# Your Student is Better Than Expected: Adaptive Teacher-Student Collaboration for Text-Conditional Diffusion Models

Supplementary Material

## A. CD-SD1.5 implementation details

In this work, we develop consistency distillation (CD) for Stable Diffusion following the official implementation [50]. For training, we prepare a subset of LAION2B [45], which consists of 80M image-text pairs. As a teacher sampler, we consider DDIM-solver using 50 sampling steps and Variance-Preserving scheme [49]. We use the teacher UNet architecture as a student model and initialize it with the teacher parameters. Classifier-free guidance is applied to the distilled model directly without merging it into the model as done in [30]. During training, we uniformly sample the guidance strength from $1$ to $8$. Thus, our model supports different guidance scales during sampling. We train the student for $\sim$200K iterations on 8 A100 GPUs using the following setting: $512$ batch size; $0.95$ EMA rate; $1\mathrm{e}{-5}$ fixed learning rate; L2 uniformly weighted distillation loss calculated in the latent space of the VAE encoder. During inference, the multistep stochastic sampler [50] is used to generate images. In most of our experiments, we use $5$ sampling steps.

Note that we use our implementation of consistency distillation for SD because, when most experiments were conducted, there were no publicly available implementations.

## B. Analysis

### B.1. Details

**Image complexity**. To calculate the image complexity, we use the recent ICNet model [7]. This model is learned on a large-scale human annotated dataset. Each image corresponds to a complexity score ranging from 0 (the simplest) to 1 (the most complex). In Figure 14, we provide examples of Stable Diffusion samples with the lowest and highest image complexity. More complex images usually depict multiple entities, often including people, and intricate backgrounds.

**Text influence**. We calculate the influence of a text prompt on student generation by using cross-attention between token embeddings and intermediate image representations. Following [51], we collect cross-attention maps for all diffusion steps and UNet [39] layers. Then, the average attention score is calculated for each text token. Finally, the highest value among all tokens is returned.

**Trajectory curvature** is estimated according to the recent work [4]. First, we calculate the trajectory deviations as L2 distance from the denoised prediction at a time step $t$ to the straight line passing through the denoising trajectory endpoints. The trajectory curvature corresponds to the highest deviation over all time steps.

**Algorithm 1:** teacher-student adaptive collaboration

```
Input: $\mathbf{E}, \mathbf{S}, \mathbf{T}-$ estimator, student, teacher;
    $\mathcal{I}-$ input; $\sigma, \tau-$ rollback value and cut-off threshold.
$\hat{\mathcal{O}} = \mathbf{S}(\mathcal{I})$                    // Student prediction
if $\mathbf{E}(\hat{\mathcal{O}}) < \tau$ then
    Strategy 1:                                   // Refinement
        $\hat{\mathcal{O}}_\sigma = \sqrt{1-\sigma} \cdot \hat{\mathcal{O}} + \sqrt{\sigma} \cdot \mathcal{Z},\ \mathcal{Z} \sim \mathcal{N}(0, 1)$
        $\hat{\mathcal{O}} = \mathbf{T}(\mathcal{I}, \hat{\mathcal{O}}_\sigma)$
    Strategy 2:                                   // Regeneration
        $\hat{\mathcal{O}} = \mathbf{T}(\mathcal{I})$
return $\hat{\mathcal{O}}$
```

In Figure 15 (Left), we visualize two denoising trajectories corresponding to high and low curvatures. We apply PCA [16] to reduce the dimensionality of the denoised predictions. In addition, Figure 15 (Right) demonstrates trajectory deviations for different time steps. The highest deviations typically occur closer to the end of the denoising trajectory.

### B.2. Other distilled text-to-image models

Here, we conduct a similar analysis as in Section 3 for consistency distillation on Dreamshaper v7 [28] and architecture-based distillation[2].

In contrast to the few-step distillation approaches [20, 24, 28, 30, 50], the architecture-based distillation removes some UNet layers from the student model for more efficient inference and trains it to imitate the teacher using the same deterministic solver and number of sampling steps.

In Figures 19, 20, we show that both methods can produce samples that significantly differ from the teacher ones for the same text prompt and initial noise. Then, Figures 21, 22 confirm that other observations in Section 3 remain valid for the Dreamshaper and architecture-based students as well.

## C. Cut-off threshold tuning

The cut-off threshold $\tau$ is used for the adaptive selection of student samples. It corresponds to the $k$-th percentile of the metric values calculated on validation student samples. $600$ and $300$ samples are generated for tuning on the COCO2014 and LAION-Aesthetics datasets, respectively.

[2] https://huggingface.co/docs/diffusers/using-diffusers/distilled_sd

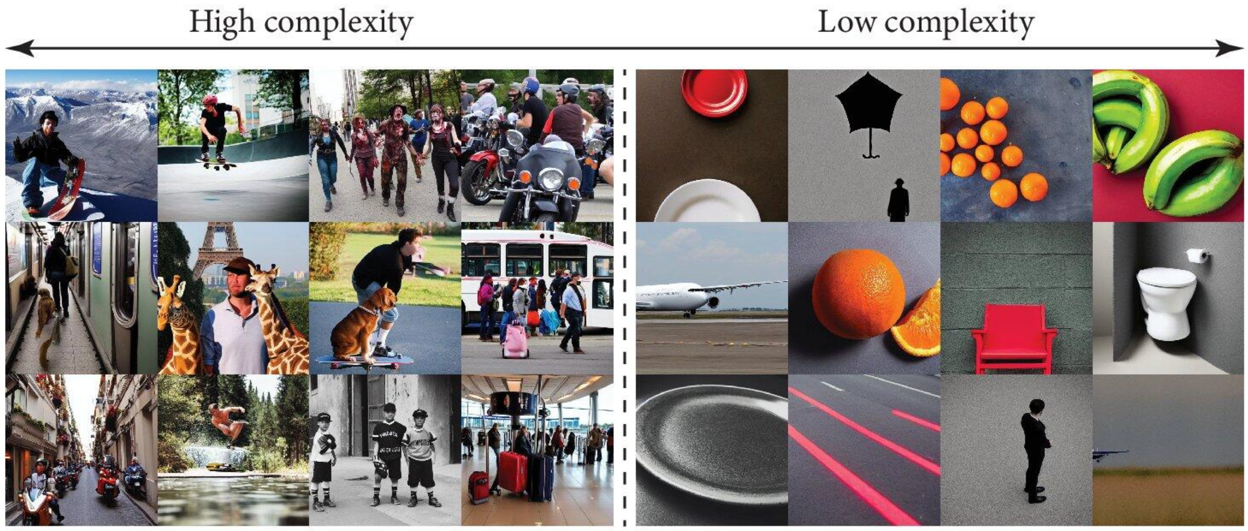


Figure 14. SD1.5 samples of different complexity according to the ICNet model [7].

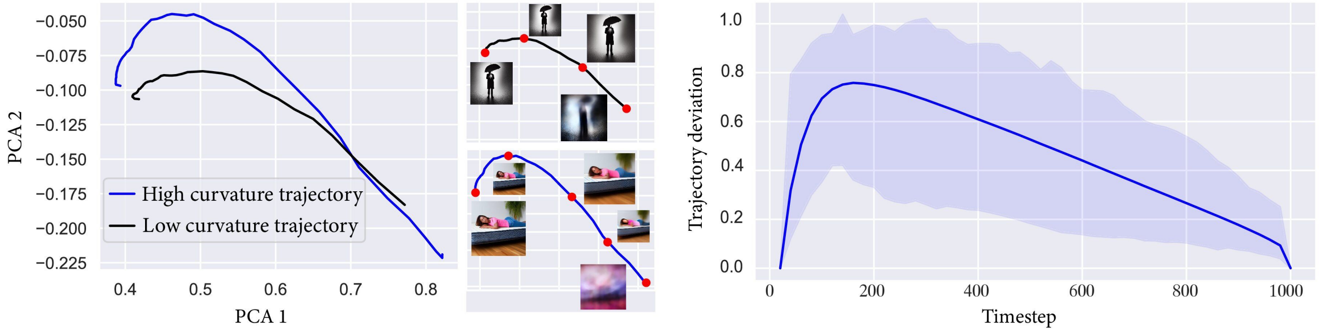


Figure 15. *Left*: Two examples of diffusion model trajectories with high and low curvatures. *Right*: Trajectory deviations according to [4].

