CONFORM: Contrast is All You Need For High-Fidelity Text-to-Image Diffusion Models

Supplementary Material

7. Benchmark Sets

In our qualitative analysis approach, we adopt benchmark sets from previous studies [6, 22]. Our benchmark protocol includes structured prompts such as 'an [animalA] and an [animalB]', 'an [animal] with a [color] [object]', and 'a [colorA] [objectA] and a [colorB] [objectB]', along with multi-instance prompts. Table 4 provides details on the benchmark sets and the number of prompts for each set. We test each prompt with 64 unique seeds, conducting 50 iterations per seed. The testing is performed using Stable Diffusion v1.5 and typically takes about 20 seconds per prompt on an NVIDIA L4 GPU.

Table 4. Description of benchmark sets and number of prompts used for qualitative evaluation.

Benchmark Set	Template	#
Animal-Animal	a [animalA] and a [animalB] *'a **horse** and a **bird**'*	66
Animal-Object	a [animal] and a [color][object] *'a **frog** and a **purple balloon**'*	144
Object-Object	a [colorA][objectA] and a [colorB][objectB] *'a **black crown** and a **red car**'*	66
Multi-Object	[numberA][animalA] and [numberB][animalB] *'**one zebra** and **two birds**'*	30

8. Algorithm

The pseudocode for the CONFORM algorithm is described in Algorithm 1.

Algorithm 1 A Single Denoising Step using CONFORM.

Input: A text prompt $\mathcal{P}$, previous attention maps A_{t+1}, a dictionary of subject and corresponding attribute token indices $\mathcal{T}$, a timestep t, a set of iterations for refinement $\{t_1, \ldots, t_k\}$, and a pretrained diffusion model ϵ_θ.
Output: A noised latent z_{t-1} for the next timestep.

$\_, A_t \leftarrow \epsilon_\theta(z_t, \mathcal{P}, t)$
$\mathrm{L} \leftarrow \{\}$ ▷ Pseudo-Labels
$\mathrm{F} \leftarrow \{\}$ ▷ Features
for $s_i, \mathcal{C} \in \mathcal{T}$ **do** ▷ Prepare features and labels
  $\mathrm{L} \leftarrow \mathrm{L} + i$ ▷ Label attention map
  $\mathrm{F} \leftarrow \mathrm{F} + A_t[:, :, s_i]$ ▷ Get attention map
  $\mathrm{L} \leftarrow \mathrm{L} + i$ ▷ Label prev attention map
  $\mathrm{F} \leftarrow \mathrm{F} + A_{t+1}[:, :, s_i]$ ▷ Get prev attention map
  for $c_j \in \mathcal{C}$ **do** ▷ Attributes of a token get the same labels
    $\mathrm{L} \leftarrow \mathrm{L} + i$ ▷ Label attention map
    $\mathrm{F} \leftarrow \mathrm{F} + A_t[:, :, c_j]$ ▷ Get attention map
    $\mathrm{L} \leftarrow \mathrm{L} + i$ ▷ Label prev attention map
    $\mathrm{F} \leftarrow \mathrm{F} + A_{t+1}[:, :, c_j]$ ▷ Get prev attention map
  end for
end for
$\mathcal{L} \leftarrow \mathcal{L}(\mathrm{F}, \mathrm{L})$
$z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$
if $t \in \{t_1, \ldots, t_k\}$ **then** ▷ If performing iterative refinement at t
  $z_t \leftarrow z'_t$
  Go to Step 1
end if
$z_{t-1}, \_ \leftarrow \epsilon_\theta(z'_t, \mathcal{P}, t)$
Return z_{t-1}

9. Ablation Study

In our ablation study, we conducted targeted experiments on a limited set of prompts and a smaller number of images to identify the most effective parameters and techniques for our final model, guided by CLIP similarity metrics. The most influential parameters were found to be: incorporating attention maps from previous iterations, fine-tuning the temperature parameter in the contrastive loss equation (referenced in Eq. 3), implementing refinement steps for iterative updates of latent before advancing to subsequent iterations, and determining whether to consistently apply optimization throughout all iterations or to stop it at a certain point.

