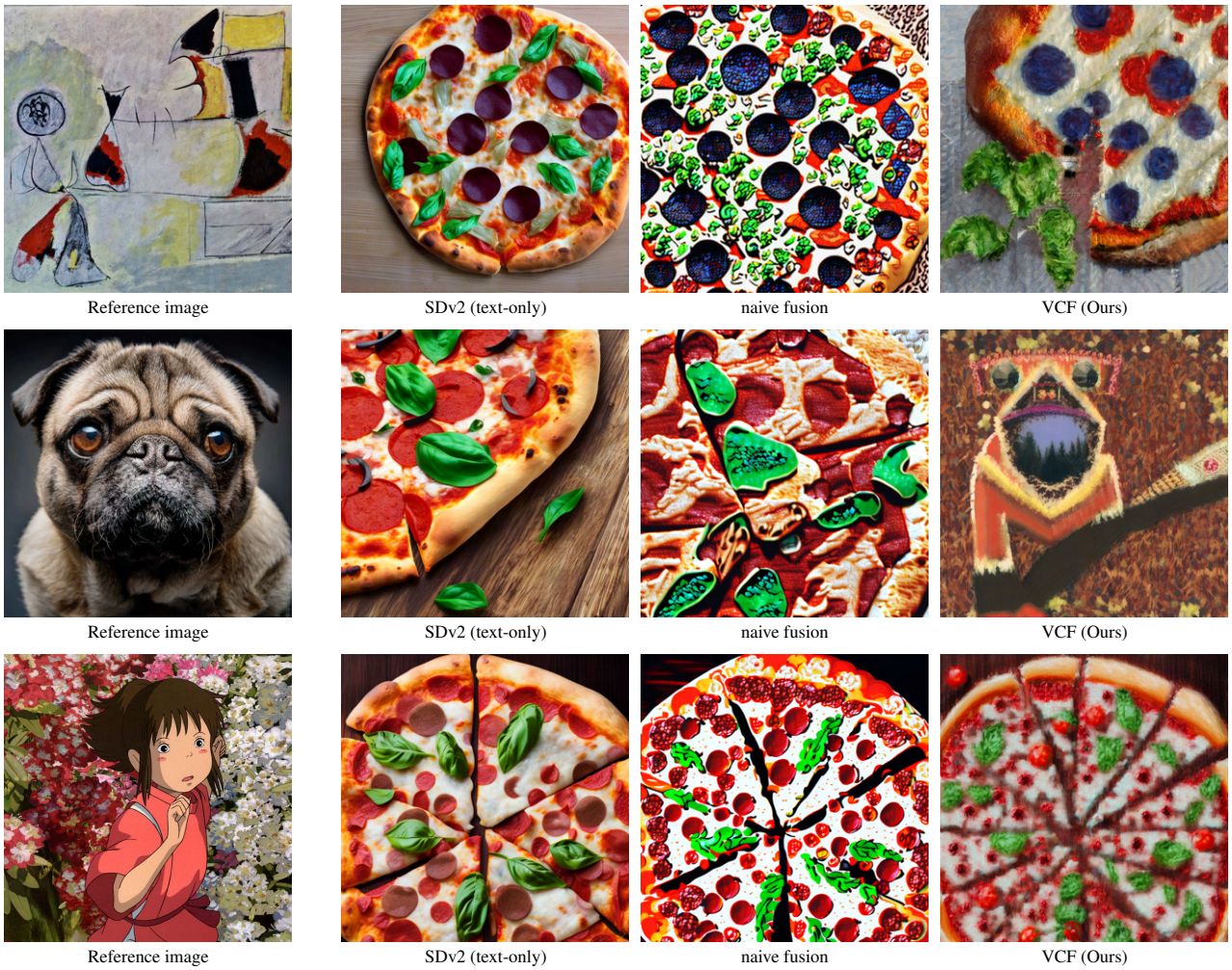

Figure 9: Additional qualitative examples of the main results using the prompt "a delicious pizza". Each row shows (left $\rightarrow$ right): the reference image, baseline text-only SDv2 output, naive fusion, and our proposed VCF.

B. Cross-Attention Fusion

As an alternative to concatenation, we experimented with a cross-attention fusion scheme. The idea is to let the text tokens query the aligned image tokens, thereby injecting fine-grained visual cues into the conditioning stream.

Fusion mechanism. Given text tokens T and aligned image tokens $\hat{I}$, we compute

$$T_{\text{fused}} = \text{Attn}(Q = T,\ K = \hat{I},\ V = \hat{I}),$$

and blend the result with the original text tokens,

$$T_{\text{final}} = (1 - \alpha)\, T + \alpha\, \gamma\, T_{\text{fused}},$$

where $\alpha \in [0, 1]$ sets the overall weight of the image signal. The factor γ rescales T_{fused} at every denoising step so that its norm remains comparable to that of T.

Qualitative observations. Representative outputs are shown in Figure 10. Cross-attention fusion does transfer some reference features, but the resulting images are noticeably noisier and less coherent than those produced by concatenation fusion, and in several cases introduce artefacts not present in either the prompt or the reference. Hence we retain this variant only for completeness and defer to concatenation fusion in the main paper.