Note that the prompts used for tuning do not overlap with the test ones. Then, we calculate an individual score, e.g., IR score, for each validation student sample and select the percentile based on an average inference budget or target metric value. For example, suppose we select the percentile given a 15 step budget and intend to perform 5 student steps and 20 steps for the improvement strategy. In this case, we have to select $\tau$ as a 50-th percentile, which results in the final average number of steps: $5 + 0.5 \cdot 20 = 15$.

During inference, we perform the adaptive selection as follows: if the score of the student sample exceeds $\tau$, we consider that this sample might be superior to the teacher one and keep it untouched. Otherwise, we perform an improvement step using the teacher model (refinement or regeneration). The proposed pipeline is presented in Algorithm 1.

We also show that hyperparameter tuning is straightforward and requires negligible effort. To verify this, we tune the threshold using the various number of prompts (Tab. 2). We can see that 500 prompts are sufficient for the threshold convergence. Thus, the process only needs to generate 500 student samples and takes $\sim$2.5 minutes.

| Prompts | 150 | 250 | 500 | 5000 | 25000 |
|---|---|---|---|---|---|
| Diffusion-DB | $0.646_{.104}$ | $0.565_{.028}$ | $0.569_{.010}$ | $0.570_{.008}$ | $0.568_{.004}$ |
| PickScore | $0.537_{.032}$ | $0.436_{.019}$ | $0.479_{.008}$ | $0.476_{.006}$ | $0.481_{.002}$ |
| Time, sec. | $44.9_{.4}$ | $73.8_{.2}$ | $148_{2}$ | $1470_{8}$ | $7448_{10}$ |

Table 2. The threshold value tuned using prompts from Diffusion-DB and PickScore datasets and the time required for tuning.

# D. Experiments

## D.1. Human evaluation

To evaluate the text-to-image performance, we use the side-by-side comparison conducted by professional annotators. Before the evaluation, all annotators pass the training and undergo the preliminary testing. Their decisions are based on the three factors: textual alignment, image quality and aesthetics (listed in the priority order). Each side-by-side comparison is repeated three times by different annotators. The final result corresponds to the majority vote.

## D.2. Experimental setup (SD1.5)

The exact hyperparameter values and number of steps used for the automated estimation (FID, CLIP score and ImageReward) of the adaptive refinement strategy in Table 3. The adaptive regeneration uses the same cut-off thresholds and number of steps, but the rollback value, $\sigma$, is equal to 1. The values used for human evaluation are presented in Table 4.

## D.3. Experimental setup (SDXL)

We evaluate two distilled models: CD-SDXL [28] and ADD-XL [44]. For the first evaluation, we use 4 steps of the CD-SDXL and then apply 12 adaptive refinement steps using the teacher model (SDXL-Base [34]) with the UniPC solver [58]. We compare our pipeline to the default teacher configuration: 50 DDIM steps. For the second evaluation, we perform 2 steps of the ADD-XL and 4 steps of the SDXL-Refiner [34] for the adaptive refinement strategy. We compare to 4 ADD-XL steps as this setting outperformed SDXL-Base in terms of image quality and textual alignment [34]. The exact hyperparameter values and number of steps for human evaluation are in Table 5.

We found that SDXL-Refiner performs slightly better than the base model for small refinement budgets (e.g., 4). The refiner typically helps to improve fine-grained details, e.g., face attributes or background details. However, it faces difficulties in providing global changes and sometimes brings artifacts for large rollback values, $\sigma$. Thus, we use the SDXL-Base teacher for more refinement steps (e.g., 12).

| Metric | $\sigma$ | $k$ | Steps | | |
|---|---|---|---|---|---|
| | | | CD | Refinement | Adaptive |
| ImageReward | 0.4 | 60 | 5 | 5 | 8 |
| | 0.55 | 60 | 5 | 10 | 11 |
| | 0.7 | 60 | 5 | 15 | 14 |
| | 0.7 | 60 | 5 | 25 | 20 |
| | 0.7 | 60 | 5 | 35 | 26 |
| | 0.7 | 60 | 5 | 45 | 32 |
| CLIP score | 0.5 | 60 | 5 | 5 | 8 |
| | 0.7 | 60 | 5 | 10 | 11 |
| | 0.75 | 60 | 5 | 15 | 14 |
| | 0.75 | 60 | 5 | 25 | 20 |
| | 0.75 | 60 | 5 | 35 | 26 |
| | 0.75 | 60 | 5 | 45 | 32 |
| FID | 0.75 | 40 | 3 | 5 | 5 |
| | 0.7 | 70 | 3 | 15 | 14 |

Table 3. Hyperparameter values used for the **automated evaluation (SD 1.5)**, Figure 10.

| Metric | $\sigma$ | $k$ | Steps | | |
|---|---|---|---|---|---|
| | | | CD | Refinement | Adaptive |
| Human evaluation | 0.7 | 50 | 5 | 10 | 10 |
| | 0.7 | 50 | 5 | 20 | 15 |
| | 0.7 | 50 | 5 | 40 | 25 |

Table 4. Hyperparameter values used for the **user preference study (SD 1.5)**, Figure 9.

| Metric | $\sigma$ | $k$ | Steps | | |
|---|---|---|---|---|---|
| | | | ADD-XL | Refinement | Adaptive |
| Human evaluation | 0.4 | 50 | 2 | 4 | 4 |
| | | | CD-SDXL | Refinement | Adaptive |
| Human evaluation | 0.85 | 70 | 4 | 12 | 13 |

Table 5. Hyperparameter values used for the **user preference study (SDXL)**, Figure 11.

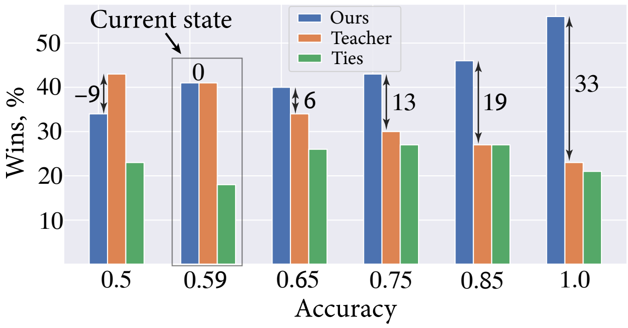


Figure 16. User preferences for different accuracy levels of the no-reference decision-making procedure. the automated sample estimator. *Current state* represents the results using ImageReward. The results for higher accuracy rates demonstrate the future gains if the oracle performance improves.

## D.4. Effect of oracle accuracy

The potential bottleneck of our approach is a poor correlation of existing text-to-image automated estimators with human preferences. For example, ImageReward usually exhibits up to 65% agreement with annotators. Moreover, it remains unclear what oracle accuracy can be achieved with no-reference decision-making, even if the estimator provides the perfect agreement. In Figure 16, we conduct a synthetic experiment examining the effect of the oracle accuracy on our scheme performance to reveal its future potential. We compare the adaptive refinement method (10 steps) to SD1.5 (50 steps) manually varying the oracle accuracy. We observe significant future gains even for the 75% accuracy rate.

## D.5. Distribution diversity

In the proposed teacher-student collaboration, the oracle aims to accept high-quality and well-aligned student samples

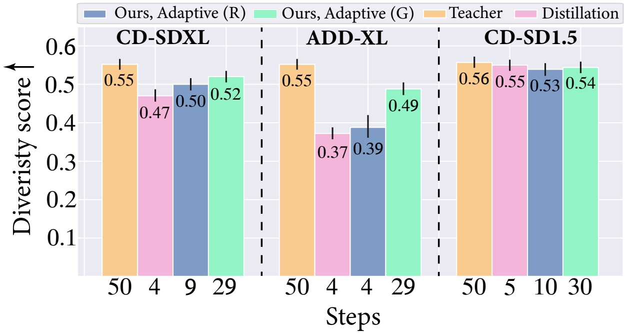


Figure 17. Diversity human scores collected for different methods.

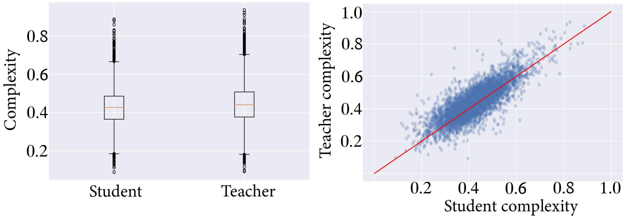


Figure 18. Image complexity of the CD-SD1.5 and SD1.5 samples in terms of ICNet [7]. *Left:* Box plot representing the complexity quantiles for both models. *Right:* Distribution of individual complexity values. Each dot corresponds to a pair of samples generated for the same prompt and initial noise. The distilled model only slightly simplifies the teacher distribution.

but does not control the diversity of the resulting image distribution. Therefore, if the student exhibits severe mode collapse or oversimplifies the teacher samples, the adaptive pipeline will likely inherit these issues to some extent.

