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

Eric Tillmann Bill [1] **Enis Simsar** [1] **Thomas Hofmann** [1]

Abstract

We introduce JEDI, a test-time adaptation method that enhances subject separation and compositional alignment in diffusion models without requiring retraining or external supervision. JEDI operates by minimizing semantic entanglement in attention maps using a novel Jensen-Shannon divergence based objective. To improve efficiency, we leverage adversarial optimization, reducing the number of updating steps required. JEDI is model-agnostic and applicable to architectures such as Stable Diffusion 1.5 and 3.5, consistently improving prompt alignment and disentanglement in complex scenes. Additionally, JEDI provides a lightweight, CLIP-free disentanglement score derived from internal attention distributions, offering a principled benchmark for compositional alignment under test-time conditions. Code and results are available at ericbill21.github.io/JEDI/.

1. Introduction

Diffusion models have achieved remarkable success in generative modeling, particularly in the domain of image synthesis (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2022). Among these, text-to-image (T2I) diffusion models (Esser et al., 2024; Ramesh et al., 2022; Podell et al., 2023) stand out for their ability to generate diverse and high-quality images conditioned on natural language prompts.

However, despite these advances, current T2I models often struggle with compositional prompts that involve multiple objects or intricate spatial relationships. For example, when given a prompt like *"A horse and a bear in a forest,"* models from the Stable Diffusion family may produce semantically inconsistent outputs: one subject may be omitted (missing object), features from both animals may blend together into a single entity (attribute mixing), or the spatial arrangement may appear incoherent, refer to Figure 1.

Figure 1. **JEDI enables test-time subject disentanglement.** For the prompt *"A **horse** and a **bear** in a forest"*, JEDI reduces attribute mixing and improves subject separation in Stable Diffusion 3.5.

Such failures are especially problematic at *test time*, where retraining or fine-tuning is often infeasible. To address these limitations, a range of test-time adaptation techniques have been proposed, which broadly fall into two categories: 1.) *Latent Optimization Methods*, which adjust the latent representations during sampling to better align with the prompt (Meral et al., 2024; Chefer et al., 2023; Wei et al., 2024). 2.) *Concept-Based Methods* which rely on external structural cues such as layouts or segmentation maps to guide the generation process (Kwon et al., 2024; Binyamin et al., 2024; Liu et al., 2022).

While concept-based methods provide structural guidance, they often require additional models and can alter the underlying generative distribution. In contrast, latent optimization methods operate entirely within the model's architecture and offer a lightweight, model-preserving alternative for test-time adaptation. In this work, we focus on latent optimization and introduce a novel, training-free test-time adaptation method called **JEDI** (**Je**nsen-Shannon Divergence for **D**isentanglement at **I**nference). By framing compositional entanglement as a probabilistic alignment problem, we propose a new divergence-based objective tailored for attention distributions. Our main contributions are as follows:

i) We introduce a novel objective based on Jensen-Shannon divergence to minimize semantic entanglement in attention maps at test-time, providing a probabilistically grounded alternative to cosine similarity.

[1]Department of Computer Science, ETH Zurich, Switzerland. Correspondence to: Eric Tillmann Bill <erbill@ethz.ch>.

PUT 2025: 2nd Workshop on Test-Time Adaptation at the 42nd *International Conference on Machine Learning*, Vancouver, Canada.
Copyright 2025 by the author(s).

ii) By leveraging adversarial optimization techniques, we reduce the number of optimization steps, making JEDI lightweight and efficient for real-world use.

iii) JEDI demonstrates strong performance across multiple architectures, including Stable Diffusion 1.5, LoRA-CLR, and Stable Diffusion 3.5, consistently improving alignment with complex prompts.

iv) JEDI provides an entanglement score derived from internal attention maps, enabling compositional evaluation without relying on external models such as CLIP.

2. Latent Optimization

Latent alignment methods steer the iterative denoising process in diffusion models by modifying the latent image during sampling. These methods often leverage model's internal attention maps, which act as soft spatial probability distributions, indicating how strongly each token (e.g., *"horse", "bear"*) influences different image regions.

At each timestep t during inference, we retrieve the updated latent $\boldsymbol{x}_{t+1}$ and the internal attention maps A_{t+1}:

$$\boldsymbol{x}_{t+1}, A_{t+1} = \text{model}(\boldsymbol{x}_t, t).$$

We then perform a test-time update of $\boldsymbol{x}_t$ by minimizing a disentanglement loss defined over A_{t+1}:

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - \alpha \nabla_{\boldsymbol{x}_t} \text{score}(A_t),$$

where $\text{score}(A_t)$ penalizes overlap between attention maps of different entities. This encourages spatial disentanglement and mitigates attribute mixing. See Algorithm 1 in Appendix C for a pseudo-code implementation.

Probabilistic View. Although attention maps are often treated as similarity scores, the use of the softmax function ensures that they are normalized and can instead be interpreted as discrete probability distributions.

Prior work (Meral et al., 2024; Wei et al., 2024) overlooked this probabilistic structure, commonly relying on cosine similarity as a measure for alignment, despite its lack of probabilistic grounding. An exception is Chefer et al. (2023), which considers attention probabilities but focuses only on maximizing individual token activation without accounting for inter-token competition.

In contrast, throughout this work we interpret attention maps as discrete probability distributions and optimize them accordingly. Our objective is to encourage unimodal, spatially localized, and non-overlapping attention for each subject in the prompt. This enables more faithful and disentangled representations, all achieved via test-time adaptation.

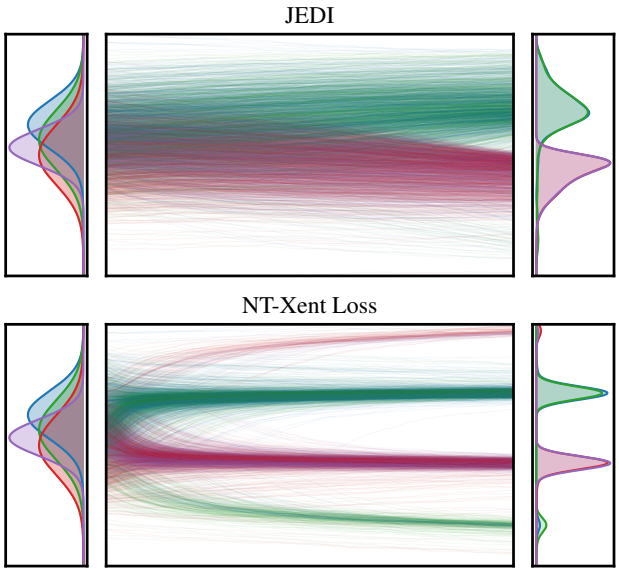

Figure 2. **Optimization Evolution of JEDI and NT-Xent.** Synthetic example with four overlapping distributions: **blue**/**green** correspond to one subject, **red**/**purple** to another. Overlaps of **blue** and **green** form teal, while **red** and **purple** form pink. JEDI preserves coherent group structure, while NT-Xent collapses modes.