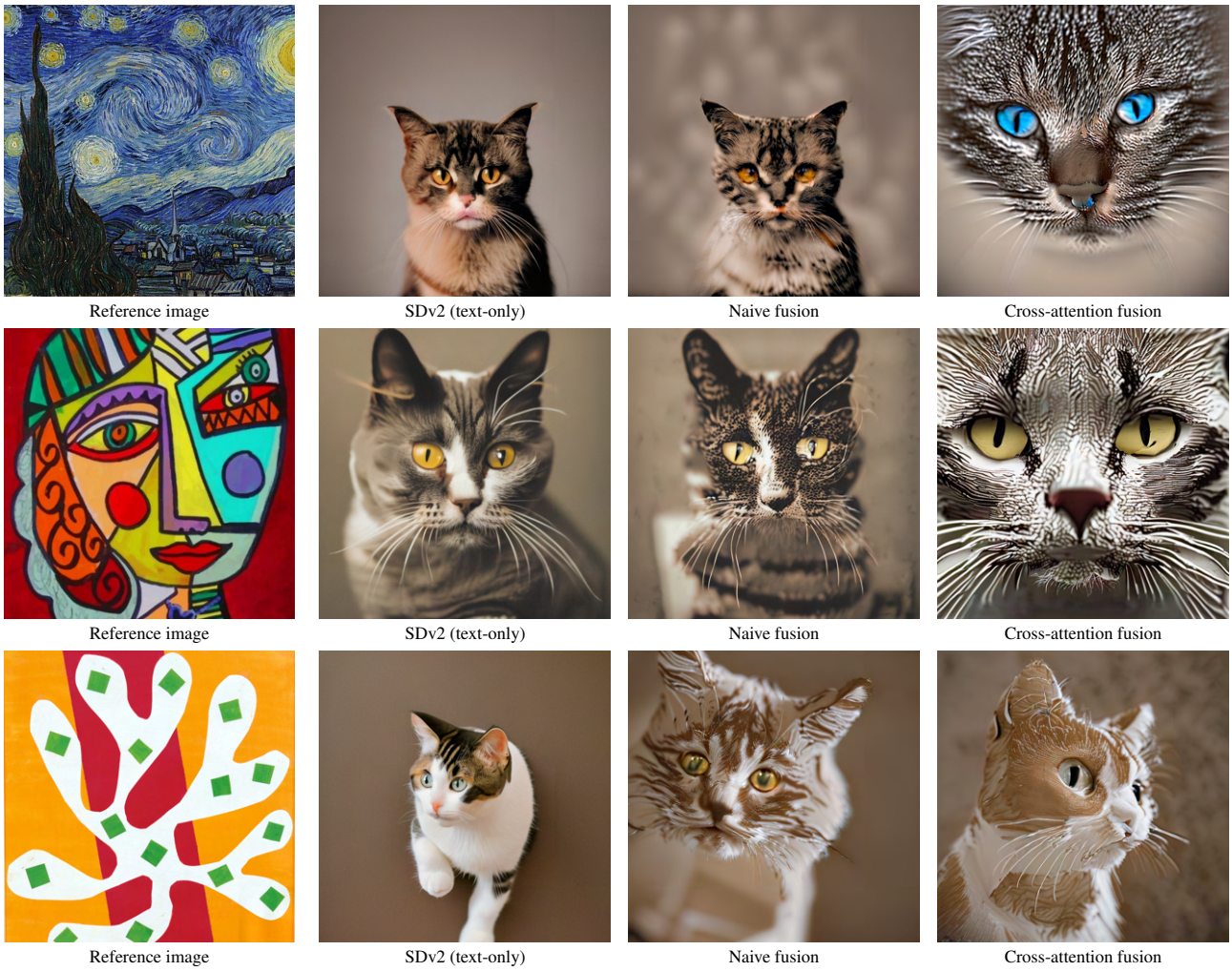

Figure 10: Qualitative ablation on fusion strategy. Each row shows (left → right): the reference image, baseline text-only SDv2 output, naive token fusion, and cross-attention fusion.

C. Conditioning Schedule Results

During the diffusion sampling process, different timesteps contribute to different aspects of image generation. Early timesteps typically determine high-level structure and composition, while later timesteps refine details and textures. We hypothesized that selectively applying visual conditioning during specific portions of the sampling process could provide more nuanced control over how reference image features are incorporated, potentially improving the balance between prompt fidelity and visual transfer.

We implemented a conditioning schedule that allows the blended conditioning signal to be applied only during specified fractions of the DDIM sampling process. Specifically, we tested four scheduling strategies:

- **Throughout:** Visual conditioning applied across all 50 sampling steps (baseline)
- **First Half:** Visual conditioning applied only during steps 1-25 **Second Half:** Visual conditioning applied only during steps 26-50 **Middle Portion:** Visual conditioning applied only during steps 15-35 (0.3-0.7 of the process)

This scheduling approach was implemented within the ”naive” (mean token) fusion mechanism, where the fusion weight α is set to 0 (text-only) outside the specified conditioning windows.