Using Attention Maps from Previous Iterations. As indicated in Tab. 5, employing attention maps from the previous iteration positively impacts both CLIP-full and CLIP-min metrics. The benefits of this technique are also evident in Fig. 3 (see main paper). While the diffusion process naturally leads to the scattering of attention, our method assists in maintaining focused attention throughout the iterations, countering the scattering effect.

Table 5. Average CLIP image-text similarities for ablation study on the effect of using previous iteration attention maps.

Method	CLIP-full	CLIP-min
Stable Diffusion	0.33	0.24
w/o Previous Iteration Attention Maps	0.35	0.25
CONFORM	**0.36**	**0.26**

Scale Factor, Temperature, Refinement Steps, Maximum Optimization Steps The temperature parameter (τ) in Eq. 3 (see main paper) is critical for achieving desirable results and requires precise calibration. We utilized a grid search ($\tau \in \{0.25, 0.5, 0.75, 1.0\}$) to identify the optimal temperature parameter and the number of refinement steps at specific iterations. Additionally, we determined the appropriate point to cease optimization and the scale factor α, as defined in Eq. 5. The selection was based on achieving the highest CLIP-full and CLIP-min similarity scores.

Beyond quantitative analysis, we carefully observed how each parameter influenced the final image output. Values of τ that are too high result in negligible changes, while values that are too low lead to unwanted artifacts. We determined $\tau = 0.5$ as the optimal setting. Implementing refinement steps at certain iterations, specifically at $i \in 0, 10, 20$ (similar to the approach in [6]), allows the model to make necessary adjustments to the attention maps. However, applying optimization at every step led to unwanted artifacts in the output. Consequently, we decided to halt optimization at $i = 25$ to avoid these issues. Lastly, we also used scale factor $\alpha = 20$ considering the results of the grid search.

10. Additional Qualitative Results

In the remainder of the Supplementary Materials, we provide additional qualitative comparisons:

- Comparison with the A-Star for the images presented in [1].
- Additional prompts involving person-person, person-object, and person-background on SD-based models: Stable Diffusion [34], Attend & Excite [6], Divide & Bind [22], and our method applied to the Stable Diffusion model.
- Additional qualitative comparisons for SD-based models: Stable Diffusion [34], Attend & Excite [6], Divide & Bind [22], and our method applied to the Stable Diffusion model.
- Additional comparisons between Imagen [37] and our method applied to the original Imagen model.

