Get in Touch!

Project Page

ETH zürich

ICML International Conference On Machine Learning

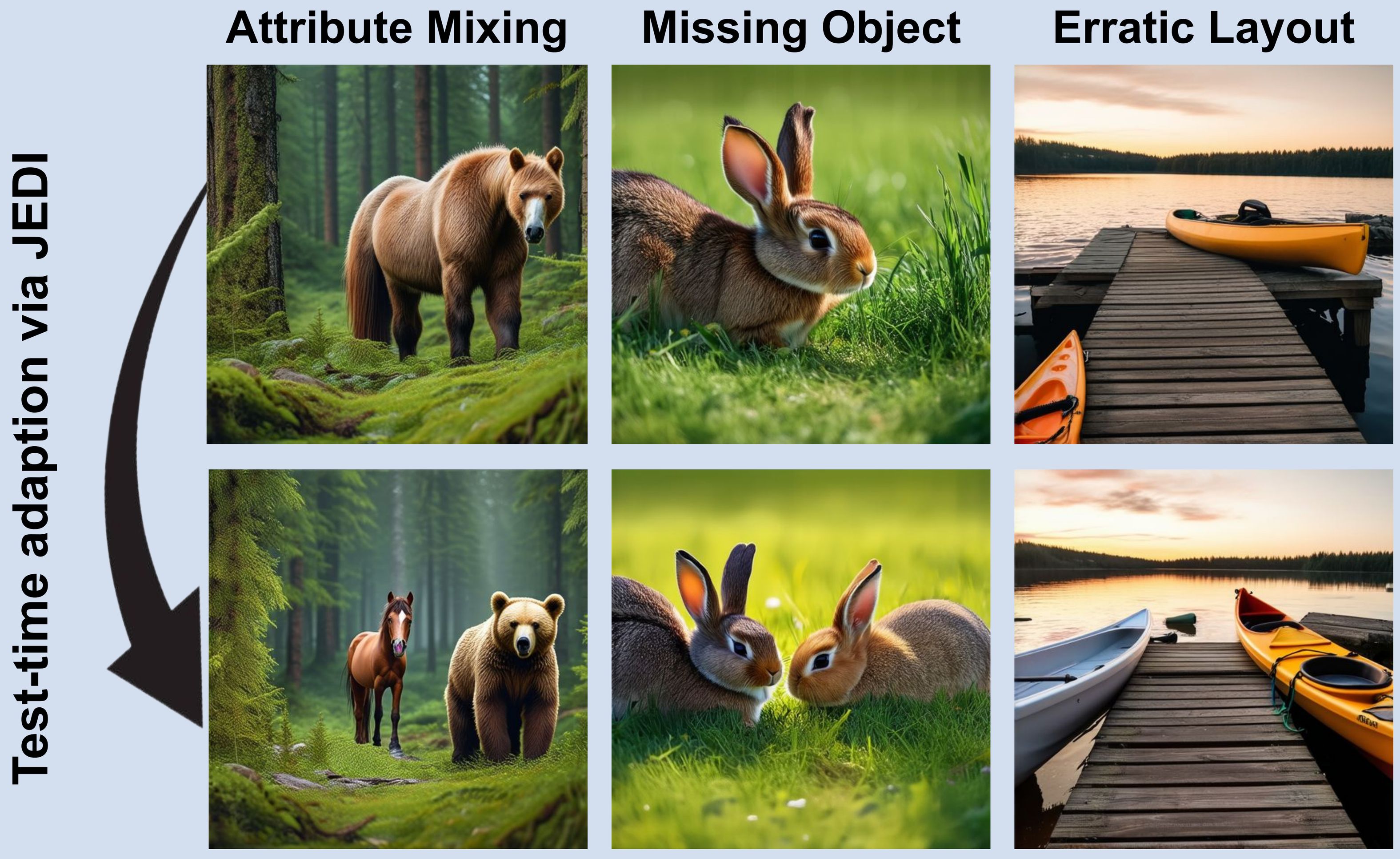
JEDI: The Force of Jensen-Shannon Divergence in Disentangling Diffusion Models

Eric Tillmann Bill, Enis Simsar, Thomas Hofmann

1 Introduction

JEDI is a **model-agnostic**, **training-free** method for improving semantic alignment **at test-time** in text-to-image models.

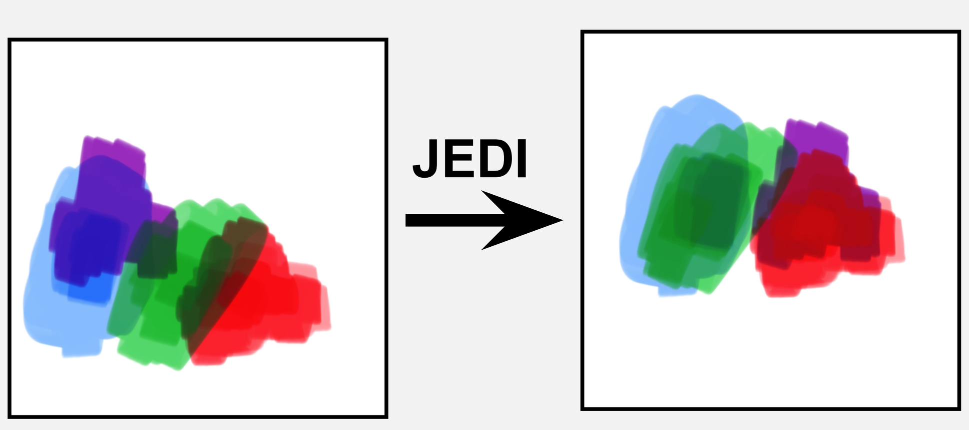
- ➔ Stays close to original base generation (no stilistic bias)
- ➔ Very efficient due to the use of adversarial optimization



