RB-Modulation: Training-Free Stylization using Reference-Based Modulation

Litu Rout

Based on joint work with: Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Wen-Sheng Chu, Constantine Caramanis, and Sanjay Shakkottai

The University of Texas at Austin, Google Research, Google DeepMind

The University of Texas at Austin Machine Learning Laboratory

The University of Texas at Austin Chandra Department of Electrical and Computer Engineering Cockrell School of Engineering

Google

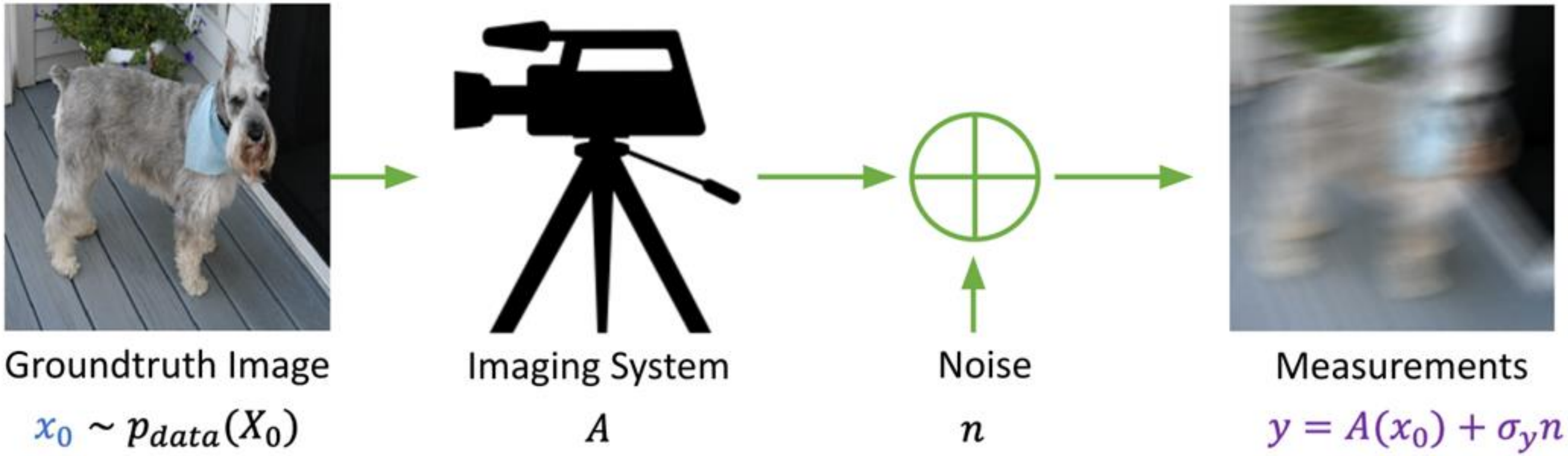
Google DeepMind

IFML

EDGE Institute

WNCG

Inverse Problem

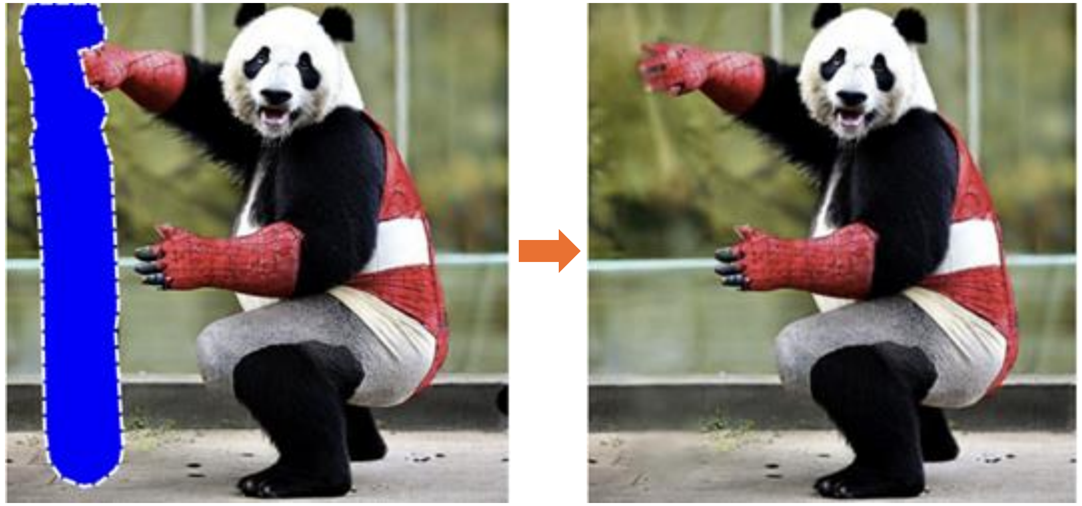


Problem: Reconstruct ground truth image x_0 from noisy measurements y

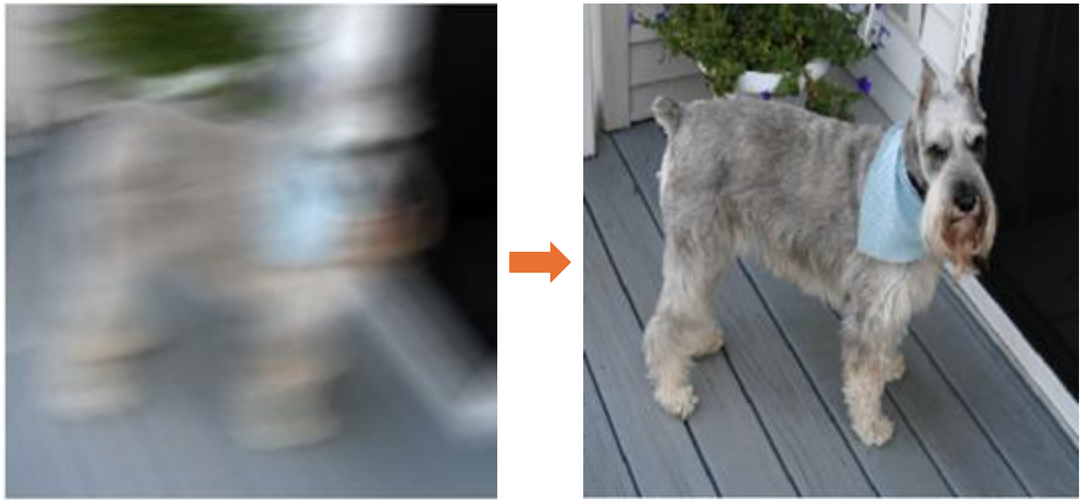
Challenge: Problem is ill-posed, that is infinitely many solutions x_0 exist

Approach: Use prior knowledge $p(x_0)$ of how the image should look like

General Inverse Problems



Free-form inpainting



Motion Deblur

Super-resolution (4X)

Gaussian Deblur

Stylization as Inverse Problem

Text-to-image generation

"A guitar"

Input

Output

Personalized text-to-image generation: stylization

"A guitar"

style

text

Output

Content-Style Composition as Inverse Problem

Text-to-image generation

"A dog"

Input

Output

Personalized text-to-image generation: content-style composition

"A dog"

Input

Output

Training-Based Approaches

DreamBooth

- Fully fine-tune the weights of the pre-trained model
- Requires ~4 images per reference subject or style
- Expensive for large-scale text-to-image models
- Catastrophic forgetting due to modified pre-trained weights

Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

IP-Adapters

- Train newly added cross/self-attention layers
- Requires ~4 images per reference subject or style
- Less expensive for large-scale text-to-image models
- Retains original contents via pre-trained weights

Ye, Hu, et al. "IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models." arXiv preprint arXiv:2308.06721 (2023).

LoRA

- Train additive residual weights of pre-trained model
- Requires ~4 images per reference subject or style
- Cost effective for large-scale text-to-image models
- Retains original contents via pre-trained weights

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Training-Free Approaches

StyleAligned

- Requires a text-prompt describing reference style image
- Merges keys and values after instance normalization
- Uses DDIM inversion to extract style features from real image
- Leaks content from reference style image

Hertz, Amir, et al. "Style aligned image generation via shared attention." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Swapping Self-Attention

- Extracts Keys and Values from reference style image
- Swaps original Keys and Values with those of reference
- Uses DDIM inversion to extract style features from real image
- Leaks content from reference style image

Jeong, Jaeseok, et al. "Visual Style Prompting with Swapping Self-Attention." arXiv preprint arXiv:2402.12974 (2024).

