Modifying Generative Distributions in Latent Diffusion Models to Improve Alignment with Desired Properties

Sven Sickert, Maria Gogolev, Niklas Penzel, Tim Büchner, Joachim Denzler

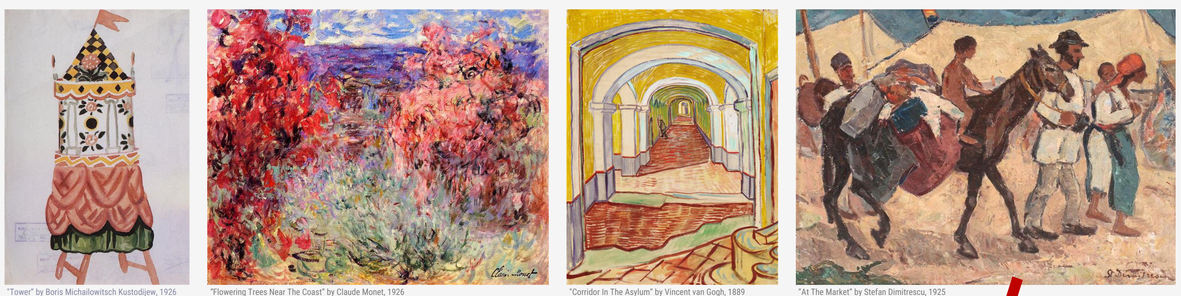
Computer Vision Group, Friedrich Schiller University, 07743 Jena, Germany

https://www.inf-cv.uni-jena.de

MVA2025

Diffusion Is Great, But Specificity Is Key!

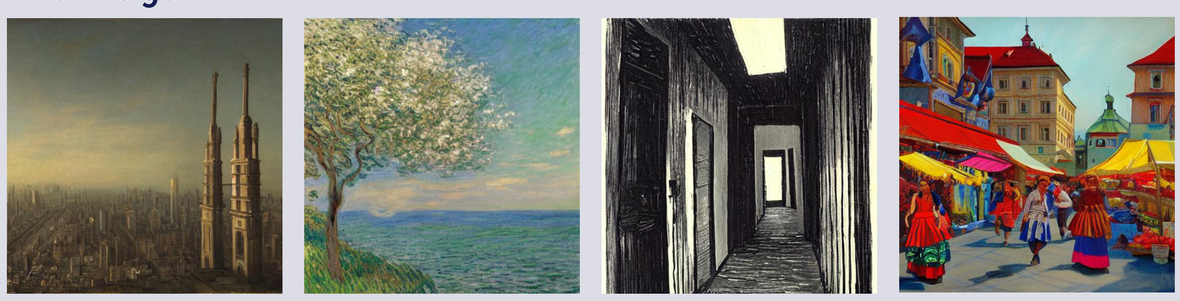
What we were looking for...



"Tower" by Boris Michailowitsch Kustodijew, 1926 | "Flowering Trees Near The Coast" by Claude Monet, 1926 | "Corridor In The Asylum" by Vincent van Gogh, 1889 | "At The Market" by Stefan Dimitrescu, 1925

Misalignment!

What we got...



Alignment Aspects

Distribution

- **Fréchet Inception Distance** (FID)[1] for image quality
- Similarity between distribution of image sets $\mathcal{I}_{real}$ and $\mathcal{I}_{gen}$
- Lower values indicate better alignment

High-Frequency Artifacts

- Generative models introduce distinct types of noise[2]
- Compute **power spectrum of residuals** (PR) for $\mathcal{I}_{real}$ and $\mathcal{I}_{gen}$
- Frobenius norm between both spectra: lower is better

Aesthetics

- Artworks exhibit f^{-2} distribution across frequency spectrum[3]
- Fit linear function in log-log space for average power spectra
- **Power spectrum slope distance** (PSD) between $\mathcal{I}_{real}$ and $\mathcal{I}_{gen}$

Modifying Distributions

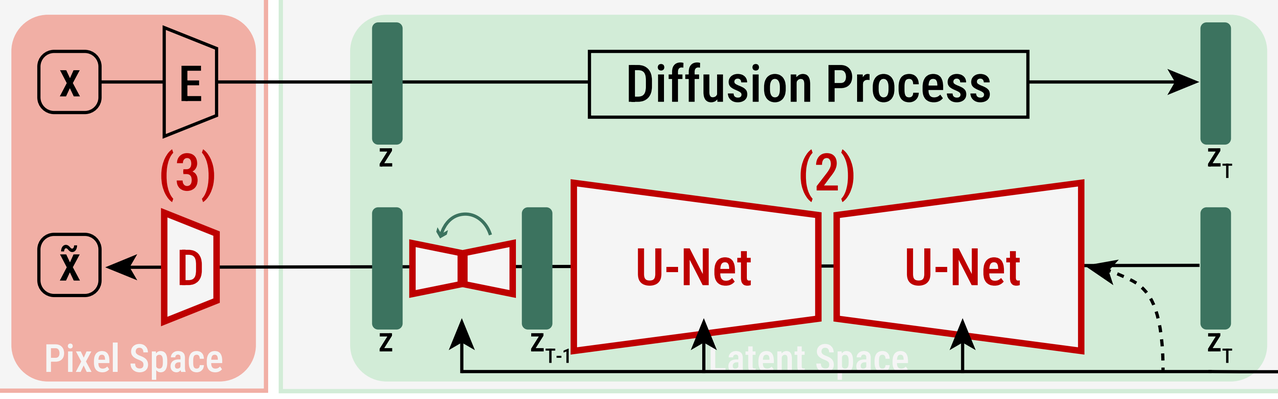
VAE Decoder Modification

- Modify learnable upscaling in VAE encoder to reduce artifacts
- $\mathcal{L}$ is weighted combination of three loss terms:
  - Focal Frequency Loss (FFL)[4]
  - Learned Perceptual Image Patch Similarity (LPIPS)[5]
  - Mean Squared Error