In this section, we investigate this potential problem for several existing distilled text-to-image models. Specifically, we consider consistency distillation [28] for SD1.5 and SDXL [34] models and ADD-XL [44]. Note that ADD-XL is a GAN-based distillation method that generates exceptionally realistic samples but has evidence to provide poor image diversity for the given text prompt [44].

We estimate the diversity of generated images by conducting a human study. In more detail, given a text prompt and a pair of samples generated from different initial noise samples, assessors are instructed to evaluate the diversity of the following attributes: angle of view, background, main object and style. For each model, the votes are collected for 600 text prompts from COCO2014 and aggregated into the scores from 0 to 1, higher scores indicate more diverse images. The results are presented in Figure 17.

CD-SDXL demonstrates significantly better diversity than ADD-XL but still produces less various images compared to the SDXL teacher. CD-SD1.5 performs similarly to the SD1.5 teacher. Also, both adaptive strategies increase the diversity of the SDXL student models, especially the regeneration one. In Figure 30, we illustrate the diversity of images generated with different distilled text-to-image models (ADD-XL, CD-SDXL and CD-SD1.5). Each column corresponds to a different initial noise (seed). We notice that ADD-XL exhibits the lowest diversity compared to the CD-based counterparts.

Then, we address whether the distilled models tend to oversimplify the teacher distribution. In this experiment, we evaluate SD1.5 using DDIM for 50 steps and the corresponding CD-SD1.5 using 5 sampling steps. In Figure 18, we compare the complexity of the student and teacher samples in terms of the ICNet score [7]. We observe that CD-SD1.5 imperceptibly simplifies the teacher distribution.

To sum up, in our experiments, the CD-based models provide the decent distribution diversity that can be further improved with the proposed adaptive approach.

### D.6. Controllable generation

For both tasks, we use the adaptive refinement strategy and set the rollback value $\sigma$ to 0.5. We perform 5 steps for the student generation and 10 steps for the refinement with the UniPC solver. The cut-off thresholds correspond to 70 and 50 ImageReward percentiles for the mask-guided and edge-guided generation, respectively. We select random 600 image-text pairs from the COCO2014 validation set for the edge-guided generation. For the mask-guided generation, we use 600 semantic segmentation masks from the ADE20K dataset [59] and use the category names as the text prompts. For evaluation, we conduct the human study similar to D.1.

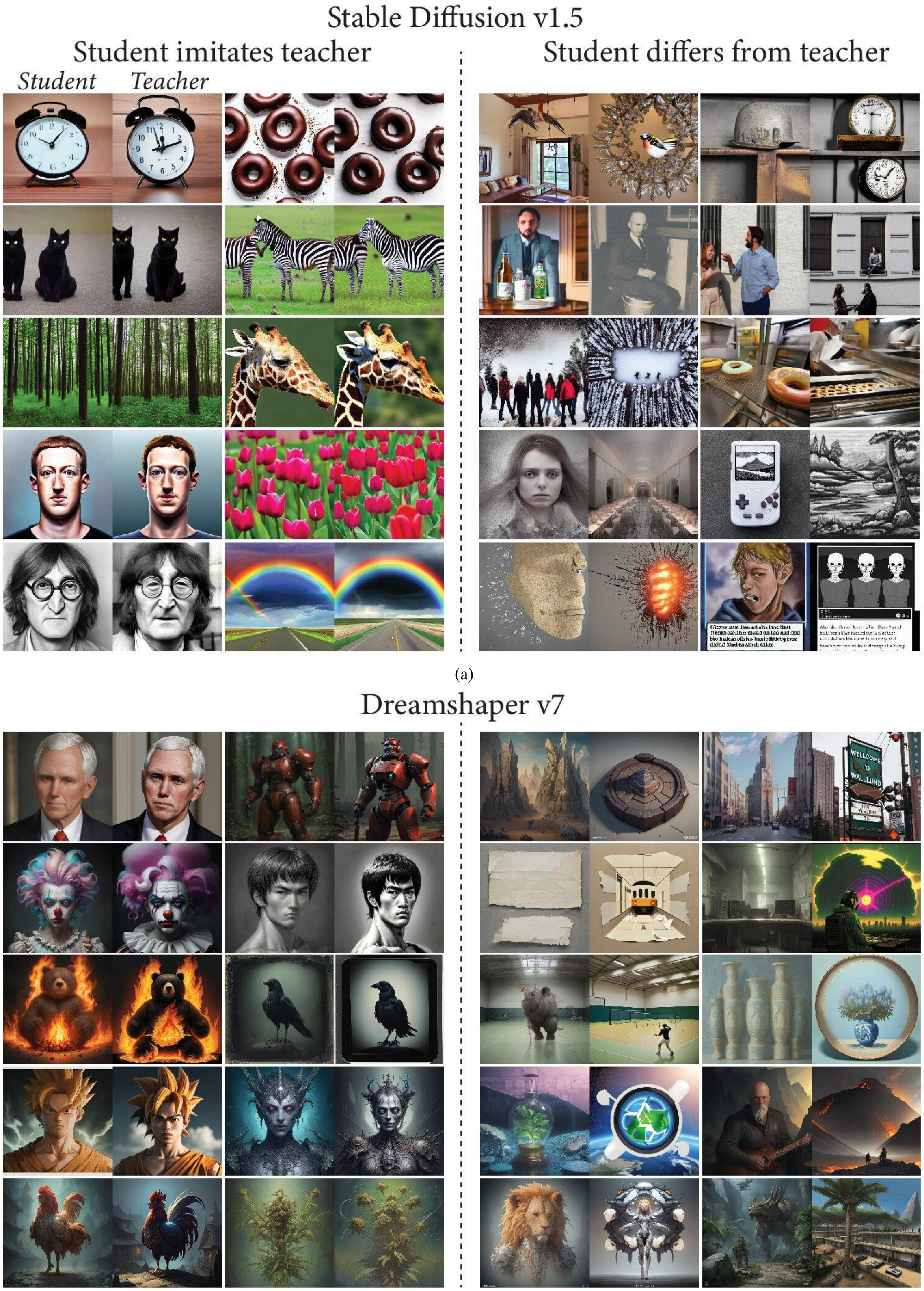
### D.7. ImageReward inference costs

We compare the absolute inference times of a single Stable Diffusion UNet step with classifier-free guidance against the ImageReward forward pass. We measure the model performance in half precision on a single NVIDIA A100 GPU. The batch size is 200 to ensure 100% GPU utility for both models. The performance is averaged over 100 independent runs. ImageReward demonstrates 0.26s while the single step of Stable Diffusion takes 3s. In the result, we consider the adaptive step costs negligible since ImageReward is more than 10× faster than a single generation step of Stable Diffision.

### D.8. Additional visualizations

In Figure 23, 24 we present qualitative verification of the first observation (the student sometimes outperforms its teacher according to the human evaluation). Figure 25 supports the second observation (the student wins are more likely where its samples differ from the teacher ones). In Figure 26 we demonstrate two adaptive strategies (refining and regeneration), the examples confirm that the refinement strategy improves the image fidelity and does not significantly alter the textual alignment, while the regeneration strategy may improve textual alignment.

Figures 27, 28, 29 provide more qualitative comparisons of our approach for different tasks.



(a)

(b)

Figure 19. Visual examples of similar (Left) and dissimilar (Right) teacher and student samples for SD1.5 (a) and Dreamshaper v7 (b).