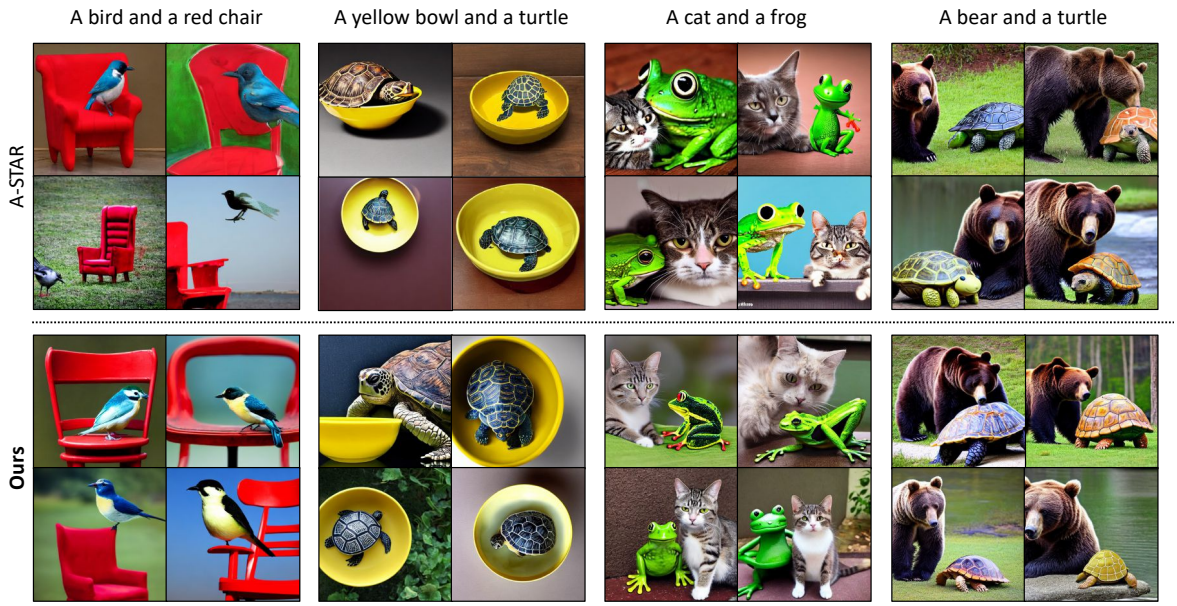

Figure 9. **Qualitative comparison of CONFORM with A-Star.** We compared the results presented in A-Star [1] paper with our method on Stable Diffusion.

Figure 10. **Qualitative comparison of CONFORM on SD with prompts involving person-person, person-object, and person-background.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

