Injecting Image Guidance into Text-Conditioned Diffusion Models at Inference

Agata Żywot Iason Skylitsis Thijmen Nijdam Zoe Tzifa-Kratira Derck Prinzhorn Konrad Szewczyk
supervised by Aritra Bhowmik

University of Amsterdam, Netherlands

The Style Transfer Gap

Effective visual transfer at inference time is a challenge. Models retrained for joint image-text conditioning are costly and often overlook key visual cues. Meanwhile, naive fusion in text-conditioned models like Stable Diffusion (SD) corrupts the output by disrupting the learned text-image alignment.

The Distribution Mismatch in Stable Diffusion (SD):

- SD is trained on CLIP text embeddings: $c \sim p_{\text{text}}(c)$.
- Naively fusing image-derived features $c' = f(I)$ at inference means $c' \not\sim p_{\text{text}}$.
- This distributional mismatch ($p_{\text{text}}(c) \neq p_{\text{image}}(c')$) leads to KL divergence and unstable generation:

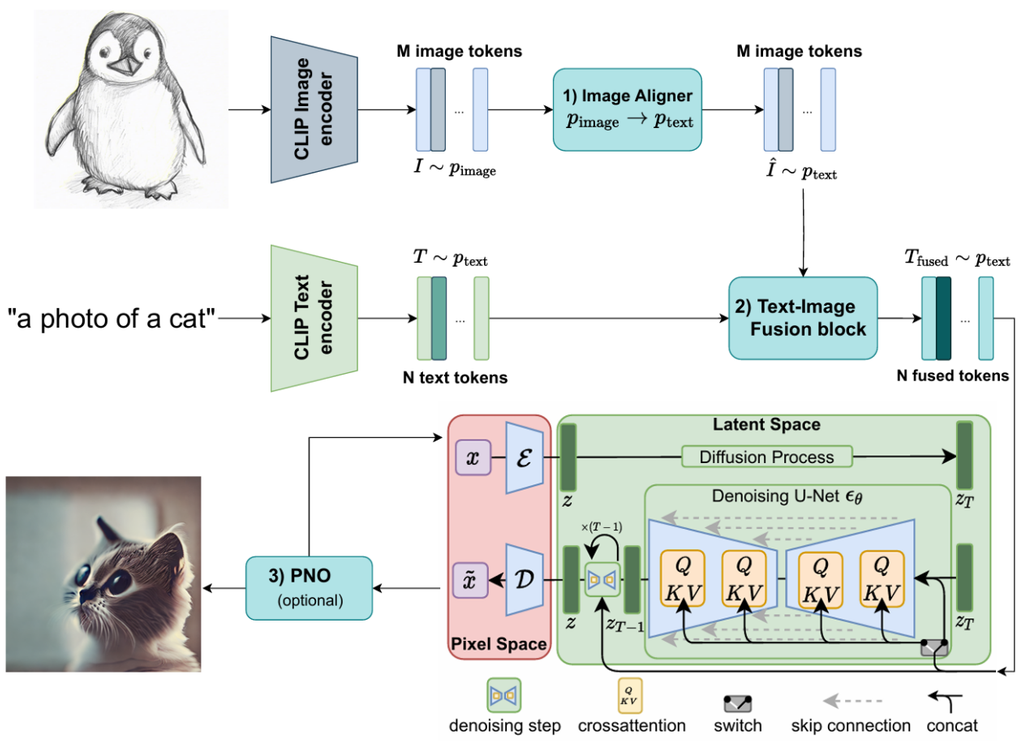
$$\Delta_{\text{KL}} = \text{KL}(p_\theta(x_0 \mid c') \| p_\theta(x_0 \mid c))$$

In short, naive image injection disrupts the learned text-image alignment, causing the model to ignore or misapply the reference.

We propose **Visual Concept Fusion (VCF)**: a pipeline to align and fuse image features with text conditioning without retraining the diffusion model, bridging this modality gap.

Methodology: Visual Concept Fusion

VCF Solution Overview



1. Aligning Modalities:

We first map the CLIP image features I into $\hat{I}$ that reside in the text embedding space. A small aligner network is trained on a subset of the COCO image-caption dataset to project $I \mapsto \hat{I}$. We have two training objectives:

- **Global (InfoNCE):** Matches mean token statistics.

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

- **Local (Cross-Attention Reconstruction):** $\hat{I}$ learns to reconstruct T.

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

2. Text-Image Fusion

Concatenation of text-image tokens: after the alignment step, the image tokens are appended to the text tokens for SD conditioning.