*"A **horse** and a **bear** [...]."* *"A **rabbit** and a **hare** [...]."* *"A **canoe** and a **kayak** [...]."*

2 Why does this happen?

Prompt conditioning via attention can lead to overlapping maps and semantic confusion:



"A ***blue cat*** and a ***red dog***"

Target of JEDI:

- ➔ Subject Separation
- ➔ Intra Subject Coherence
- ➔ Equal support

3 JEDI Objective

For each target, we have an additive component to minimize. Let S denote the set of subjects, and P_s the set of attention maps for each subject $s \in S$:

1. **Intra-group Coherence:** Encourage similarity within each group:

$$\frac{1}{|S|}\sum_{s \in S} \hat{D}_{\text{JS}}(P_s).$$

2. **Inter-group Separation:** Encourage subject-wise distinction by minimizing:

$$1 - \hat{D}_{\text{JS}}(M),$$

where

$$M = \{\mathbf{m}_s = \frac{1}{|P_s|}\sum_{\mathbf{p} \in P_s} \mathbf{p} \mid s \in S\}$$

are mixture distributions for each subject.

3. **Diversity Regularization:** Promote spatial spread by maximizing entropy:

$$\lambda \cdot \frac{1}{|S|}\sum_{s \in S}\left(1 - \hat{H}(\vec{m}_s)\right).$$

4 Latent Optimization

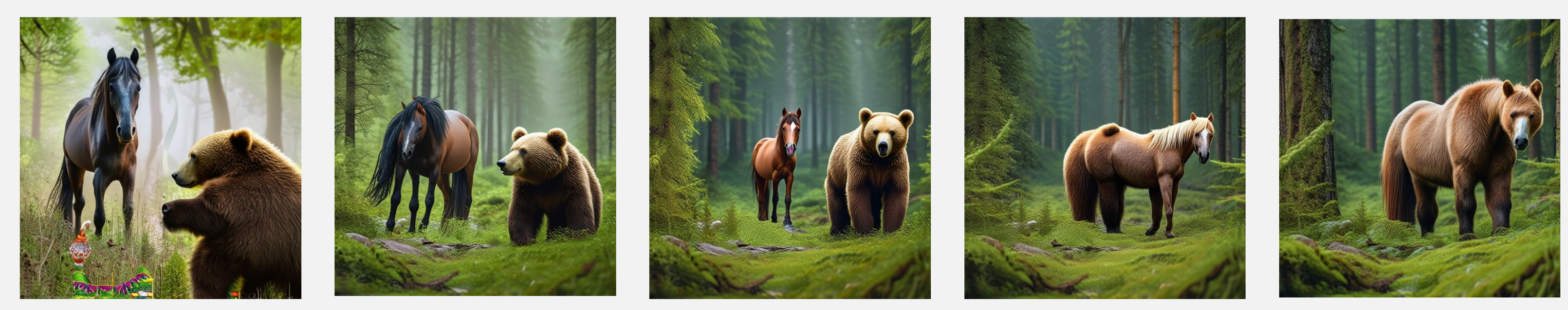
JEDI steers the iterative denoising process by updating the latent image $\boldsymbol{x}_t$:

- ➔ Only applied during the first K=18 steps.
- ➔ Minimal updates via signed gradients.

Sampling Process with JEDI Optimization

1: $x_0 \sim p_{\text{prior}}$
2: **for** $t = 0$ to $T-1$ **do**
3: **if** $t \leq K$ **then**
4: $\_, A_t \leftarrow$ Model$(\boldsymbol{x}_t, \boldsymbol{c})$
5: $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - \alpha \cdot$ sign$(\nabla_{\boldsymbol{x}_t}$JEDI$(A_t, \boldsymbol{c}))$
6: $\boldsymbol{x}_{t+1}, \_ \leftarrow$ Model$(\boldsymbol{x}_t, \boldsymbol{c})$
7: **return** x_T

- ➔ Signed gradients allow for finer control of stilistic drift via α:



1e-2 | 5e-3 | 3e-3 | 1e-3 | Base

5 Qualitative Results

Stable Diffusion 3.5

*"A **dachshund** and a **corgi** sitting [...]"*

*"A **street bike** and a **dirt bike** [...]"*

*"A **sparrow** and a **finch** perched on [...]"*

*"A **black cat**, an **orange cat**, and a **white cat** lounging on a windowsill"*

*"A **horse** , a **bear**, and a **moose** in forest clearing"*

*"A **Labrador**, a **Golden Retriever** and a **German Shepherd playing** [...]"*

Stable Diffusion 1.5

*"A **violin** and a **viola** on a wooden stage under soft spotlights"*

*"A **sparrow** and a **finch** perched on a blossoming branch"*

*"A **canoe** and a **kayak** tied to a wooden dock at dawn"*

*"A **street bike** and a **dirt bike** leaning against a garage wall"*

LoRACLR (Simsar et al., 2024)

*"**<LeBron>** and **<Margot>** at [...]"*

*"**<Pitt>** and **<Taylor>** in Venice [...]"*

*"**<Messi>** and **<Taylor>** in front of [...]"*

6 Conclusion

- ➔ **JEDI improves subject disentanglement** at test time without retraining or external models.
- ➔ **Efficient and lightweight:** Only 18 optimization steps with minimal stilistic drift from base model.
- ➔ **Model-agnostic:** Works across Stable Diffusion 1.5, 3.5, and LoRACLR.
- ➔ **Built-in disentanglement score:** JEDI provides a free measure of entanglement (alternative to CLIP).