3. Methodology

We propose JEDI (**Je**nsen-Shannon Divergence for **D**isentanglement at **I**nference), a test-time adaptation method that improves subject separation in diffusion models by adjusting latent representations using attention statistics. Our objective combines Jensen-Shannon divergence (JSD) and Shannon Entropy to encourage intra-group coherence, inter-group separation, and spatial diversity.

Jensen-Shannon Divergence. To measure the overlap among a set $P = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\} \subset \mathbb{R}^d$ of spatial attention distributions, we use the Jensen-Shannon divergence:

$$D_{\text{JS}}(P) = \frac{1}{|P|} \sum_{\boldsymbol{p} \in P} D_{\text{KL}}(\boldsymbol{p} \parallel \boldsymbol{m}), \quad \boldsymbol{m} = \frac{1}{|P|} \sum_{\boldsymbol{p} \in P} \boldsymbol{p},$$

where D_{KL} is the Kullback-Leibler divergence, defined as:

$$D_{\text{KL}}(\boldsymbol{p} \parallel \boldsymbol{q}) = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}.$$

Since $D_{\text{JS}}(P) \in [0, \log n]$ is bounded, we normalize it by dividing by $\log n$, yielding $\hat{D}_{\text{JS}}(P) \in [0, 1]$, which enables comparison across groups of different sizes. For a formal proof of this bound, see Lemma B.1 in Appendix B.

Shannon Entropy. To control the sharpness of individual attention maps, we incorporate the Shannon entropy:

$$H(\boldsymbol{p}) = -\sum_{i=1}^{d} p_i \log p_i.$$

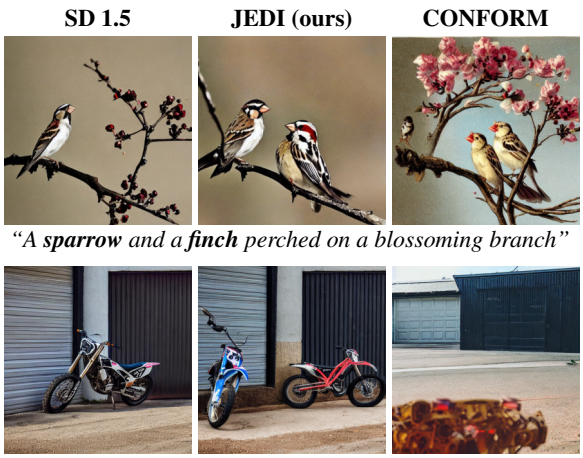

Figure 3. **Comparison of JEDI and CONFORM on Stable Diffusion 1.5.** Each image triplet was generated under identical conditions. For more details and examples, refer to Appendix G.

Entropy ranges from 0 (single peak) to $\log d$ (uniform). We normalize it by $\log d$, yielding $\hat{H}(\boldsymbol{p}) \in [0, 1]$, which allows scale-independent balancing, where high entropy indicates spatial spread; low entropy implies tight localization. For a proof of the bound refer to Lemma B.2 in Appendix B.

Objective Formulation. Let S denote the set of subjects in the text prompt, and let P_s be the set of attention maps associated with subject $s \in S$. The total loss consists of three additive components:

1. **Intra-group Coherence:** Encourages attention maps within each group (e.g., between an attribute and its subject) to be similar by minimizing their JSD:

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

2. **Inter-group Separation:** For each subject s, we compute its mixture distribution: $\boldsymbol{m}_s = \frac{1}{|P_s|} \sum_{\boldsymbol{p} \in P_s} \boldsymbol{p}$. Let $M = \{\boldsymbol{m}_s \mid s \in S\}$. To encourage separation between subjects, we maximize the divergence between these mixtures, by minimizing:

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

3. **Diversity Regularization:** To avoid overly sharp or degenerate maps, we encourage spatial spread by maximizing the normalized entropy of each mixture distribution. To this end, we minimize:

$$\lambda \cdot \frac{1}{|S|} \sum_{s \in S} \Big(1 - \hat{H}(\boldsymbol{m}_s)\Big),$$

where λ is a hyperparameter controlling the strength of the regularization term. In practice, we set $\lambda = 0.01$.

We provide further analysis of the effect of each component in the form of an ablation study in Appendix D.

Update Formulation. To efficiently update the latent representation, we follow the Fast Gradient Sign Method (Goodfellow et al., 2014) and perform:

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - \alpha \cdot \text{sign}\left(\nabla_{\boldsymbol{x}_t} \text{score}(A_t)\right),$$

where $\text{sign}(\cdot)$ is applied element-wise. This formulation accelerates updates while enabling finer control over the latent shift. We analyze the effect of α in Figure 6; unless otherwise stated, we use $\alpha = 3 \times 10^{-3}$ throughout.

4. Experiments

We evaluate JEDI in three settings: 1.) a synthetic comparison against the NT-Xent loss used in the latent optimization technique CONFORM (Meral et al., 2024); 2.) qualitative results on Stable Diffusion 3.5 (SD3.5)[1]; and 3.) quantitative experiments on Stable Diffusion 1.5 (SD1.5)[2], including a comparison to CONFORM and an evaluation of JEDI applied to LoRACLR, a variant of SD1.5 (Simsar et al., 2024), to demonstrate its broader applicability. All experiments were conducted on a NVIDIA GeForce GTX TITAN X. Implementation details are provided in Appendix E.

Synthetic Comparison. Contrastive objectives for aligning attention maps—bringing same-subject maps closer while pushing different ones apart—were first explored in CONFORM (Meral et al., 2024), which uses the NT-Xent loss (Oord et al., 2018; He et al., 2020; Chen et al., 2020) based on cosine similarity. While effective in embedding spaces, cosine similarity is not well-suited for optimizing probability distributions: it tends to collapse mass into narrow peaks and fails to capture broader structural relationships.

To illustrate this limitation, we construct a toy example with four overlapping 1D Gaussians: blue and green represent one subject, red and purple another. The objective is to align distributions within the same group while separating those across groups. As shown in Figure 2, JEDI preserves the support of each group, producing coherent mixtures, while NT-Xent distorts shapes and leads to over-concentrated and fragmented modes.

Since attention maps are soft spatial probability fields, preserving their continuity and avoiding artificial multimodality is critical, for example, to prevent the same subject from being generated in multiple places. JEDI's objective maintains this structure, encouraging stable and semantically grounded attention patterns for generation.