3. Prompt-Noise Optimization (PNO)

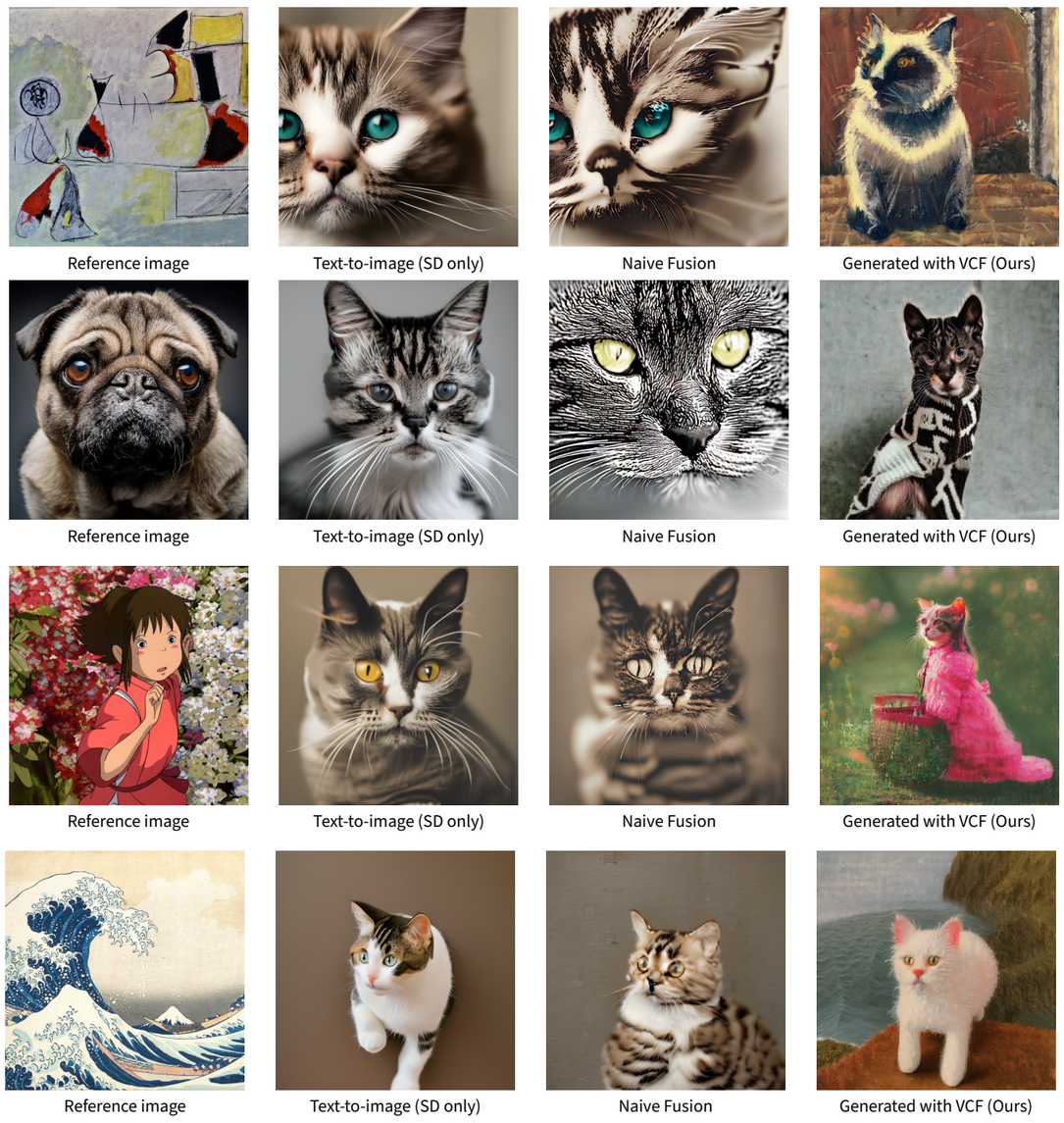
Inference-time refinement of the conditioning c' and initial noise x_T.

- Optimizes for semantic alignment between the generated image x_0 and the reference image in CLIP space.
- Performed for a small number of gradient steps (10-50) before sampling.

$$\min_{c', x_T} \mathcal{L}_{\text{PNO}} = \lambda \mathcal{L}_{\text{reg}}(x_T) - \cos(\text{CLIP}(x_0), \text{CLIP}(\text{reference\_image}))$$

Results

Qualitative results



Qualitative comparison of generation methods for the prompt: `a photo of a cat`.

Quantitative results

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

CLIP Score: Measures semantic alignment between the generated image and the text prompt (higher is better). **LPIPS:** Measures perceptual similarity between the generated image and the reference style image (lower is better).

Discussion

- We propose **Visual Concept Fusion**, a method to inject image guidance into text-to-image diffusion models at inference.
- VCF aligns image features with the text embedding space and fuses them via concatenation, preserving semantics and reference fidelity.
- **Future work**:
  - Explore improved alignment losses and mappers for $\hat{I} \sim p_{\text{text}}(T)$ to better align image features with the text distribution.
  - Add spatial control over the visual fusion process.
  - Thoroughly analyze PNO's interaction with VCF-aligned conditioning c' for optimal refinement.