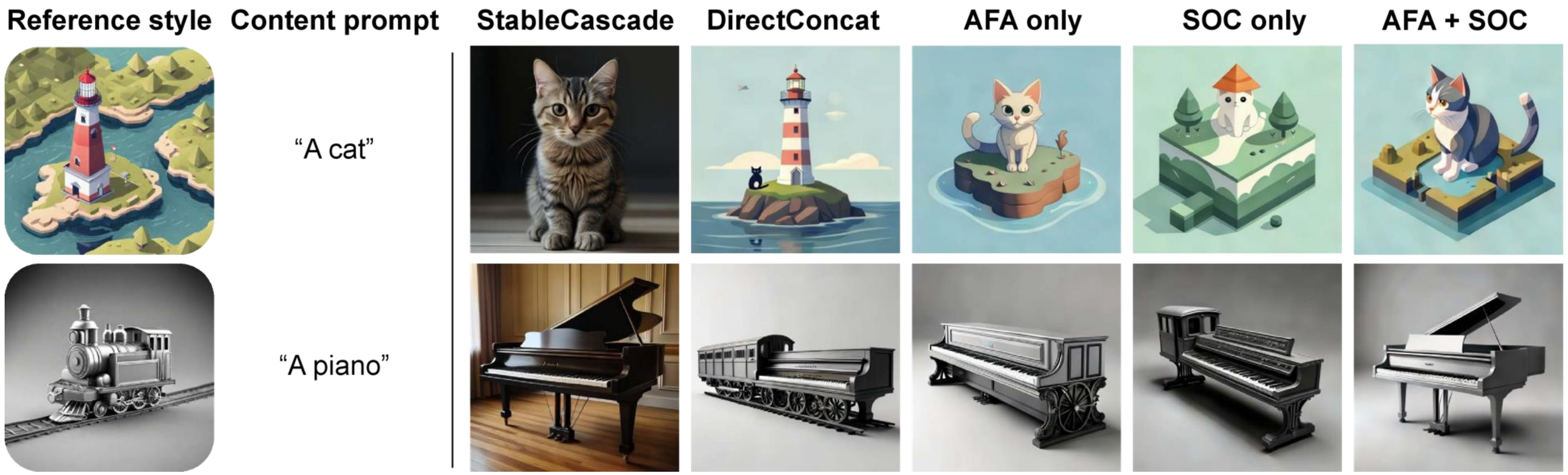
InstantStyle

- Injects style features through a specific layer of an IP-Adapter
- Avoids DDIM inversion and memory intensive reverse SDEs
- Identifying a layer is a complex task and may not generalize
- Limited diversity due to ControlNet and leaks content

Wang, Haofan, et al. "Instantstyle: Free lunch towards style-preserving in text-to-image generation." arXiv preprint arXiv:2404.02733 (2024).

Our Approach: Modulation of Reverse Diffusion Processes

- RB-Modulation has two key elements
  - Stochastic Optimal Controller (SOC) and Attention Feature Aggregation (AFA)
  - SOC: An optimal control formulation-based sampler, implemented as a test-time optimization algorithm
  - AFA: Personalize the score and disentangle content-style from the reference images through an alternate cross-attention processor