Stable Diffusion 3.5. We apply JEDI to SD3.5 and assess image quality on a custom dataset of prompts involving

[1] huggingface.co/stabilityai/stable-diffusion-3.5-medium
[2] huggingface.co/stable-diffusion-v1-5

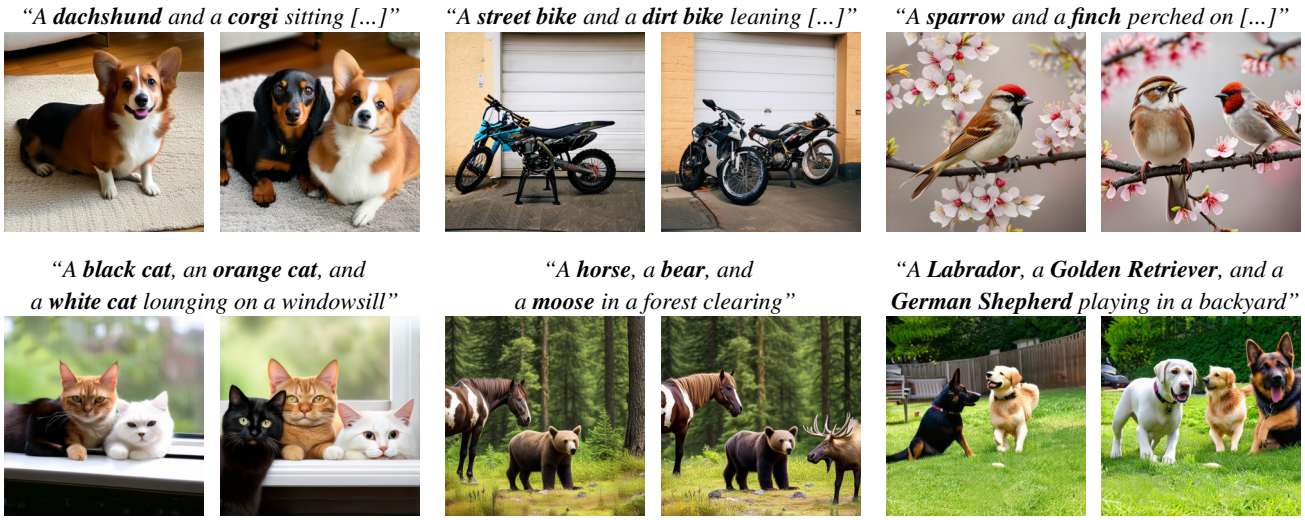

Figure 4. **Side-by-side comparison of Stable Diffusion 3.5 (left) and Stable Diffusion 3.5 + JEDI (right).** The base model often mixes attributes or omits subjects, while JEDI corrects these issues. See Figures 11 and 12 in Appendix G for full prompts and more examples.

Figure 5. **Comparison between LoRACLR (left) and LoRACLR + JEDI (right).** The base model without a control shows attribute mixing, while JEDI produces clearer subject separation.

visually similar object pairs (e.g., *"apple"* and *"pear"*). For each prompt, we generate two images: one using vanilla SD3.5 and one using SD3.5 + JEDI.

As shown in Figure 4, JEDI consistently improves over the base model by correctly rendering both subjects and reducing attribute mixing. Moreover, since we set the learning rate to $\alpha = 3 \times 10^{-3}$, the overall composition and background remain nearly unchanged (e.g., in *"A street bike and a dirt bike [...]"*). For additional examples, see Appendix G.

Comparison to CONFORM. To compare directly with CONFORM, originally designed for SD1.5, we adapt their implementation by replacing the CONFORM component with our JEDI objective. CONFORM performs optimization up to the 29th timestep, applying 20 iterative latent updates at steps 0, 10, and 20—totaling 69 updates. In contrast, JEDI achieves comparable or better results with just 18 updates, making it approximately 67% faster in practice.

Additionally, JEDI operates with a smaller learning rate, resulting in images that remain closer to the base model's distribution. By comparison, CONFORM begins with a much higher rate ($\alpha = 20$, tapering to 16.85), resulting in greater stylistic drift. Visual comparisons in Figure 3 highlight JEDI's superior subject separation and overall image quality. Additional examples are provided in Appendix G.

Extension to LoRACLR. We further test JEDI on LoRACLR (Simsar et al., 2024), a multi-concept model known to suffer from attribute mixing. On a model combining 14 distinct concepts, JEDI significantly improves subject separation (see Figure 5), highlighting its flexibility across architectures. Implementation details and additional examples are in Appendices E and G.

5. Discussion and Future Work

Unbiased Disentanglement Score. We find that the inter-group loss term in the JEDI objective naturally serves as an effective metric for measuring subject disentanglement during generation. For the images in Figure 1, the disentangled image achieves a mean JSD of 0.40 ± 0.15, compared to 0.17 ± 0.10 for the entangled counterpart, with full progression over time shown in Figure 10. Unlike CLIP-based metrics, this score is computed directly from internal attention maps, making it lightweight, model-internal, and free from external supervision or bias. This presents a promising alternative for evaluating subject separation in multi-object prompts, particularly in test-time settings.

Efficiency and Time Complexity. While JEDI is highly efficient relative to existing methods, it roughly doubles inference time compared to the base model. One potential solution is to shift its objective to the training or fine-tuning stage, as it is naturally defined over all diffusion steps and requires no supervision. This would reduce inference time while preserving the benefits of the JEDI framework.

References

Binyamin, L., Tewel, Y., Segev, H., Hirsch, E., Rassin, R., and Chechik, G. Make it count: Text-to-image generation with an accurate number of objects. *arXiv preprint arXiv:2406.10210*, 2024.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, Y., Wang, X., Wu, J. Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kwon, G., Jenni, S., Li, D., Lee, J.-Y., Ye, J. C., and Heilbron, F. C. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8880–8889, 2024.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Meral, T. H. S., Simsar, E., Tombari, F., and Yanardag, P. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9005–9014, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Simsar, E., Hofmann, T., Tombari, F., and Yanardag, P. Loraclr: Contrastive adaptation for customization of diffusion models. *arXiv preprint arXiv:2412.09622*, 2024.

Wei, T., Chen, D., Zhou, Y., and Pan, X. Enhancing mmdit-based text-to-image models for similar subject generation. *arXiv preprint arXiv:2411.18301*, 2024.

A. Hyperparameter

The learning rate α serves as a critical hyperparameter in our optimization, especially due to the use of the $\mathrm{sign}(\cdot)$ function, which restricts the gradient to unit magnitude.

As such, α directly controls the extent to which the latent image is updated at each step. To illustrate this effect, we show in Figure 6 the outputs of JEDI applied to Stable Diffusion 3.5 under varying values of α. As expected, large learning rates lead to overly aggressive updates, causing the optimization to diverge and fail to produce coherent images. Conversely, excessively small values have a negligible effect, resulting in little to no noticeable changes.