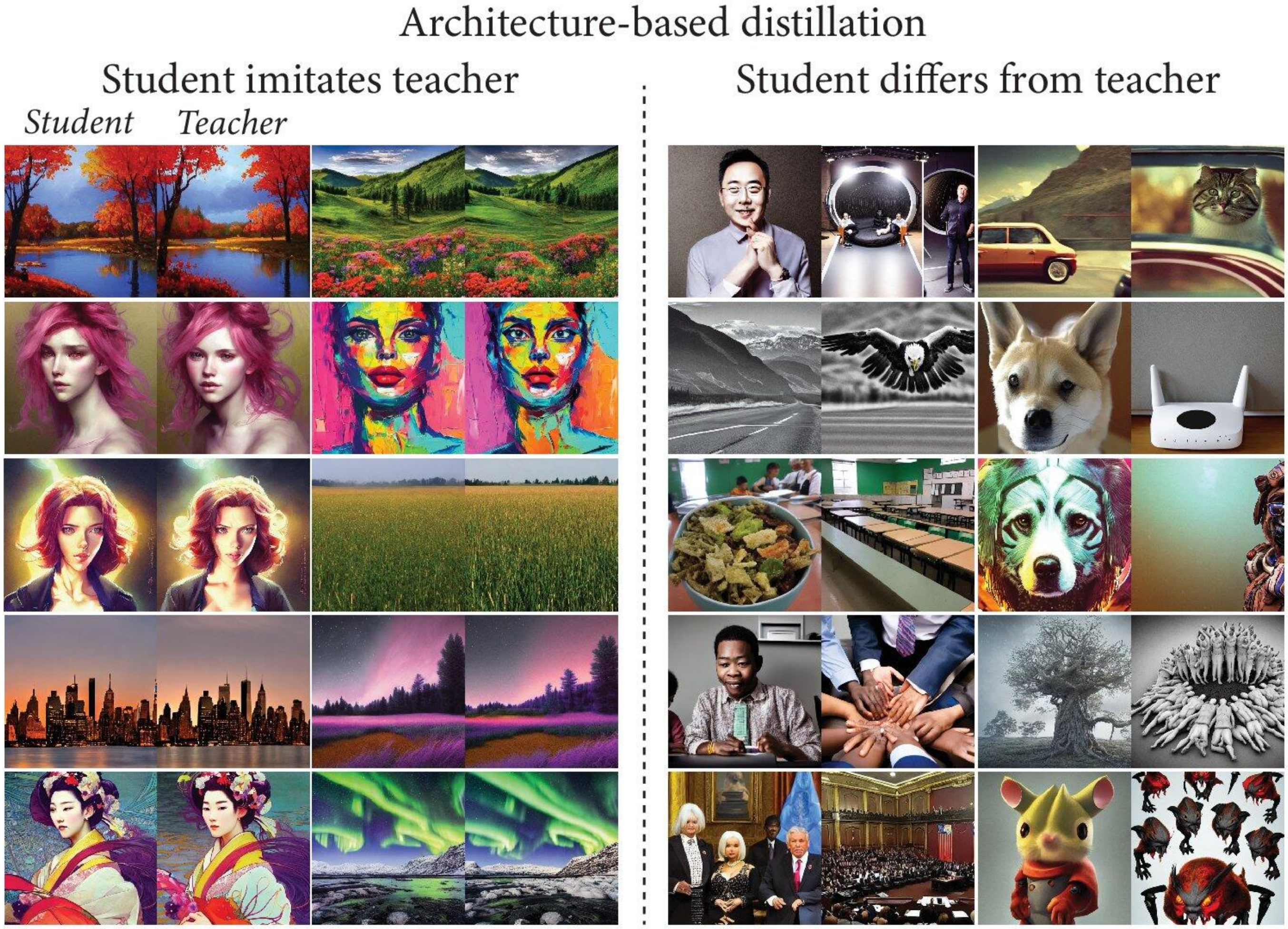
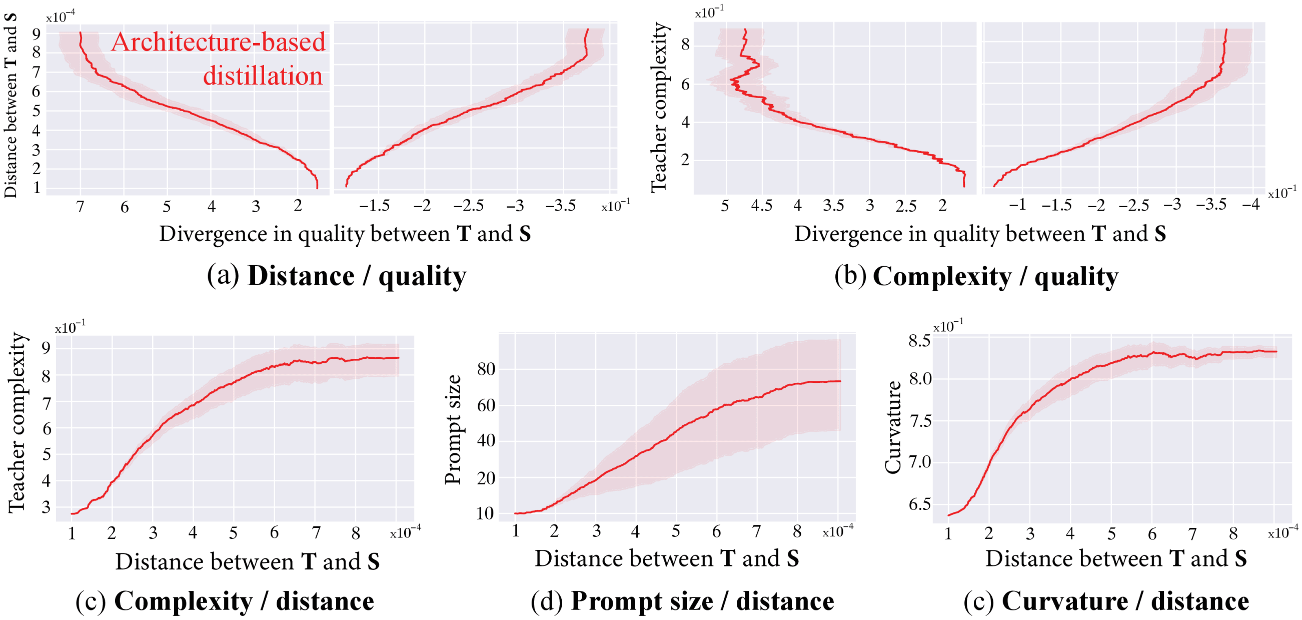


Figure 20. Visual examples of similar (Left) and dissimilar (Right) teacher and student samples for the architecture-based distillation.



(a) **Distance / quality**

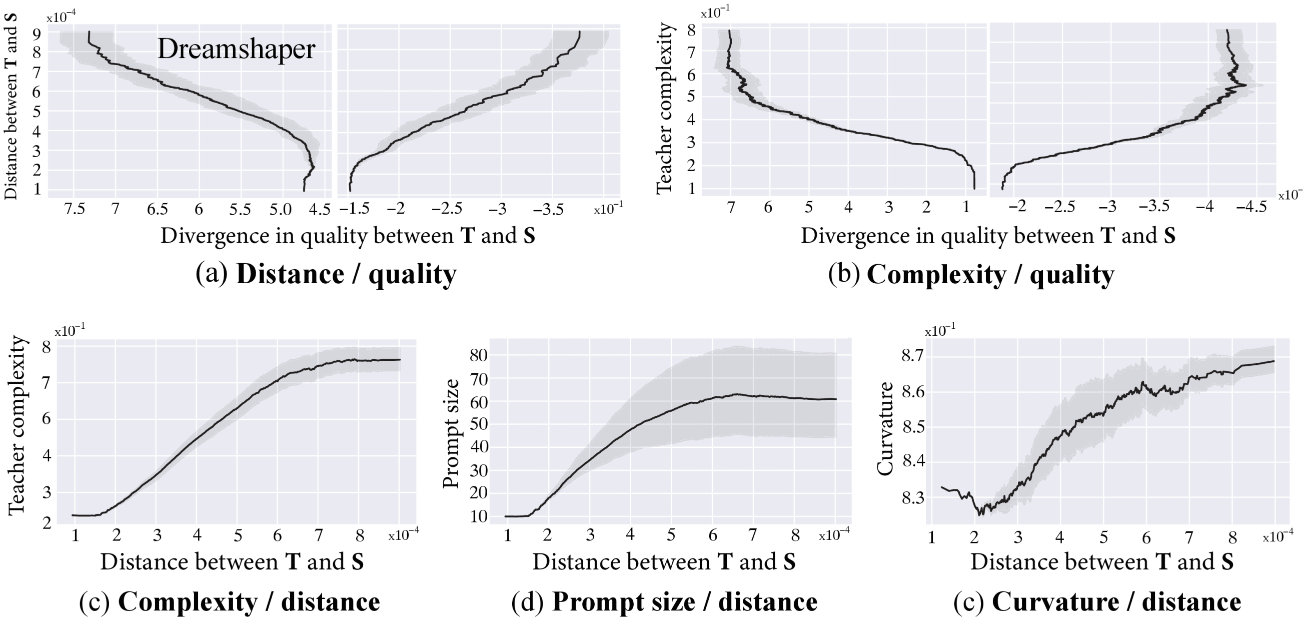
(b) **Complexity / quality**

(c) **Complexity / distance**

(d) **Prompt size / distance**

(c) **Curvature / distance**

Figure 21. Analysis results for the architecture-based distillation.