Background: Diffusion Models

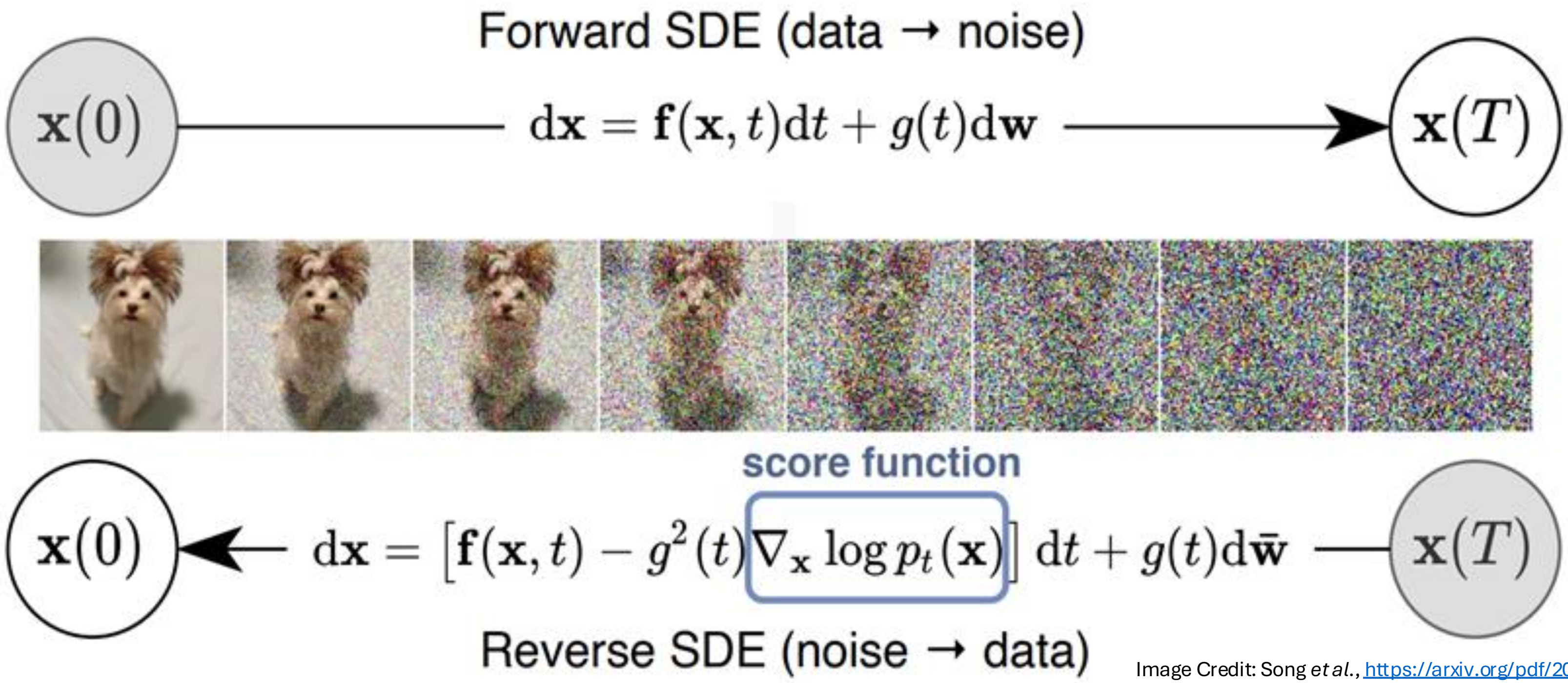


Image Credit: Song *et al.*, https://arxiv.org/pdf/2011.13456.pdf, ICLR'21

- Goal: Design a Markov process-based sampler (a transition kernel) such that stationary distribution samples images
- Approach: Learn annealed score that is affine in the conditional expectation of $X(0)$ (clean image) given X(t) (noisy image) by Tweedie's Formula

References: Deep Unsupervised Learning using Diffusion (Sohl-Dickstein et al.' 2015); Score-based Generative Models (Song & Ermon' 2019); Diffusion Probabilistic Models (Ho et al.'2020); Score-based Generative Models through SDEs (Song et al.' 2021)

Background: Inverse Problems as Posterior Sampling

Problem: Sample from $p_0(x_0|y)$ instead of $p(x_0)$

$$dX_t = (-X_t - 2\nabla \log p_t(X_t|y))\, dt + \sqrt{2} d\bar{W}_t, t = T, \cdots, 0$$