B. Proofs

Lemma B.1 (Upper Bound of Jensen-Shannon Divergence)**.** *Let $P = \{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(n)}\} \subset \mathbb{R}^d$ be a set of probability distributions. Then, $D_{\mathrm{JS}}(P)$ is upper bounded by $\log n$.*

Proof. Define P as in Lemma B.1, then the JSD is defined as follows:

$$D_{\mathrm{JS}}(P) = \frac{1}{n}\sum_{k=1}^{n} D_{\mathrm{KL}}\Big(\boldsymbol{p}^{(k)} \parallel \boldsymbol{m}\Big), \quad \boldsymbol{m} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{p}^{(k)}.$$

We can upper bound each D_{KL}-term as follows:

$$\begin{aligned} D_{\mathrm{KL}}(\boldsymbol{p}^{(k)} \parallel \boldsymbol{m}) &= \sum_{i=1}^{d} p_i^{(k)} \log \frac{p_i^{(k)}}{m_i} \\ &= \sum_{i=1}^{d} p_i^{(k)} \log \frac{p_i^{(k)}}{\frac{1}{n}\sum_{\ell=1}^{n} p_i^{(\ell)}} \\ &= \sum_{i=1}^{d} p_i^{(k)} \log \left(n \cdot \frac{p_i^{(k)}}{\sum_{\ell=1}^{n} p_i^{(\ell)}} \right) \\ &\leq \sum_{i=1}^{d} p_i^{(k)} \log n \\ &= \log n. \end{aligned}$$

Plugging this bound back into the definition of the JSD, yields the desired results:

$$\frac{1}{n}\sum_{k=1}^{n} D_{\mathrm{KL}}\Big(\boldsymbol{p}^{(k)} \parallel \boldsymbol{m}\Big) \leq \frac{1}{n}\sum_{k=1}^{n} \log n = \log n$$

□

Lemma B.2 (Upper Bound of Shannon Entropy)**.** *Let $\boldsymbol{p} \in \mathbb{R}^d$ be a discrete probability distribution, such that $p_i \geq 0$ and $\sum_i^d p_i = 1$. Then, its entropy $H(\boldsymbol{p})$ is upper bounded by $\log d$.*

Proof. Let $\boldsymbol{p}$ be defined as in Lemma B.2. We define a Lagrangian as follows:

$$\mathcal{L}(\boldsymbol{p}, \lambda) = H(\boldsymbol{p}) + \lambda \cdot \left(1 - \sum_{i}^{d} p_i\right),$$

where $\lambda \in \mathbb{R}$ is a Langrage multiplier. Taking the derivative with respect to each p_i and setting it to zero yields:

$$\nabla_{p_i}\mathcal{L}(\boldsymbol{p}, \lambda) = 0 \iff \log(p_i) = \lambda - 1.$$

Thus, all p_i must be equal at the maximum. Using the constraint $\sum_{i=1}^{d} p_i = 1$, it follows that $p_i = \frac{1}{d}$ for all i.

Substituting this result back into the definition of entropy gives:

$$H(\boldsymbol{p}) = -\sum_{i=1}^{d} p_i \log(p_i) = -\sum_{i=1}^{d} \frac{1}{d}\log(\frac{1}{d}) = \log(d).$$

□

C. Pseudo-code of JEDI

Algorithm 1 illustrates a minimal implementation of JEDI's test-time adaptation procedure, integrated into a standard iterative denoising loop of a diffusion model. The modifications introduced by JEDI are highlighted in blue, while the rest of the loop corresponds to the denoising process.

At each timestep, the model produces a denoised latent $\boldsymbol{x}_{t+1}$ along with the corresponding internal attention maps A_t.

Algorithm 1 JEDI Test-time Adaptation

Input: Condition prompt $\boldsymbol{c}$
$\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = 0$ to $T-1$ **do**
  if $t \leq K$ **then**
    $_, A_t \leftarrow \texttt{Model}(\boldsymbol{x}_t, \boldsymbol{c})$
    $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - \alpha \cdot \texttt{sign}(\nabla_{\boldsymbol{x}_t}\texttt{JEDI}(A_t, \boldsymbol{c}))$
  end if
  $\boldsymbol{x}_{t+1}, _ \leftarrow \texttt{Model}(\boldsymbol{x}_t, \boldsymbol{c})$
end for
return $\boldsymbol{x}_T$

D. Ablation

The JEDI objective comprises three additive components: *Intra-group Coherence*, *Inter-group Separation*, and a *Diversity Regularizer*. To evaluate the individual contribution of each term, we conduct an ablation study by systematically removing one component at a time. The results are shown in Figure 7.

Overall, the best results are achieved when all three components are included. Among them, *Inter-group Separation*

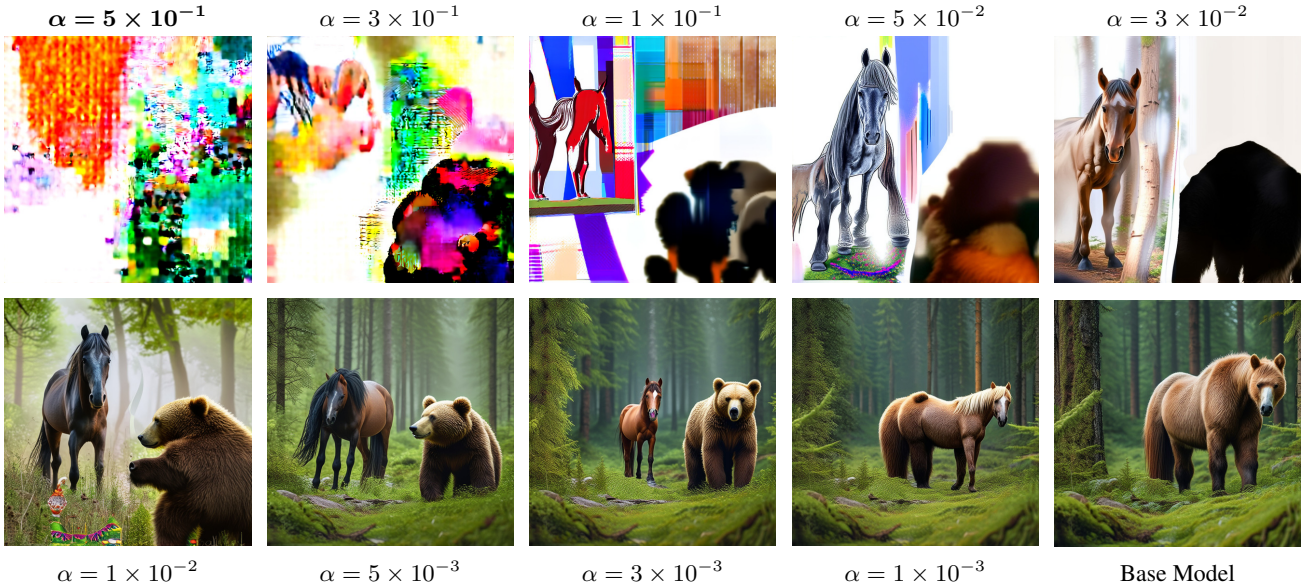

Figure 6. **Effect of learning rate α on image generation.** Outputs generated for the same prompt, "*A horse and a bear in a forest*," using Stable Diffusion 3.5 under identical settings, varying only the learning rate. Higher α values lead to excessive changes in the latent space, deteriorating image quality, while lower values result in minimal updates and limited visual difference.

has the most pronounced effect. This term encourages the model to spatially disentangle subjects, thereby reducing attribute mixing. Whenever it is removed, we observe noticeable shifts in image style and a significant increase in attribute mixing and spatial overlap between entities.