(a) **Distance / quality**

(b) **Complexity / quality**

(c) **Complexity / distance**

(d) **Prompt size / distance**

(c) **Curvature / distance**

Figure 22. Analysis results for the consistency distillation on Dreamshaper v7.

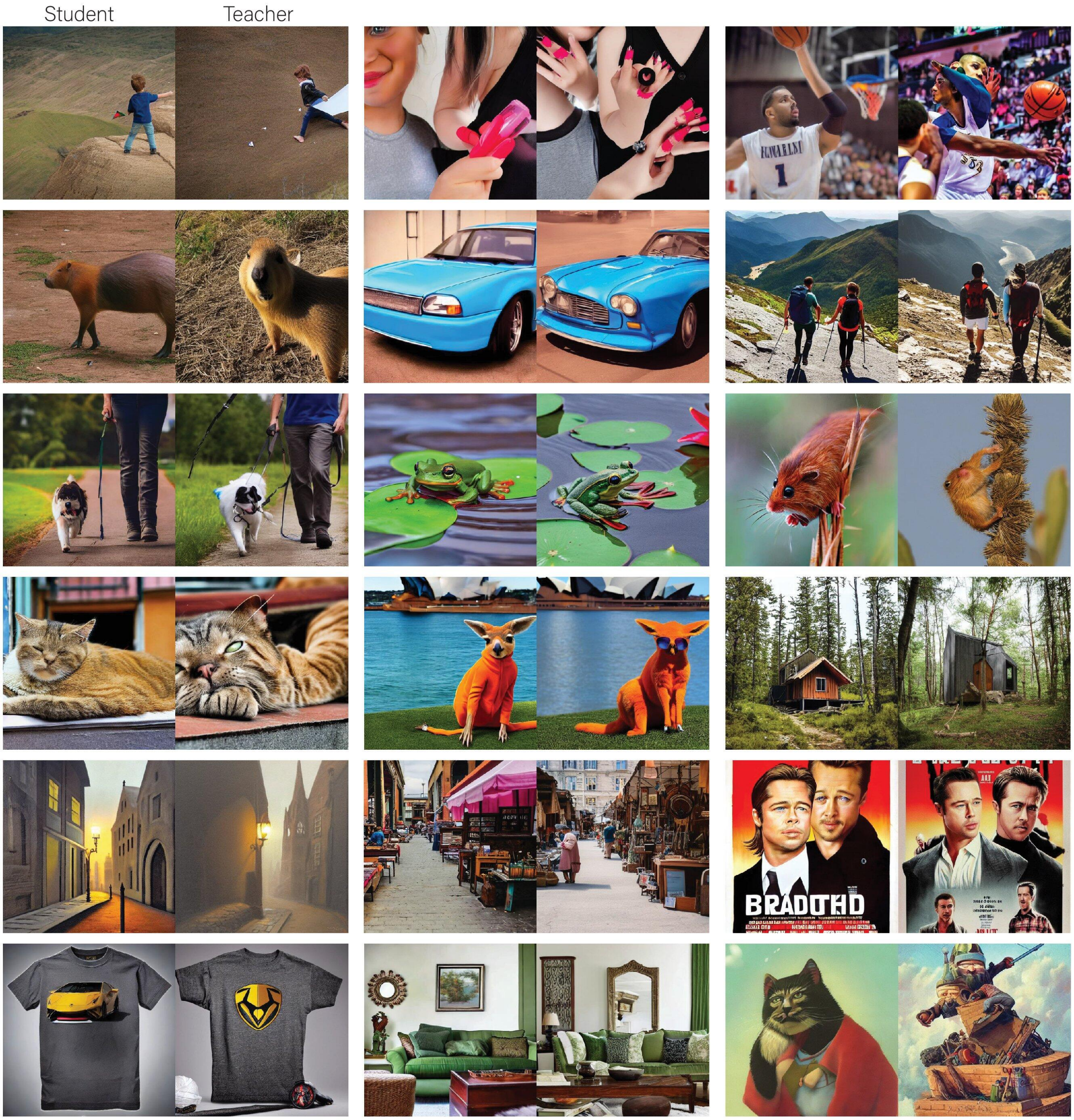


Figure 23. Additional examples where the student (CD-SD1.5) outperforms its teacher (SD1.5) according to the human evaluation.

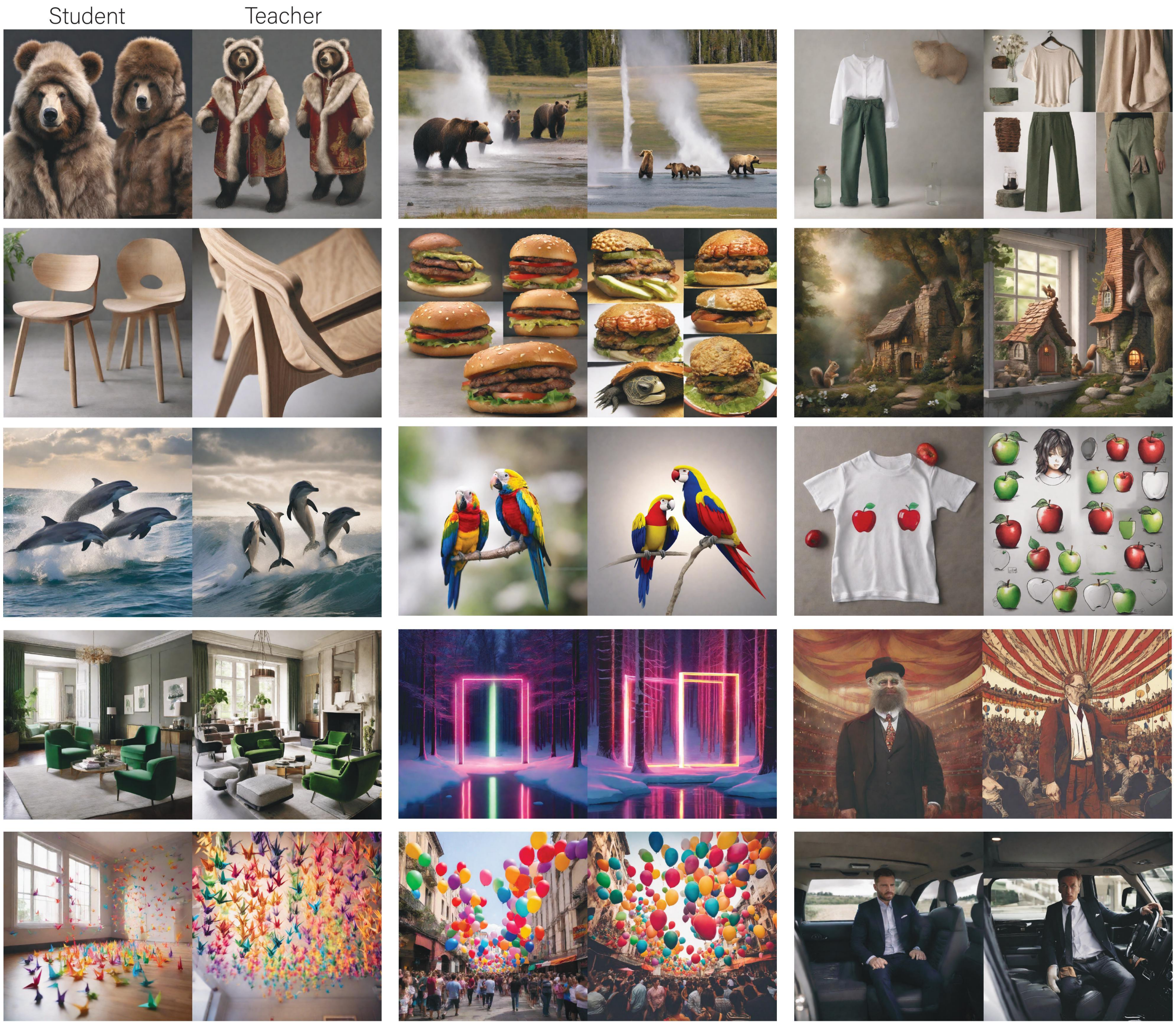


Figure 24. Additional examples where the student (CD-SDXL) outperforms its teacher (SDXL) according to the human evaluation.