Unknown

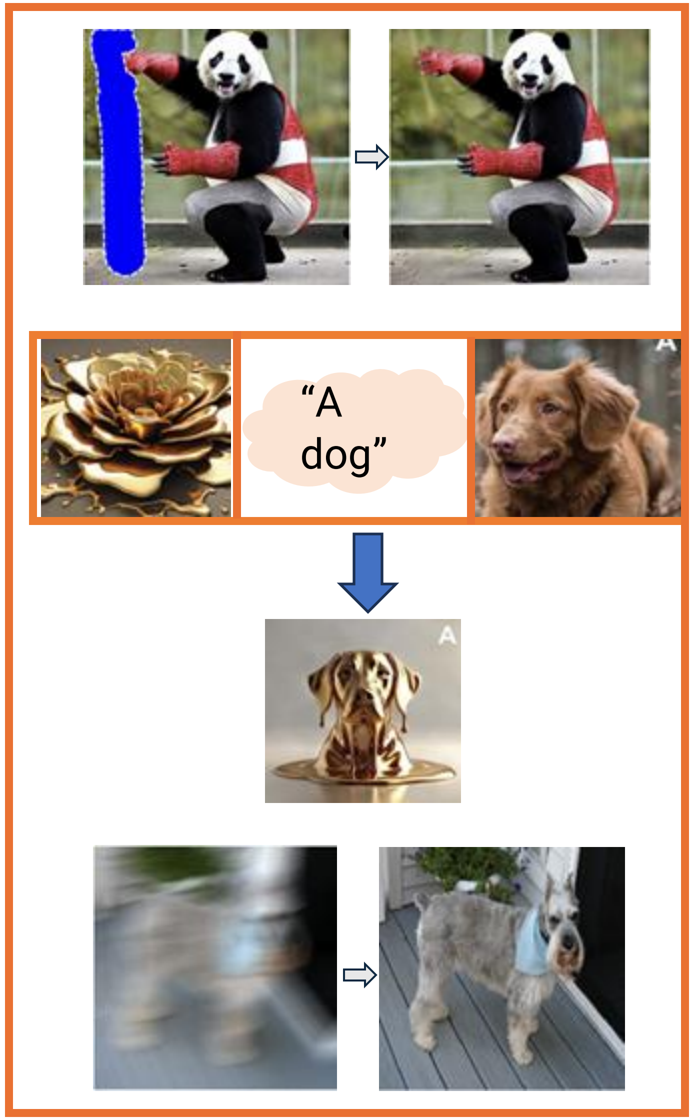
Bayes rule:

$$\log p_t(x_t|y) = \log p_t(y|x_t) + \log p_t(x_t) - \log p_t(y)$$

$$dX_t = (-X_t - 2\nabla \log p_t(y|X_t) - 2\nabla \log p_t(X_t))\, dt + \sqrt{2} d\bar{W}_t$$

Unknown Known: $\nabla \log p_t(X_t) \approx s_\theta(X_t, t)$

How well can we approximate $\nabla \log p_t(y|x_t)$?



DDRM: Kawar et al. https://arxiv.org/pdf/2201.11793.pdf, NeurIPS’21; DPS: Chung *et al.*, https://arxiv.org/pdf/2209.14687.pdf, ICLR’23
See also Delbracio & Milanfar, https://openreview.net/pdf?id=VmyFF5lL3F, TMLR’23 for an alternate formulation of supervised inverse problems.

Our Approach: Posterior Sampling using RB-Modulation

Goal: Interpret posterior sampling as a stochastic control problem

Posterior Sampling: Sample $p_0(\cdot\,|y)$ using conditional reverse SDE

$$dX_t = (-X_t - 2\nabla \log p_t(y|X_t) - 2\nabla \log p_t(X_t))\, dt + \sqrt{2} dW_t, \qquad t = T, \cdots, 0$$

- Prior approaches[1,2,3] rely on first- or second-order Taylor's approximation
- We replace $\nabla \log p_t(y|X_t)$ with a controller $u(X_t, t, y)$ and solve a stochastic optimal control problem

$$\min_{u \in U} E\left[\int_T^0 [\|u(X_t^u, t, y)\|^2 + f(X_t^u, t)]\, dt + g(X_0^u, y)\right]$$

where $dX_t^u = (-X_t^u + u(X_t^u, t, y) - 2\nabla \log p_t(X_t^u))dt + \sigma(t) dW_t\,, \qquad X_T^u \sim p_T$

[1] Chung, Hyungjen et. al. "Diffusino Posterio Sampling for Noisy Inverse Problems", Internation Conference on Learning Representations (2023).
[2] Rout, Litu, et. al. "Solving linear inverse problems provably via posterior sampling with latent diffusion models." Advances in Neural Information Processing Systems 36 (2024).
[3] Rout, Litu, et. al. "Beyond first-order tweedie: Solving inverse problems using latent diffusion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