The effect of *Intra-group Coherence* is more subtle but still important. For example, in the generation of the *"moose"* subject, removing this term results in unnatural proportions. We attribute this degradation to misalignment between attention distributions produced by Stable Diffusion 3.5's dual text encoders (T5 and CLIP) which differ substantially in architecture and semantic representation. The coherence term helps align these internal representations, yielding more consistent subject rendering.

Finally, the contribution of the *Diversity Regularizer* is minimal in this setting. We scale this term with a small coefficient of $\lambda = 1 \times 10^{-2}$, which limits its influence during optimization. However, we found it to be beneficial in synthetic scenarios where we noticed attention map collapse. For this reason, we retain it as a safeguard.

E. Implementation Details

To facilitate reproducibility, we describe the key implementation details for each architecture evaluated. Full source code and experimental configurations are available on our project website: ericbill21.github.io/JEDI/.

Stable Diffusion 1.5. To enable a direct comparison with CONFORM (Meral et al., 2024), we adopt their implementation and replace the CONFORM module with our JEDI objective. Following their experimental setup, we sample the model for 50 timesteps using a guidance scale of 7.5.

During each forward pass, we extract cross-attention maps at a resolution of 16×16 and compute the JEDI objective over these maps. We then backpropagate the loss and update the latent variables using signed gradients, with a learning rate of $\alpha = 3 \times 10^{-3}$. Optimization is applied only during the first 18 timesteps, which already yields strong results. We do not perform extensive hyperparameter tuning, as our primary focus is on evaluating JEDI with SD3.5.

For LoRACLR (Simsar et al., 2024), which is built on the Mix-of-Show codebase (Gu et al., 2023), we extend the CONFORM implementation to operate within this framework by applying the same setup used for SD 1.5. The only modification is an extended optimization window of 30 timesteps to account for the observed increased attribute mixing in LoRACLR.

Stable Diffusion 3.5. Modern T2I models like Stable Diffusion 3.5 are based on the Diffusion Transformer (DiT) architecture by Peebles & Xie (2023), which replaces the traditional U-Net with a sequence of DiT blocks. Unlike U-Nets, DiT does not use explicit cross-attention between image and text tokens, making it more challenging to extract spatial attention distributions for individual prompt tokens. To approximate this behavior, we took inspiration from Wei et al. (2024).

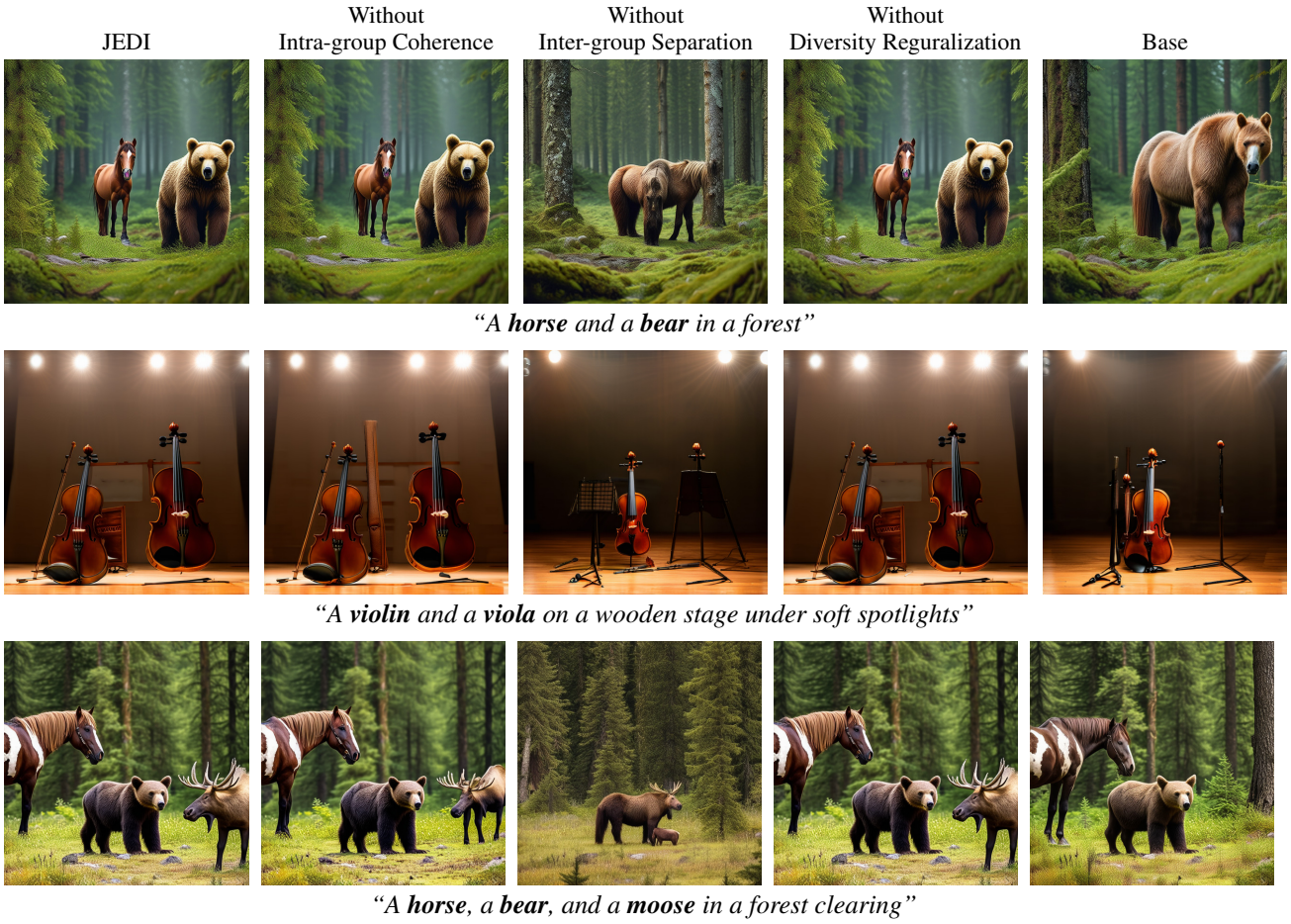

Figure 7. **Effect of individual components in the JEDI objective.** Outputs generated from the same prompt using Stable Diffusion 3.5 under identical sampling settings. The left column shows results with all components enabled. The middle columns each omit one component of the JEDI objective. The right column shows outputs from the base model without any JEDI adaptation.