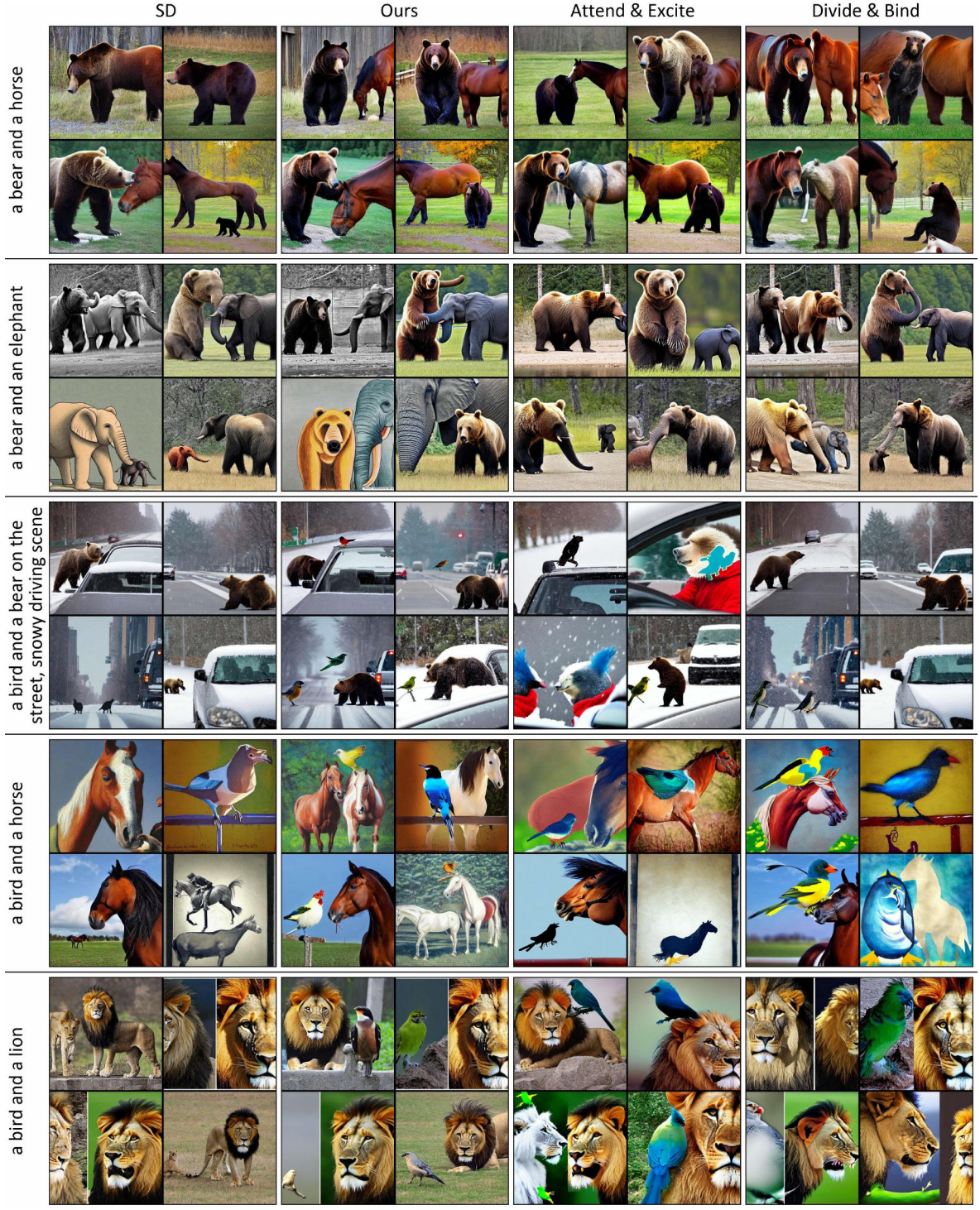

Figure 11. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

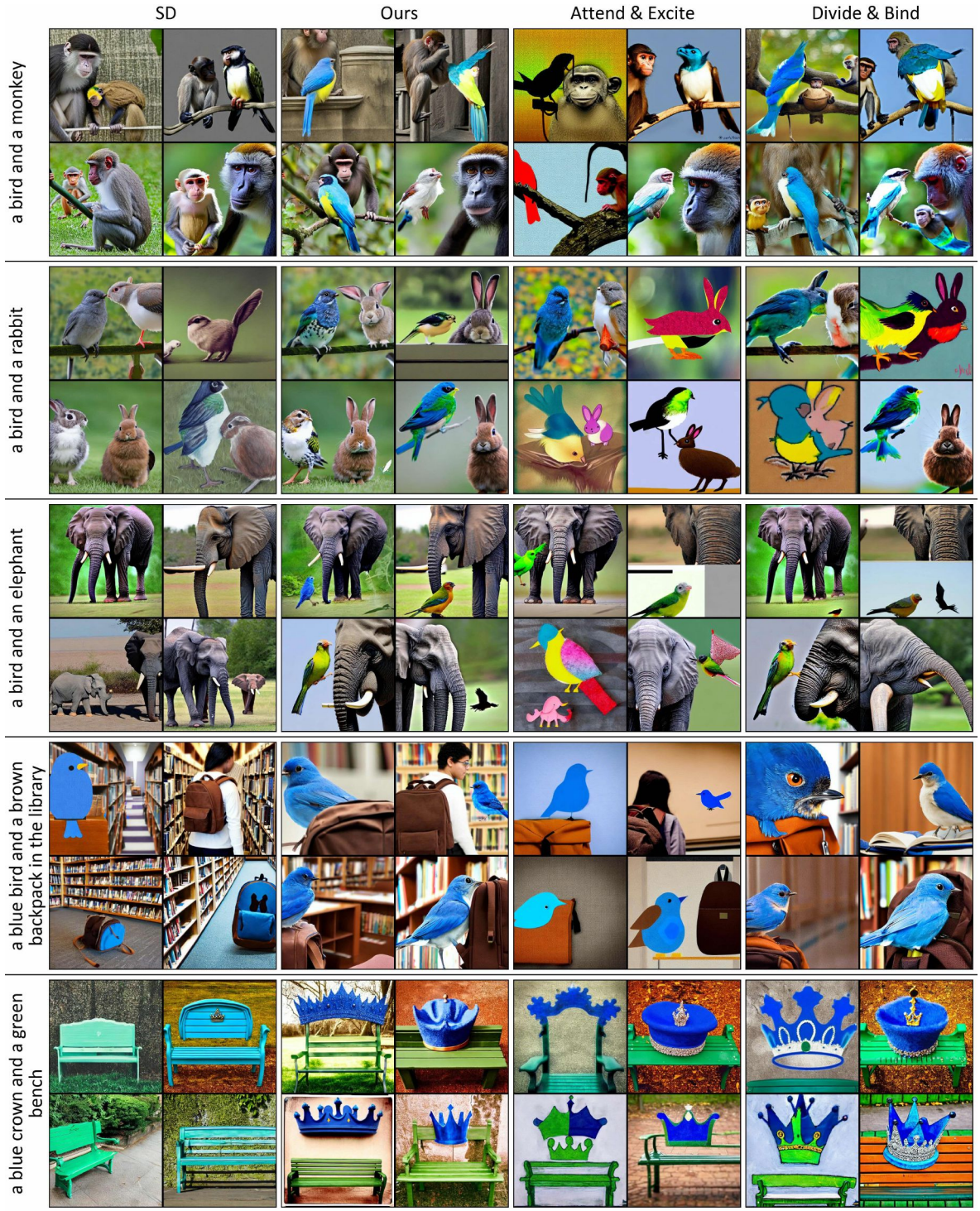