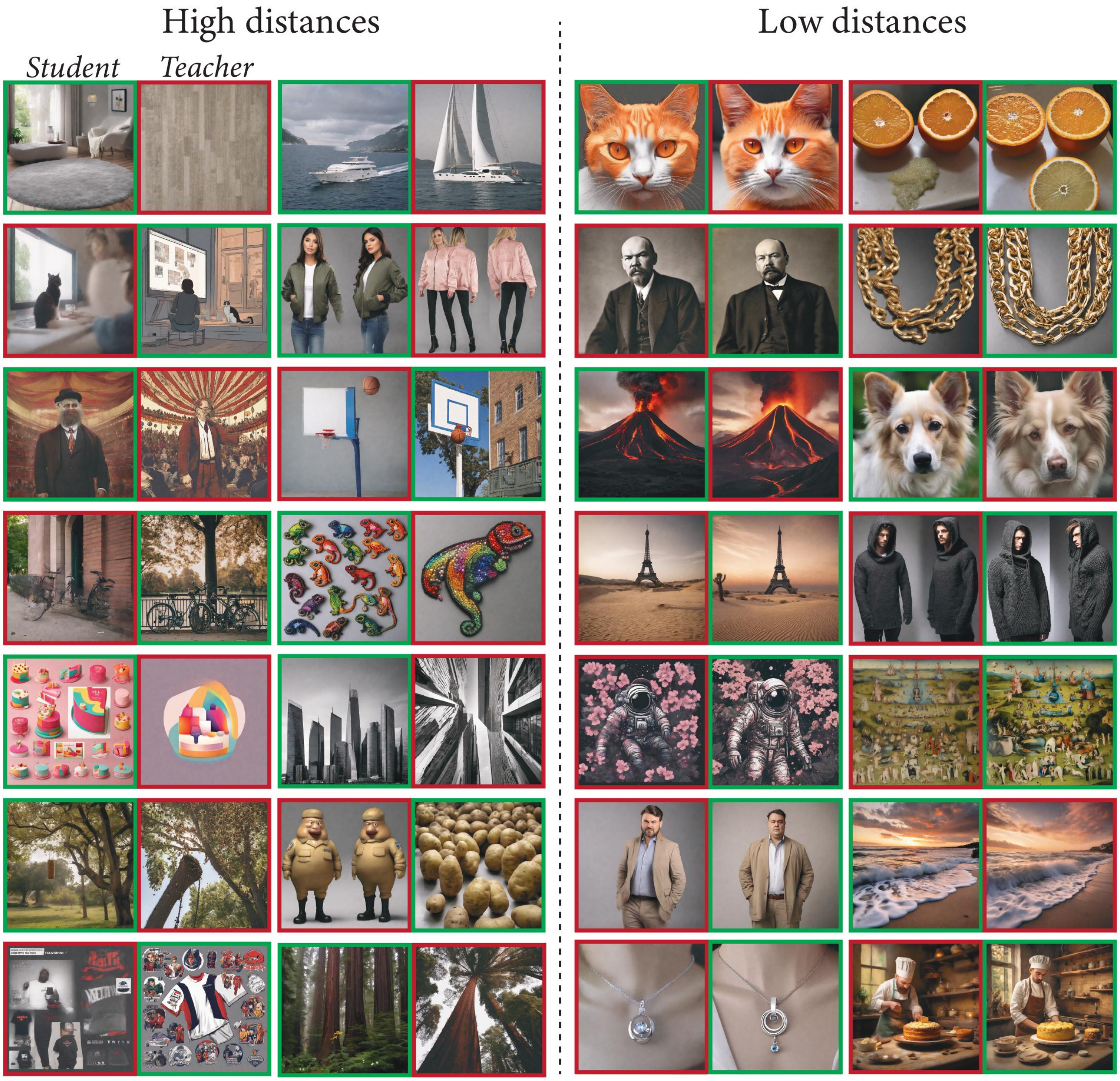
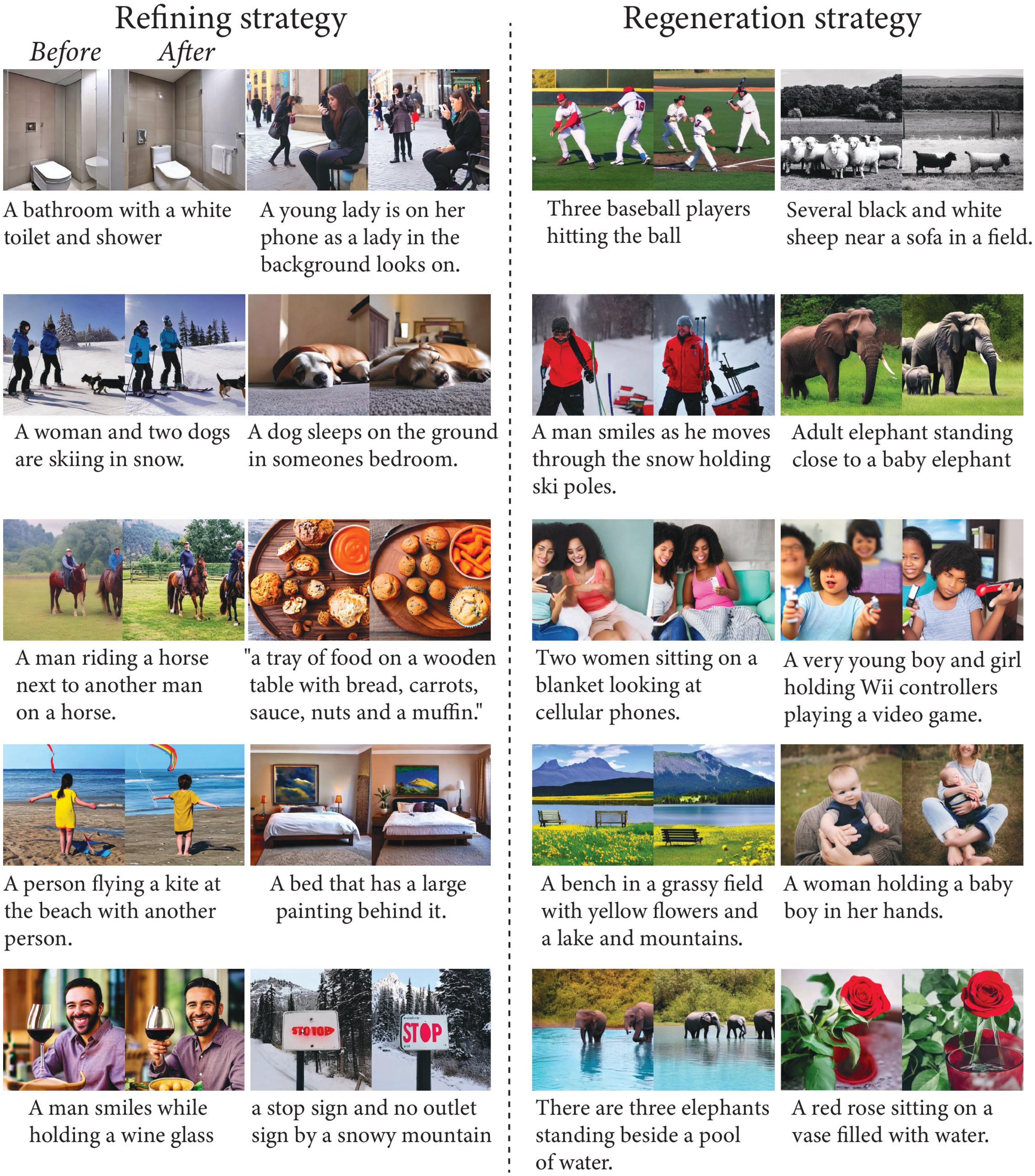


Figure 25. Visualization of the student (CD-SDXL) and teacher (SDXL) samples for low and high distance ranges. The green outline corresponds to wins, while the red one - losses.



A bathroom with a white toilet and shower

A young lady is on her phone as a lady in the background looks on.

A woman and two dogs are skiing in snow.

A dog sleeps on the ground in someones bedroom.

A man riding a horse next to another man on a horse.

"a tray of food on a wooden table with bread, carrots, sauce, nuts and a muffin."

A person flying a kite at the beach with another person.

A bed that has a large painting behind it.

A man smiles while holding a wine glass

a stop sign and no outlet sign by a snowy mountain

Three baseball players hitting the ball

Several black and white sheep near a sofa in a field.

A man smiles as he moves through the snow holding ski poles.

Adult elephant standing close to a baby elephant

Two women sitting on a blanket looking at cellular phones.

A very young boy and girl holding Wii controllers playing a video game.