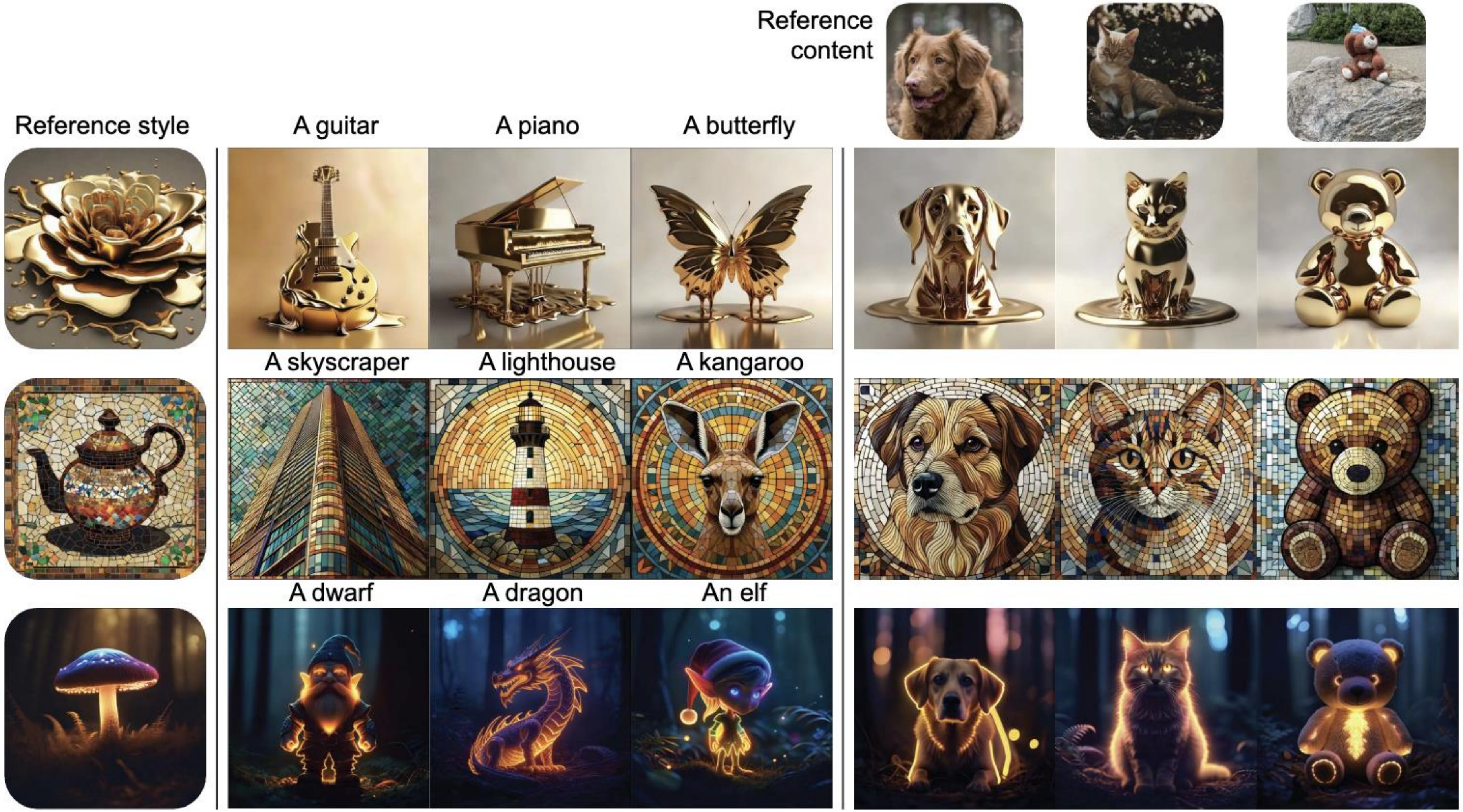
RB-Modulation: Training vs Test-time Optimization

- Training-time optimization (DreamBooth, LoRA, IP-Adapter)
  - Approximately 10s of samples per conditioning (style/content)
    - Single sample leads to catastrophic forgetting
  - Gradient computed with respect to weights of base model
  - LoRA finetuning takes ~20 min per style (40 min for content-style)
  - Full finetuning takes hours

- Test-time optimization (DPS, PSLD, P2L, STSL)
  - Single sample suffices (no catastrophic forgetting)
  - Gradient computed with respect to input to base model
  - Takes ~10 min for PSLD (1B), ~20 min for P2L(1B) (longer for Flux-12B)