Each DiT block processes image tokens $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and prompt tokens $\boldsymbol{C} \in \mathbb{R}^{m \times d}$ separately, producing respective query, key, and value matrices:

$$\boldsymbol{Q}_x, \boldsymbol{K}_x, \boldsymbol{V}_x \text{ (image)} \quad \text{and} \quad \boldsymbol{Q}_c, \boldsymbol{K}_c, \boldsymbol{V}_c \text{ (text)}.$$

These matrices are then concatenated to form the full attention inputs:

$$\begin{aligned} \boldsymbol{Q} &= \text{concat}[\boldsymbol{Q}_x, \boldsymbol{Q}_c], \\ \boldsymbol{K} &= \text{concat}[\boldsymbol{K}_x, \boldsymbol{K}_c], \\ \boldsymbol{V} &= \text{concat}[\boldsymbol{V}_x, \boldsymbol{V}_c]. \end{aligned}$$

Self-attention is applied over the combined sequence, yielding the attention matrix $\boldsymbol{A} = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^\top) \in \mathbb{R}^{(n+m)\times(n+m)}$. To estimate the spatial influence of prompt token i on the image, we compute:

$$\frac{1}{\sqrt{2}}\left(\boldsymbol{A}_{n+i,:n} + \boldsymbol{A}_{:n,n+i}^\top\right).$$

Since this expression is not guaranteed to form a normalized distribution, we consider two options: 1.) renormalize the result, or 2.) bypass the softmax during attention and apply it only during extraction, using raw logits. Empirically, we find the second approach yields more stable and consistent results.

SD3.5 contains 24 DiT blocks, each producing an attention map per prompt token. However, not all blocks provide equally useful information. Based on visual analysis and computational efficiency, we select blocks 5 to 15 for both extraction and optimization. See Figures 8 and 9 for example visualizations.

We find that applying latent optimization during the first 18 timesteps is sufficient. All experiments use these settings, along with 28 inference steps in total and a guidance scale of 4.5, following the official recommendations.

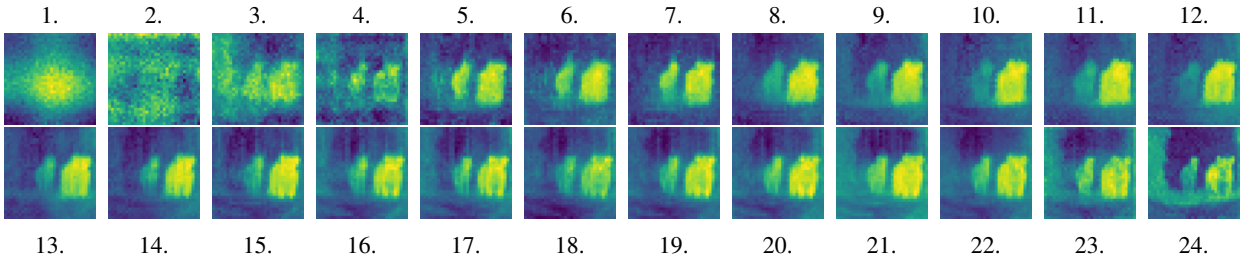

Figure 8. **Attention maps for the subject *"bear"*.** Extracted from diffusion timestep 13 of 28 across all 24 DiT blocks for the prompt *"A horse and a bear in a forest"*, using Stable Diffusion 3.5 with JEDI. The final generated image is shown in Figure 1.

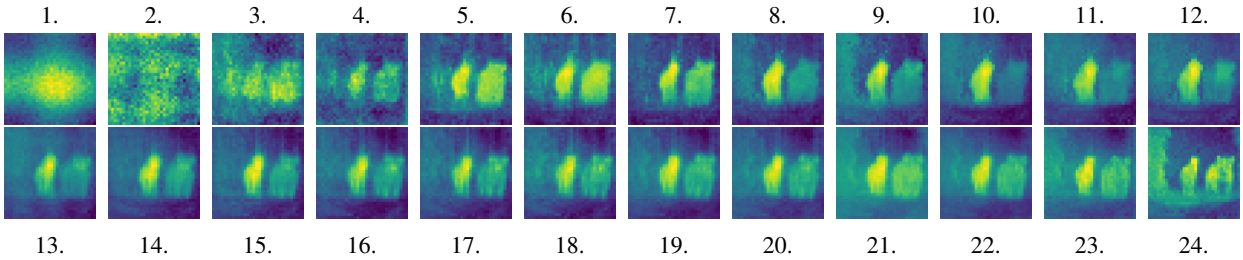

Figure 9. **Attention maps for the subject *"horse"*.** Extracted from diffusion timestep 13 of 28 across all 24 DiT blocks for the prompt *"A horse and a bear in a forest"*, using Stable Diffusion 3.5 with JEDI. The final generated image is shown in Figure 1.

F. Score

Figure 10 presents the inter-group JSD between the two subjects from Figure 1, computed across DiT blocks 7 to 15 over all diffusion timesteps. The image without attribute mixing exhibits consistently higher inter-group JSD values from timestep 5 onward, indicating stronger subject disentanglement.

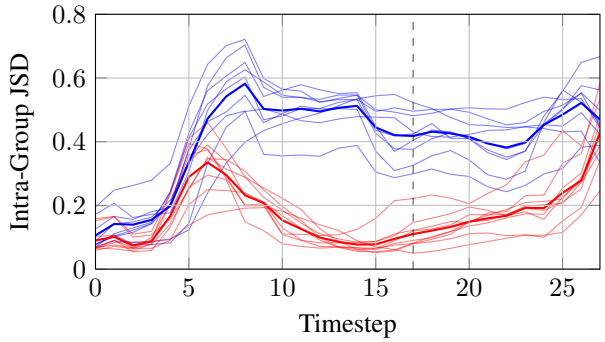

Figure 10. **Inter-group JSD across diffusion timesteps for the base model (red) and JEDI (blue).** Thick lines show the mean JSD across blocks. JEDI is applied only during the first 18 timesteps, indicated by the dashed vertical line.

G. Samples

We present additional qualitative results and highlight notable behaviors across different model variants.