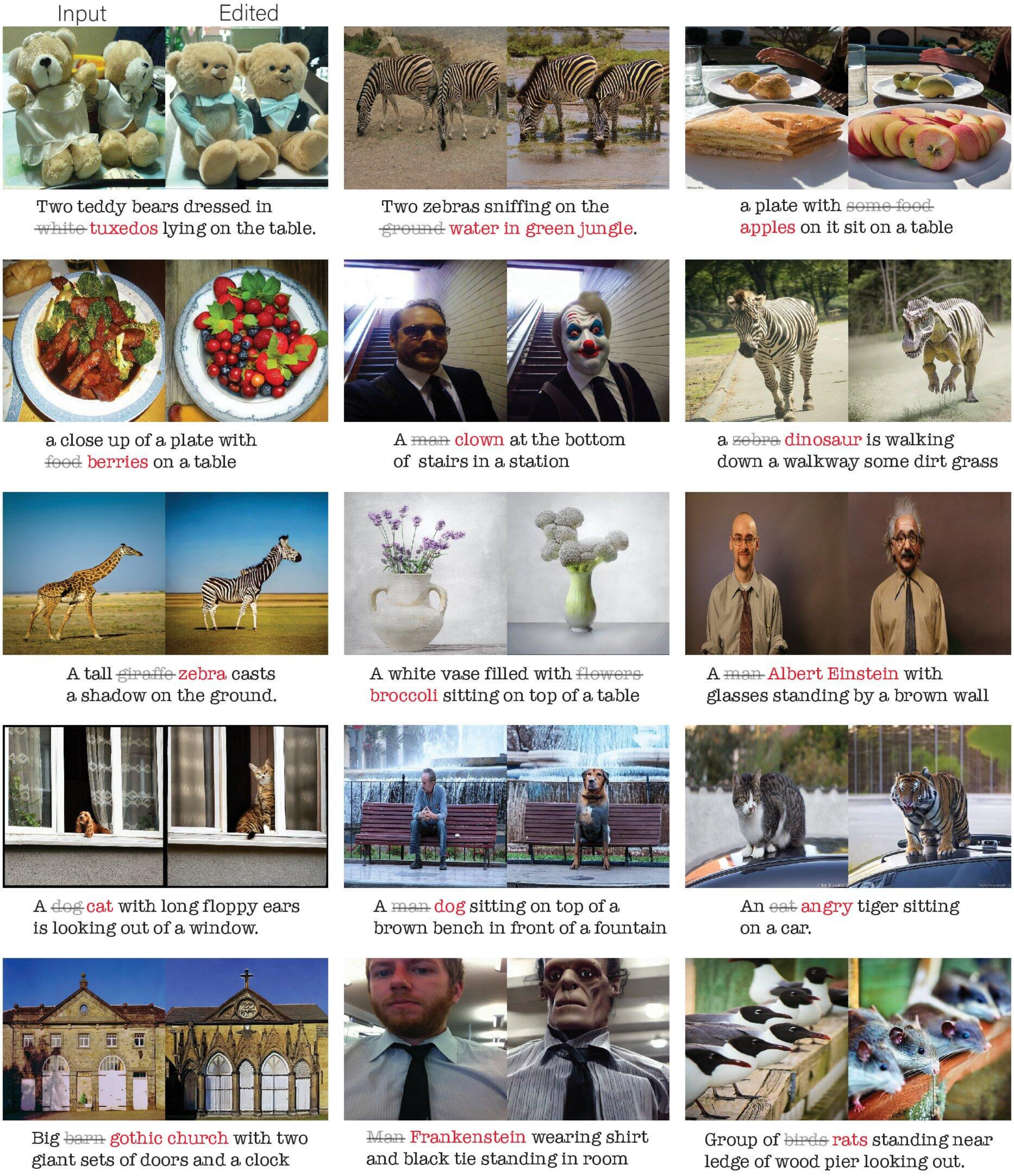
A bench in a grassy field with yellow flowers and a lake and mountains.

A woman holding a baby boy in her hands.

There are three elephants standing beside a pool of water.

A red rose sitting on a vase filled with water.

Figure 26. Visual examples of the refining and regeneration adaptive strategies.



Two teddy bears dressed in ~~white~~ tuxedos lying on the table.

Two zebras sniffing on the ~~ground~~ water in green jungle.

a plate with ~~some food~~ apples on it sit on a table

a close up of a plate with ~~food~~ berries on a table

A ~~man~~ clown at the bottom of stairs in a station

a ~~zebra~~ dinosaur is walking down a walkway some dirt grass

A tall ~~giraffe~~ zebra casts a shadow on the ground.

A white vase filled with ~~flowers~~ broccoli sitting on top of a table

A ~~man~~ Albert Einstein with glasses standing by a brown wall

A ~~dog~~ cat with long floppy ears is looking out of a window.

A ~~man~~ dog sitting on top of a brown bench in front of a fountain

An ~~cat~~ angry tiger sitting on a car.

Big ~~barn~~ gothic church with two giant sets of doors and a clock

~~Man~~ Frankenstein wearing shirt and black tie standing in room

Group of ~~birds~~ rats standing near ledge of wood pier looking out.

Figure 27. Additional image editing results produced with our approach.

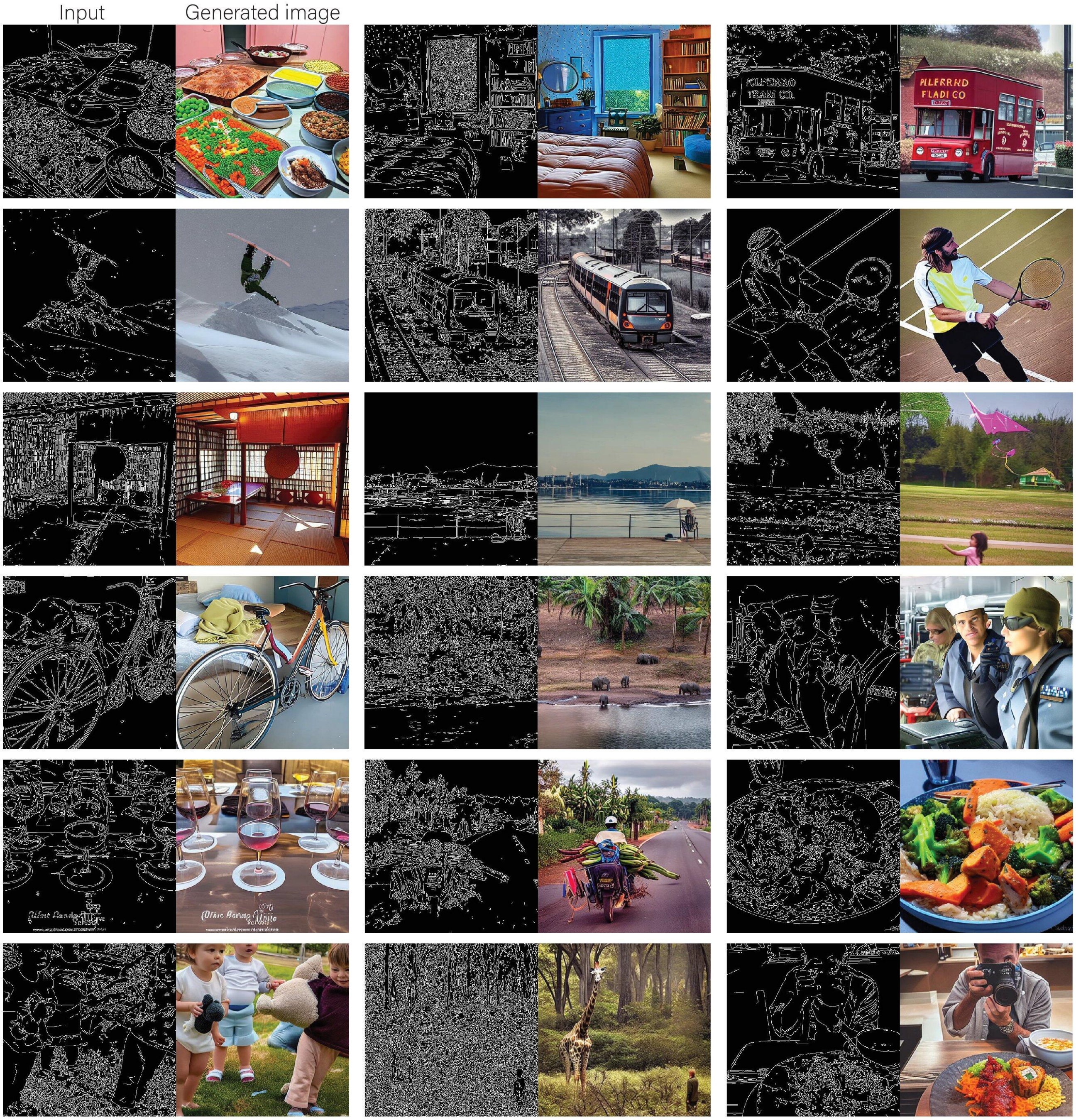


Figure 28. Additional results on Canny edge guided image generation with our approach.