Figure 12. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

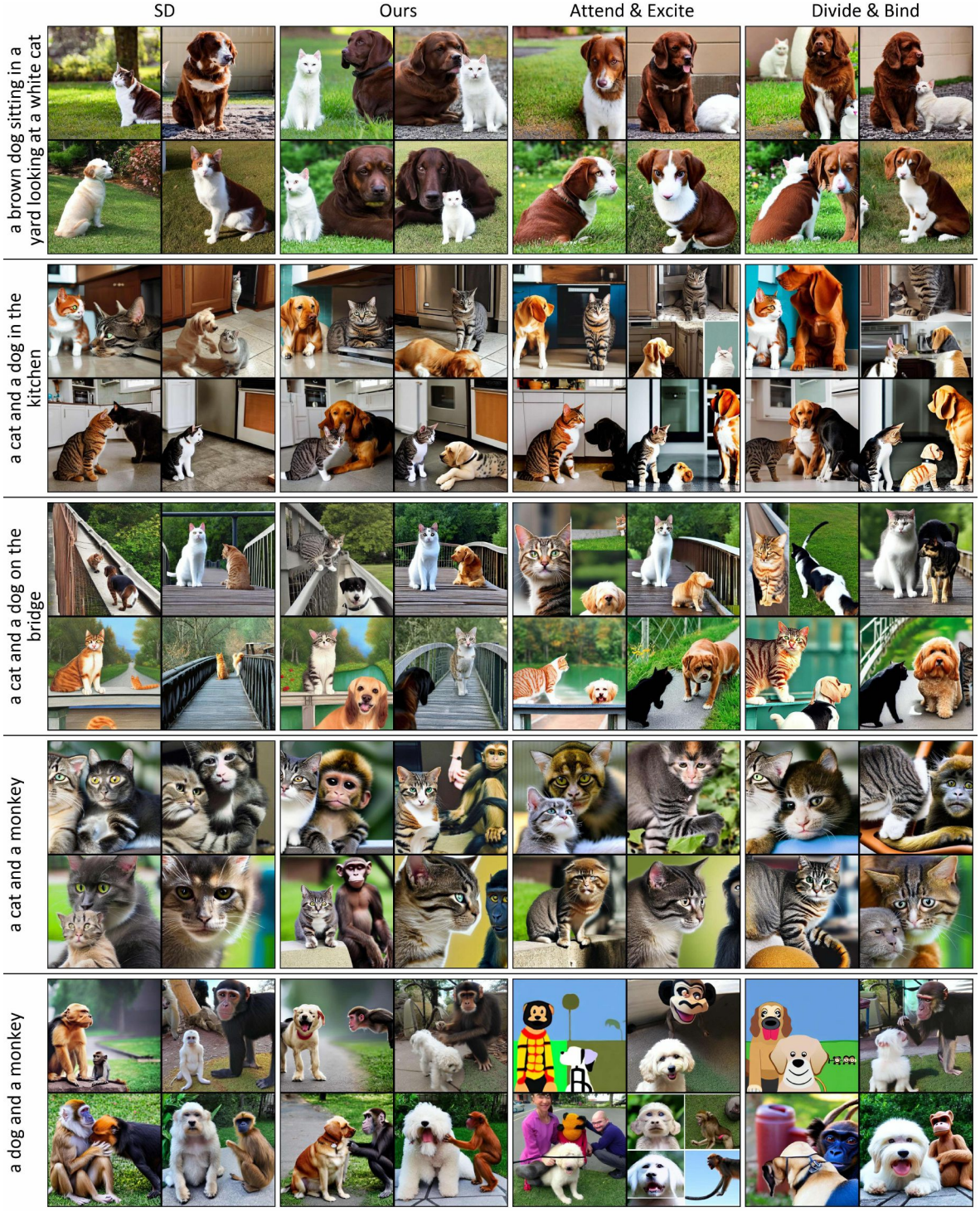