Figure 11 presents qualitative results across three reference images with different styles. We observe that conditioning throughout the sampling process produces the strongest visual transfer (most prominent color transfer is observed here) but sometimes at the cost of prompt faithfulness, as seen in the middle row where the abstract art reference heavily influences the cat's appearance. Additionally and as expected, conditioning only on the first half tends to have more pronounced stylistic results than conditioning only on the second half, which maintains the structural composition established by the text prompt in early timesteps. Qualitatively, First Half blended conditioning yields the best results, as they are smoother and less noisy than if we condition on the entire sampling process. The Middle portion conditioning setting presents as a less noisy and stylistically equivalent approach to the Second Half conditioning.

While these experiments were conducted using the naive fusion approach, the conditioning schedule method could readily be incorporated into the concatenation fusion method that formed our main results. This would provide researchers with additional control over the trade-off between prompt fidelity and visual transfer, potentially enabling more sophisticated applications such as structure-preserving style transfer or content-aware visual conditioning.

Figure 11: Qualitative ablation on the naive fusion type, where the blended conditioning is only applied for a fraction of the diffusion sampling process. Each row shows (left $\rightarrow$ right): the reference image, baseline text-only SDv2 output, blended conditioning applied on the whole diffusion procces, on the first half only, on the second half only, and on the middle portion of the diffusion process (0.3-0.7).

D. Attention maps visualization

We visualized attention maps for output blocks 4 and 11 across different fusion weights ($\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$). In our setting, the reference image conveys the style, while the object is specified by the text prompt. For low fusion weights, attention maps remain focused and align well with the object described in the prompt. As α increases and the model relies more on the reference style, attention becomes increasingly diffuse and random, mirroring the distortions observed in the generated outputs. This suggests that excessive fusion with the reference image negatively impacts the model's ability to represent objects specified in the prompt. No unexpected or novel attention patterns were observed.

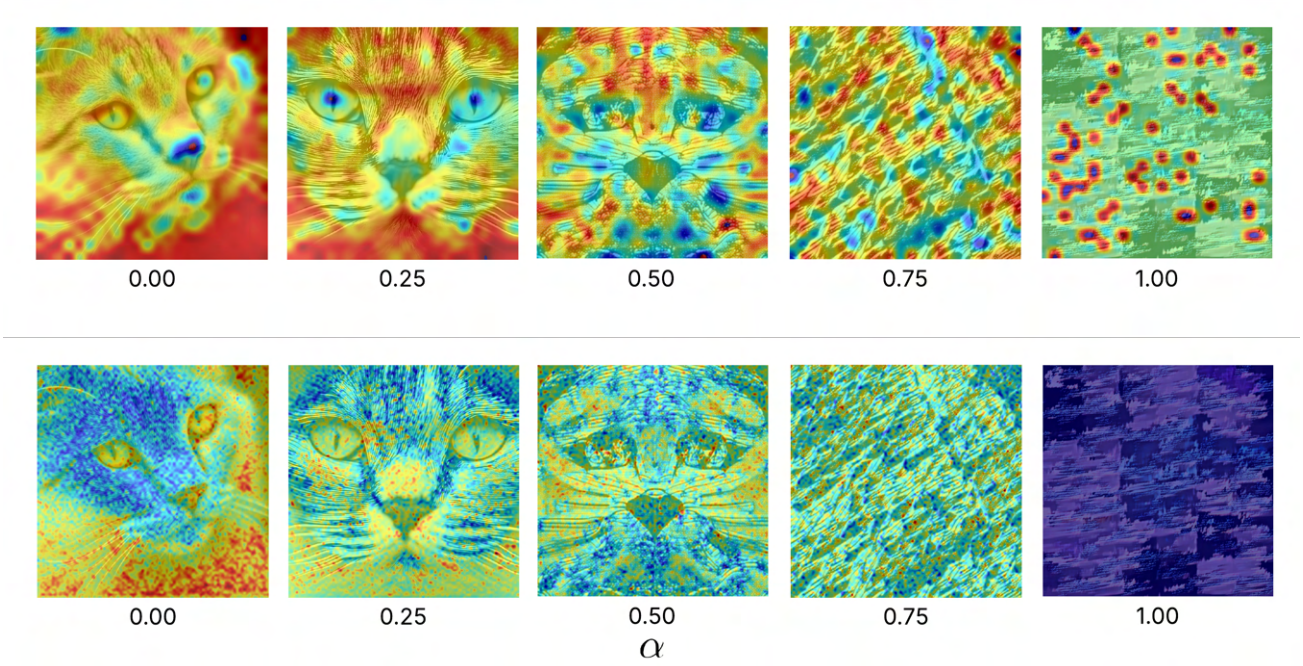

Figure 12: Cross-attention maps for output blocks 4 (top) and 11 (bottom) at different fusion weights ($\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$). For low fusion weights, attention is focused and object-aligned. For high α, the attention maps become increasingly random, mirroring distortions in the generated outputs.