Stable Diffusion 3.5. Figures 11 and 12 show samples across a broader range of object categories. A noteworthy observation is that when subjects are already well disentangled (i.e., no visible attribute mixing), JEDI leaves the image unchanged. This occurs because the JEDI loss approaches zero in such cases. For example, see the image pair with *"dachshund"* and *"corgi"*.

Stable Diffusion 1.5. Additional samples are shown in Figure 13. Due to CONFORM's relatively high learning rate, generated images occasionally deviate from the base model's distribution. For instance, a *"violin"* may appear with an unnatural blue color. This phenomenon is absent in JEDI, as the choice of $\alpha = 3 \times 10^{-3}$ prevents excessive deviation from the base model.

LoRACLR. Further LoRACLR results are shown in Figure 14. Since the LoRACLR model combines 14 concepts—many involving famous figures from film or sports—it occasionally disregards background details, leading to misalignment between the prompt and the image. However, this issue is inherent to the LoRACLR model and not introduced by JEDI. We solely demonstrate that JEDI can successfully disentangle the subjects of the prompt.

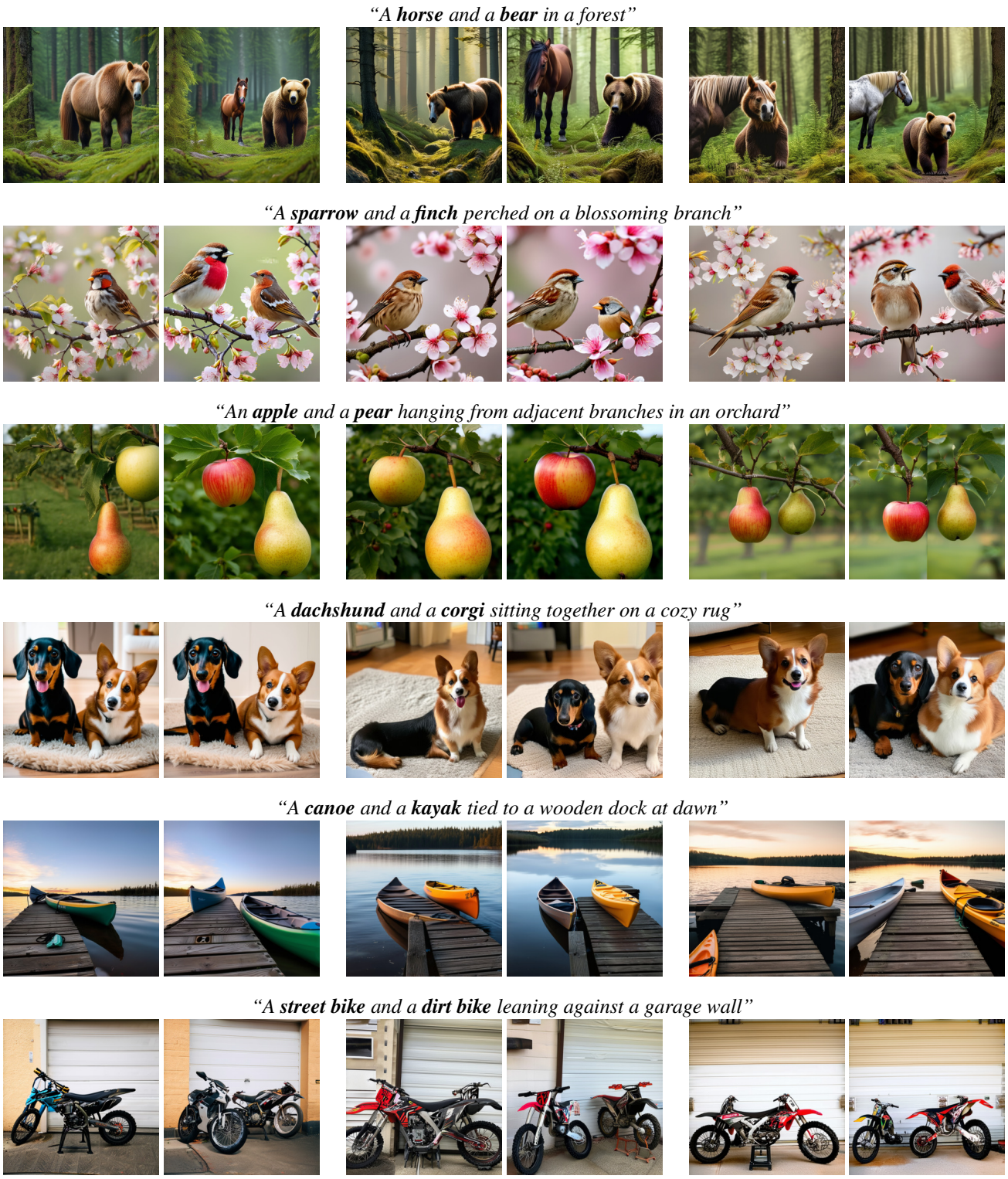

Figure 11. **Side-by-side comparison of Stable Diffusion 3.5 (left) and Stable Diffusion 3.5 + JEDI (right).** Each image pair was generated under identical conditions with a guidance scale of 4.5 and 28 inference steps.