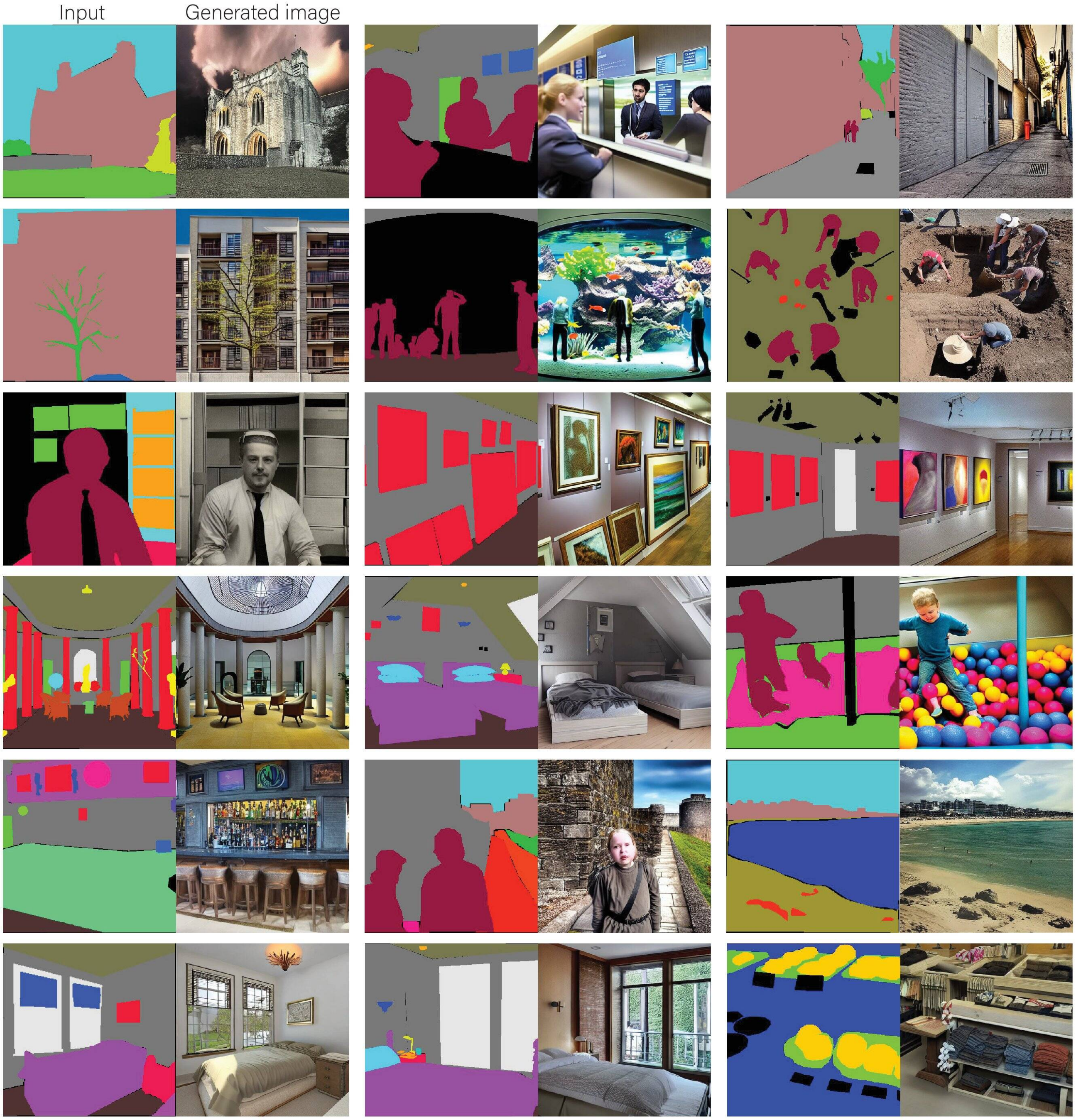


Figure 29. Additional results on segmentation mask guided image generation with our approach.

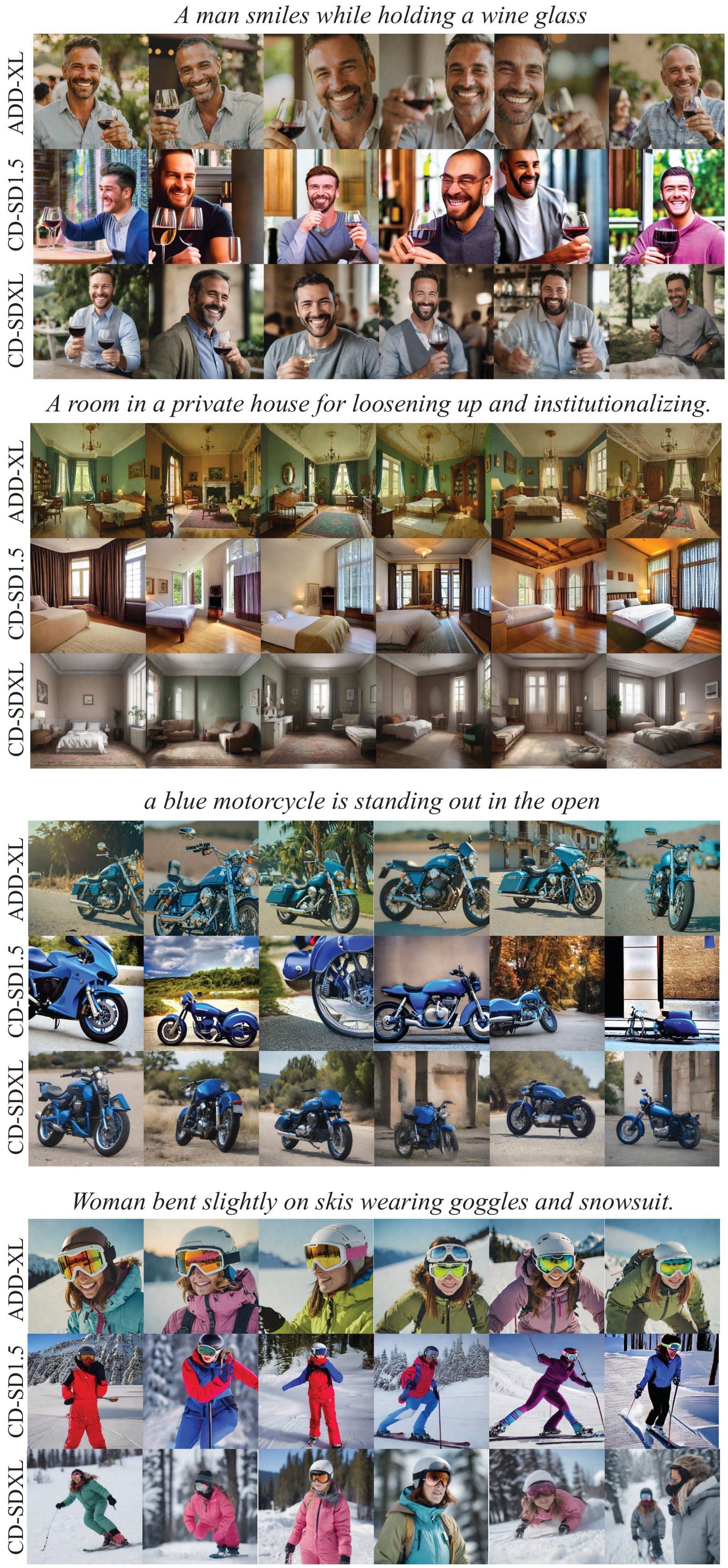

A man smiles while holding a wine glass
ADD-XL
CD-SD1.5
CD-SDXL
A room in a private house for loosening up and institutionalizing.
ADD-XL
CD-SD1.5
CD-SDXL
a blue motorcycle is standing out in the open
ADD-XL
CD-SD1.5
CD-SDXL
Woman bent slightly on skis wearing goggles and snowsuit.
ADD-XL
CD-SD1.5
CD-SDXL

Figure 30. Visual examples generated with various distilled text-to-image models for different seed values. CD-based students generate more diverse images than ADD-XL.