Figure 13. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

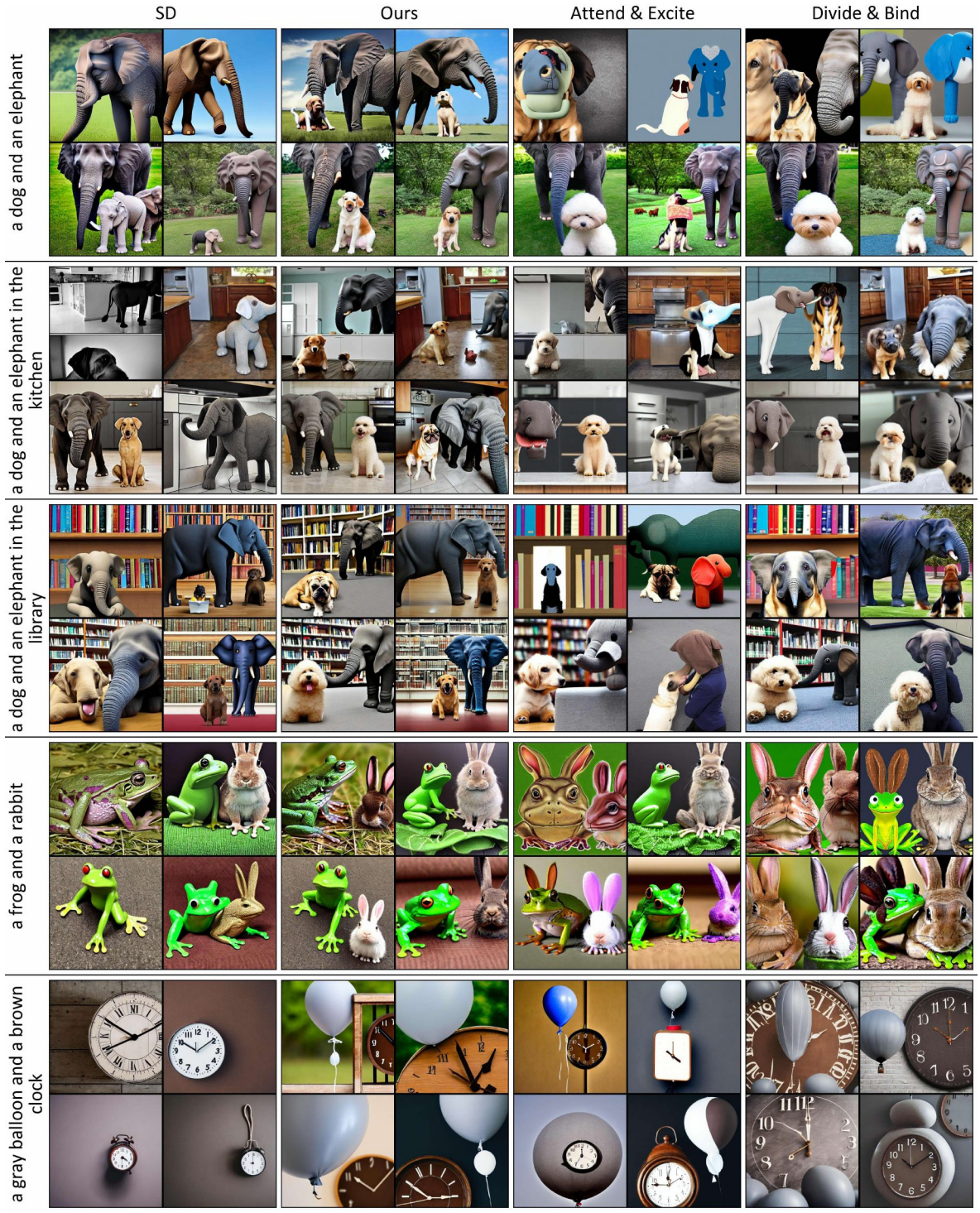

Figure 14. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