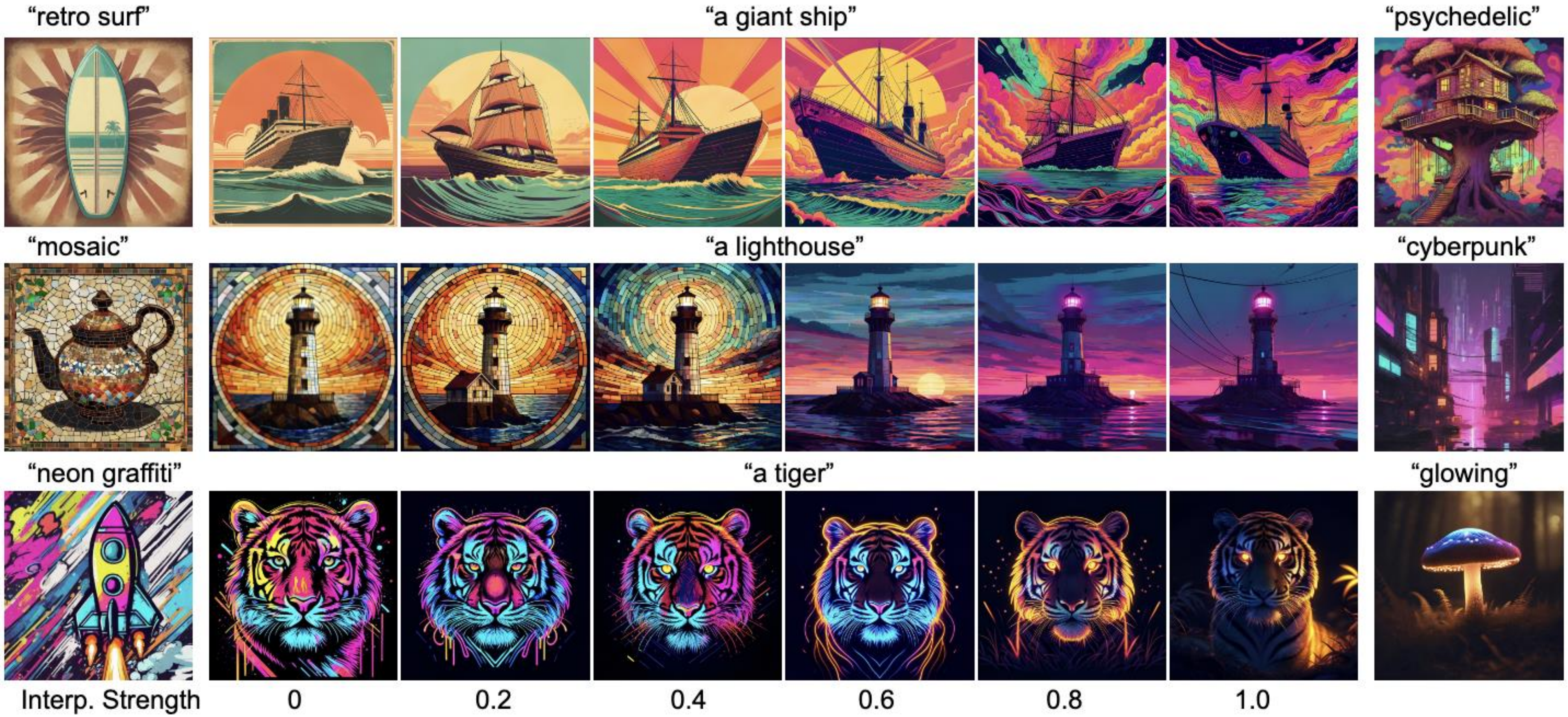
- Proximal test-time optimization (RB-Modulation)
  - Takes 40 sec using StableCascade (4B)