Fine-Tuning Denoising U-Net

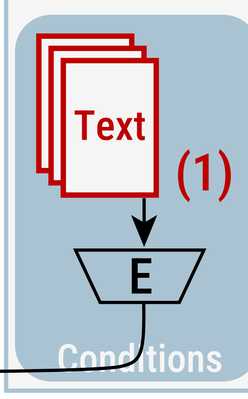
- Direct method to influence generative image distribution
- Three variants of how to include text prompts:

(i) Unconditional (ii) Conditional (iii) Mixed (20/80)



Text Prompt Optimization

- Modify condition by adding specific information (styles, attributes, ...)
- Prompt enrichment strategy in three levels and using CFG[6]



- Level 0: Art descriptions only
- Level 1: + Art style and artist
- Level 2: + Evoked emotions[7]

Evaluation & Results

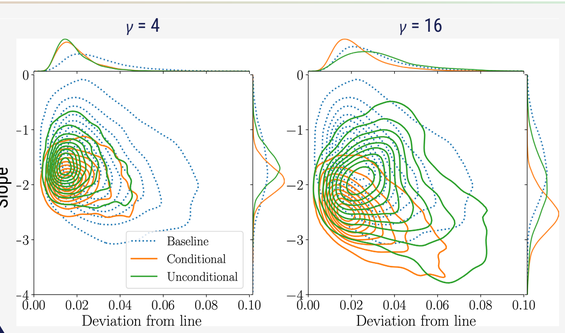
ArtEmis Dataset[7]

- 80K artworks from WikiArt
- Controlled shift from regular data

Experimental Setup

- SD2.1[8] base: VAE downsampling factor 8; 865M parameter U-Net; OpenCLIP ViT-h/14 text encoder
- 75K copyright free artworks; 75/25 train/test split; CFG scales (γ): 4, 8, 12, 16; Sampling five times

Rk	$\mathcal{L}$	Fine-Tuning	FID ↓	PR ↓	PSD ↓	FID ↓	PR ↓	PSD ↓
				γ = 4			γ = 16	
		Baseline	75.63±10.39	73.91±26.60	0.17±0.18	66.81±11.12	58.41±22.66	0.52±0.19
4		Unconditional	69.57±2.30	60.28±9.99	0.13±0.08	70.67±7.40	55.61±9.80	0.36±0.09
4		Conditional	72.75±21.43	69.06±12.32	0.29±0.13	34.94±4.22	64.09±8.16	1.00±0.05
4		Mixed (20/80)	63.47±14.49	73.62±11.25	0.25±0.15	41.56±2.84	74.39±8.40	0.88±0.06
8		Unconditional	68.49±9.43	53.34±8.57	0.13±0.09	56.91±7.61	36.01±8.82	0.35±0.12
8		Conditional	73.30±21.73	62.07±11.36	0.26±0.12	33.54±3.91	53.80±7.96	0.82±0.07
8		Mixed (20/80)	67.86±17.47	63.49±10.56	0.28±0.14	34.77±2.73	57.17±8.34	0.79±0.07
8	✓	Unconditional	68.34±9.09	54.80±8.48	0.14±0.14	57.45±7.67	37.54±8.80	0.39±0.12
8	✓	Conditional	72.73±21.70	62.31±12.16	0.37±0.12	33.90±3.85	54.68±8.20	0.89±0.07
8	✓	Mixed (20/80)	67.37±17.42	63.96±11.44	0.37±0.13	35.26±2.65	58.92±8.51	0.85±0.07



- Fine-tuning using LoRA[9] ranks 4 and 8
- Denoising U-Net fine-tuning is best
- $\mathcal{L}$ does not further improve results

10% Improvement

- Higher CFG has negative impact on alignment with respect to aesthetic (PSD)

Now we get...

γ	Prompt	FID ↓	PR ↓	PSD ↓
4	Level 0	83.17±9.98	81.20±27.67	0.18±0.19
4	Level 1	75.22±10.00	73.61±25.68	0.14±0.15
4	Level 2	77.23±9.95	72.30±26.21	0.14±0.16
16	Level 0	71.37±11.84	74.55±28.89	0.52±0.21
16	Level 1	66.19±10.98	55.65±21.97	0.47±0.19
16	Level 2	66.96±11.29	55.69±22.47	0.47±0.20

- Style and artist info improve alignment
- No further improvement when adding evoked emotions and human attribution

References

[1] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", NeurIPS, 2017
[2] Riccardo Corvi et al. "Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models", CVPR, 2023
[3] Michael Koch et al. "1/f2 Characteristics and Isotropy in the Fourier Power Spectra of Visual Art, Cartoons, Comics, Mangas, and Different Categories of Photographs", PLOS ONE 5.8, 2010
[4] Liming Jiang et al. "Focal Frequency Loss for Image Reconstruction and Synthesis", ICCV, 2021
[5] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations", ICLR, 2021
[6] Jonathan Ho and Tim Salimans. "Classifier-Free Diffusion Guidance", NeurIPS, 2021
[7] Panos Achlioptas et al. "ArtEmis: Affective Language for Visual Art", CVPR, 2021
[8] Robin Rombach et al. "High-Resolution Image Synthesis With Latent Diffusion Models", CVPR, 2022
[9] Edward J Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models", ICLR, 2022

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Computer Vision Group