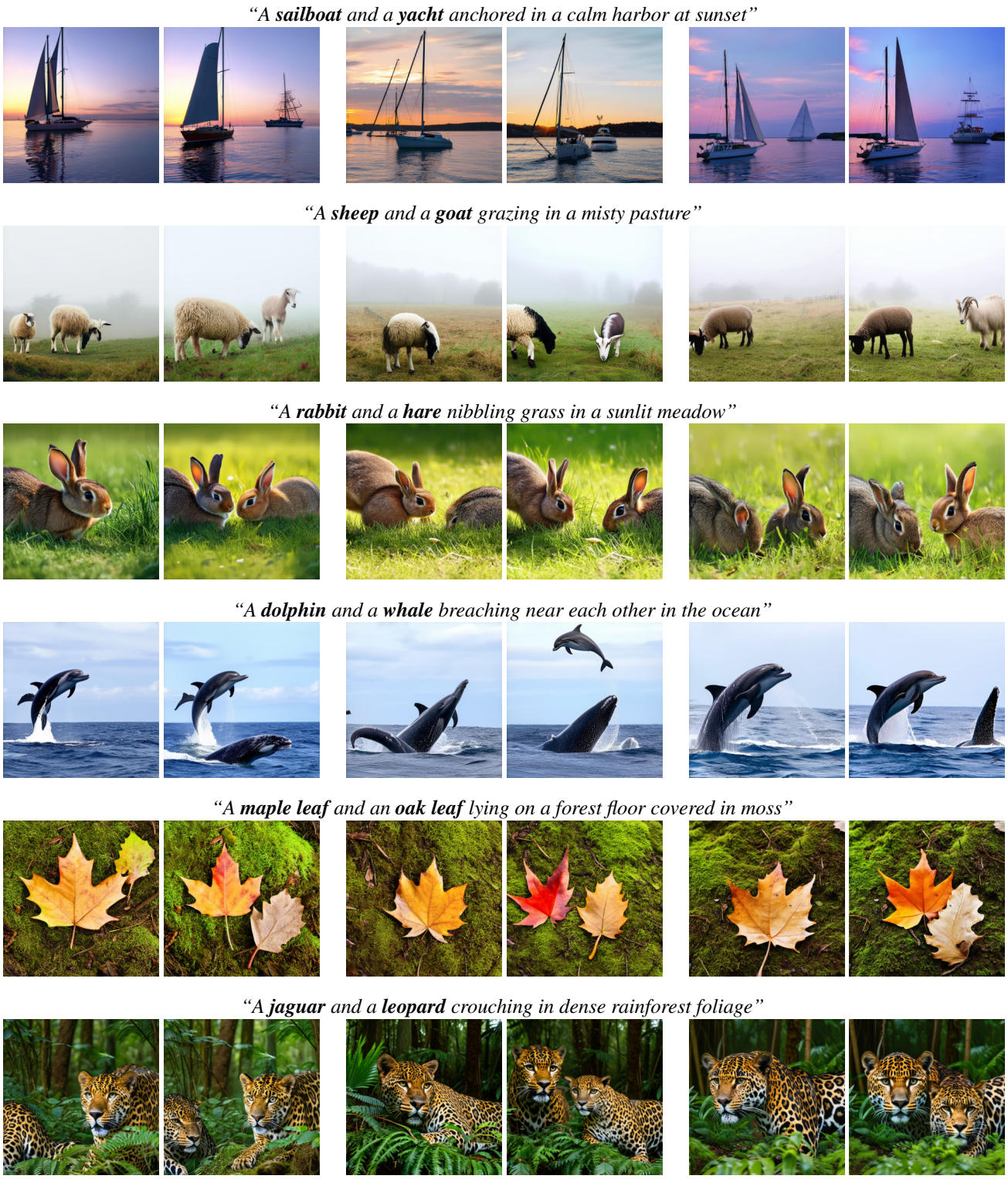

Figure 12. **Side-by-side comparison of Stable Diffusion 3.5 (left) and Stable Diffusion 3.5 + JEDI (right).** Each image pair was generated under identical conditions with a guidance scale of 4.5 and 28 inference steps.

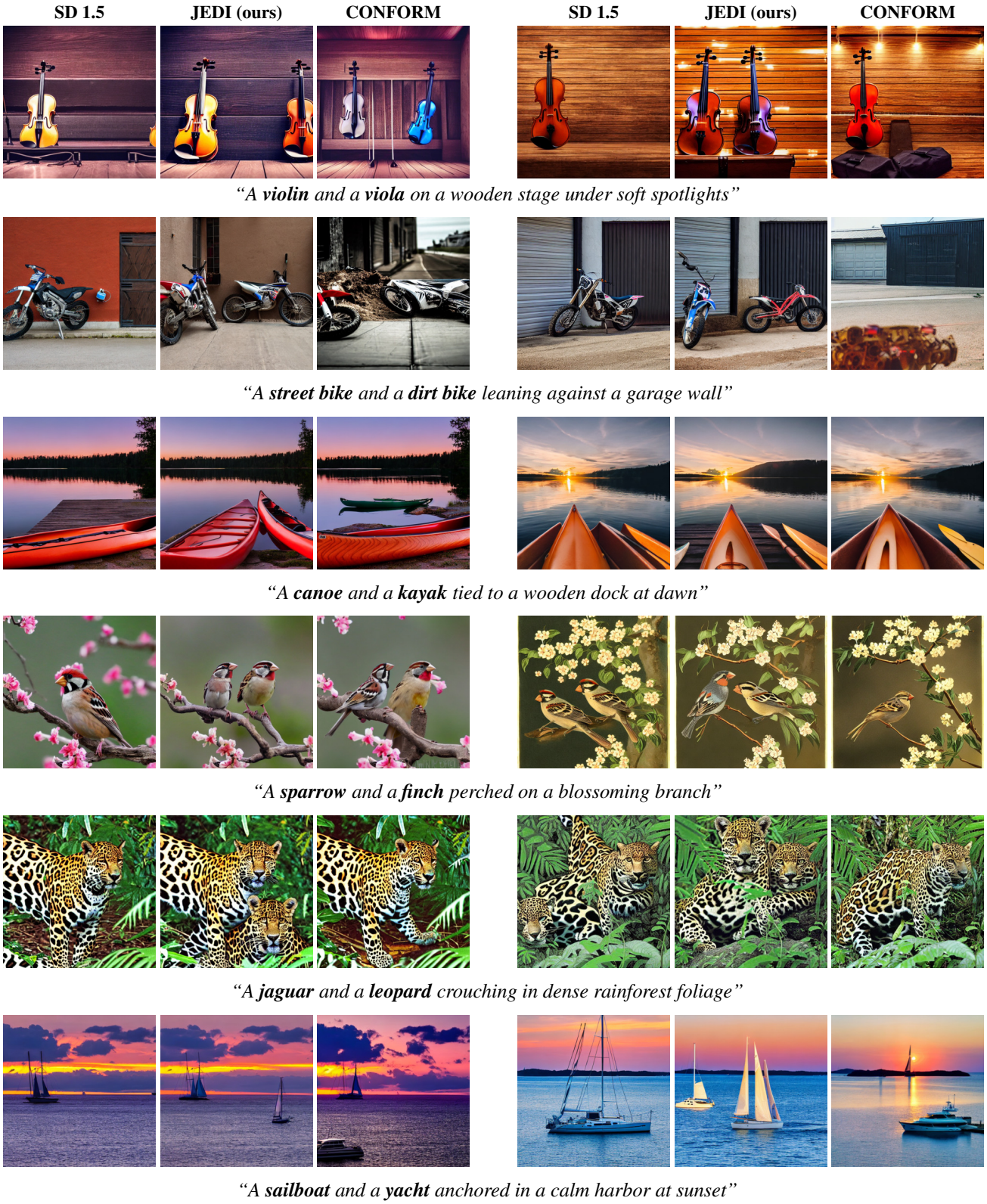

Figure 13. **Comparison of JEDI and CONFORM on Stable Diffusion 1.5.** Each image triplet was generated under identical conditions with 50 inference steps and a guidance scale of 7.5. For details of each method, refer to Appendix E.

Figure 14. **Comparison between LoRACLR (left) and LoRACLR + JEDI (right).** The baseline model exhibits attribute mixing between subjects (e.g., "Taylor" appears in football attire), whereas LoRACLR + JEDI achieves clearer subject disentanglement and preserves subject-specific features.