Experiments: Training-free Personalization



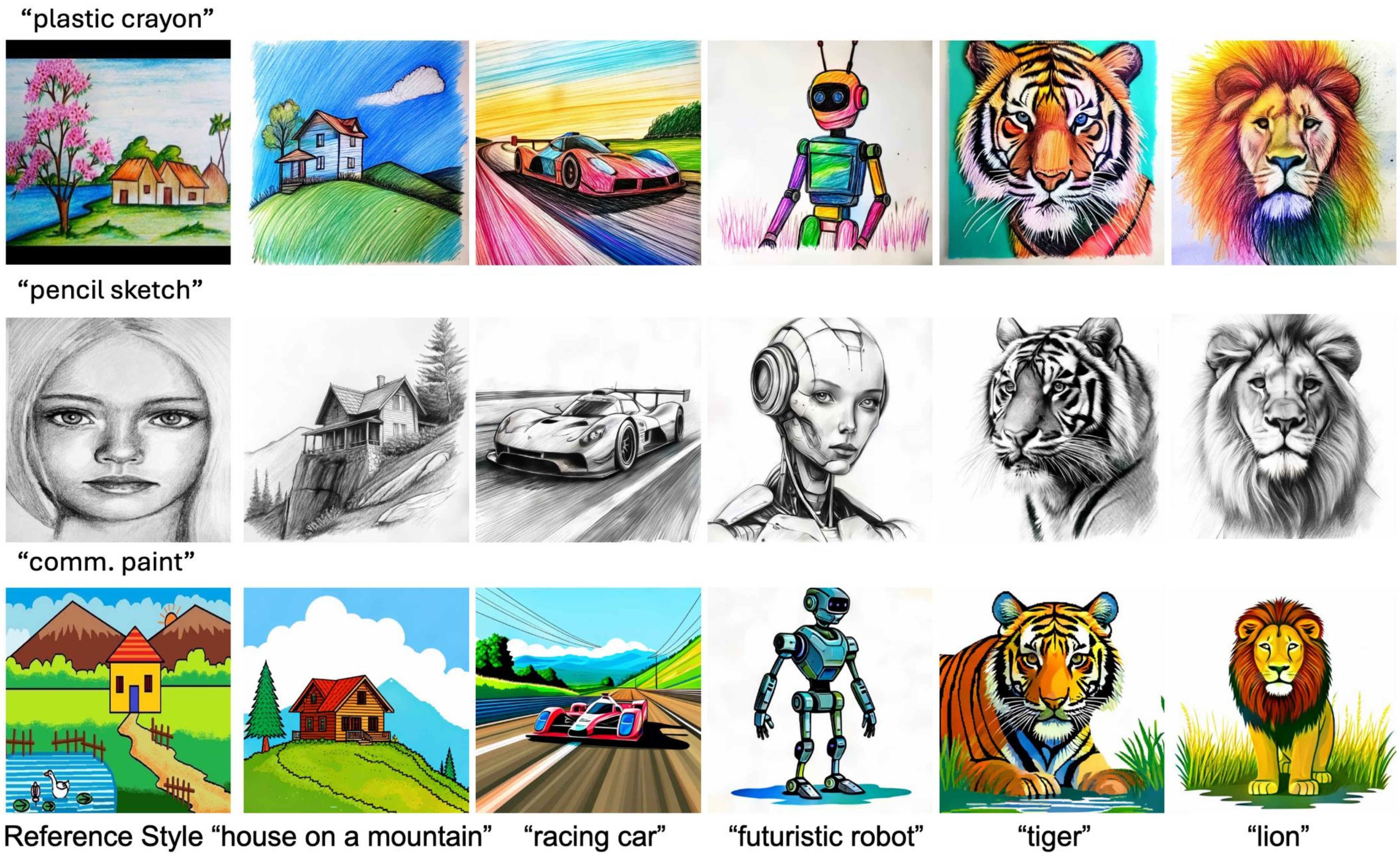
RB-Modulation as a plug-and-play solution for (a) stylization (b) content-style composition

Novel Style Synthesis: Interpolating Reference Styles



Training based methods cannot interpolate novel styles due to lack of prior examples

Stylization: Hand Drawn Reference Images



Please see: https://openreview.net/forum?id=bnINPG5A32 for reference image credits

Project page: https://rb-modulation.github.io/

RB-Modulation: Training-Free Stylization using Reference-Based Modulation

Litu Rout[1,2] Yujia Chen[2] Nataniel Ruiz[2]
Abhishek Kumar[3] Constantine Caramanis[1] Sanjay Shakkottai[1] Wen-Sheng Chu[2]

[1] University of Texas, Austin [2] Google [3] Google DeepMind

ICLR 2025 (Oral: 1.8% acceptance ratio)

[Paper] [OpenReview] [ArXiv] [Code] [Demo]