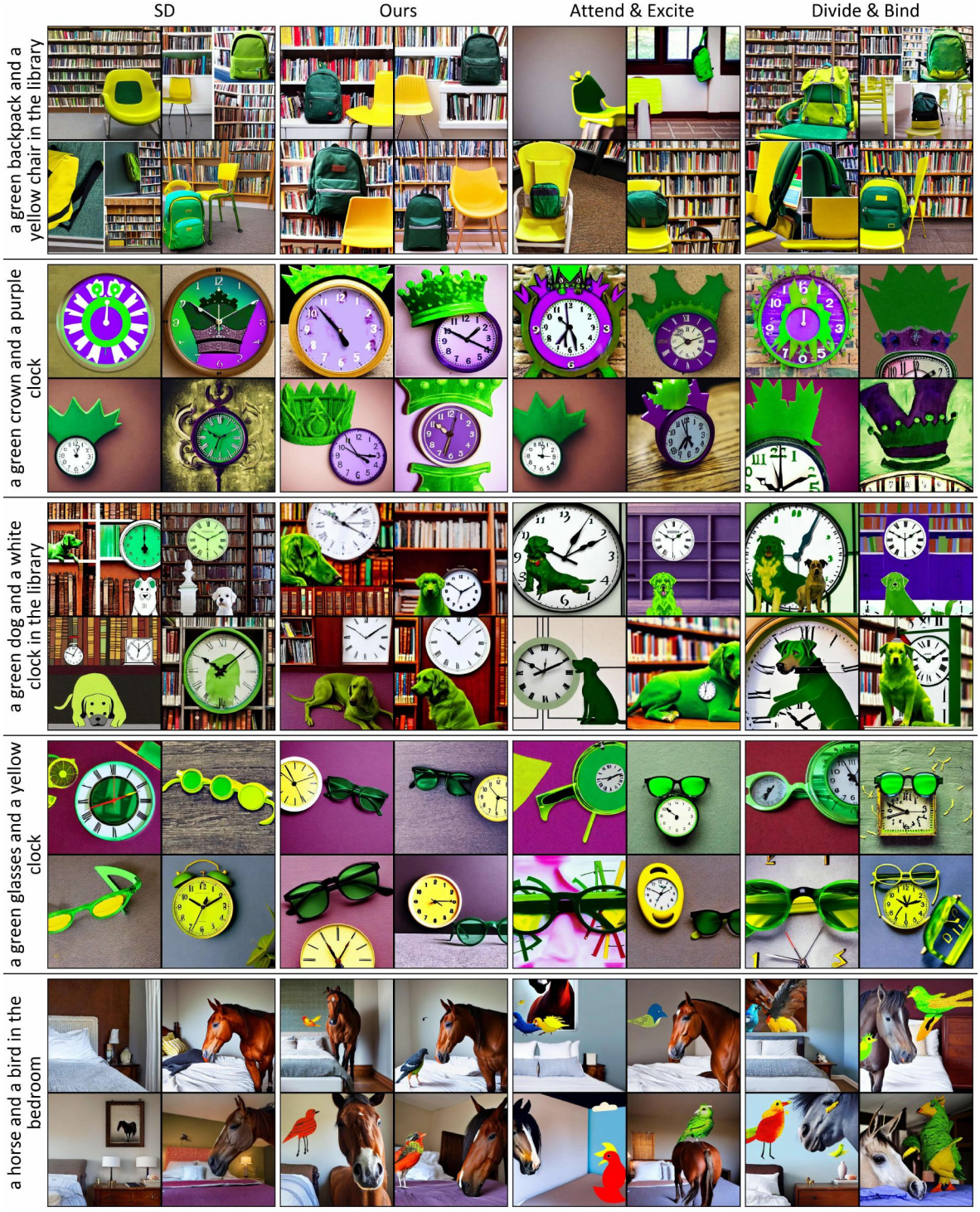

Figure 15. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

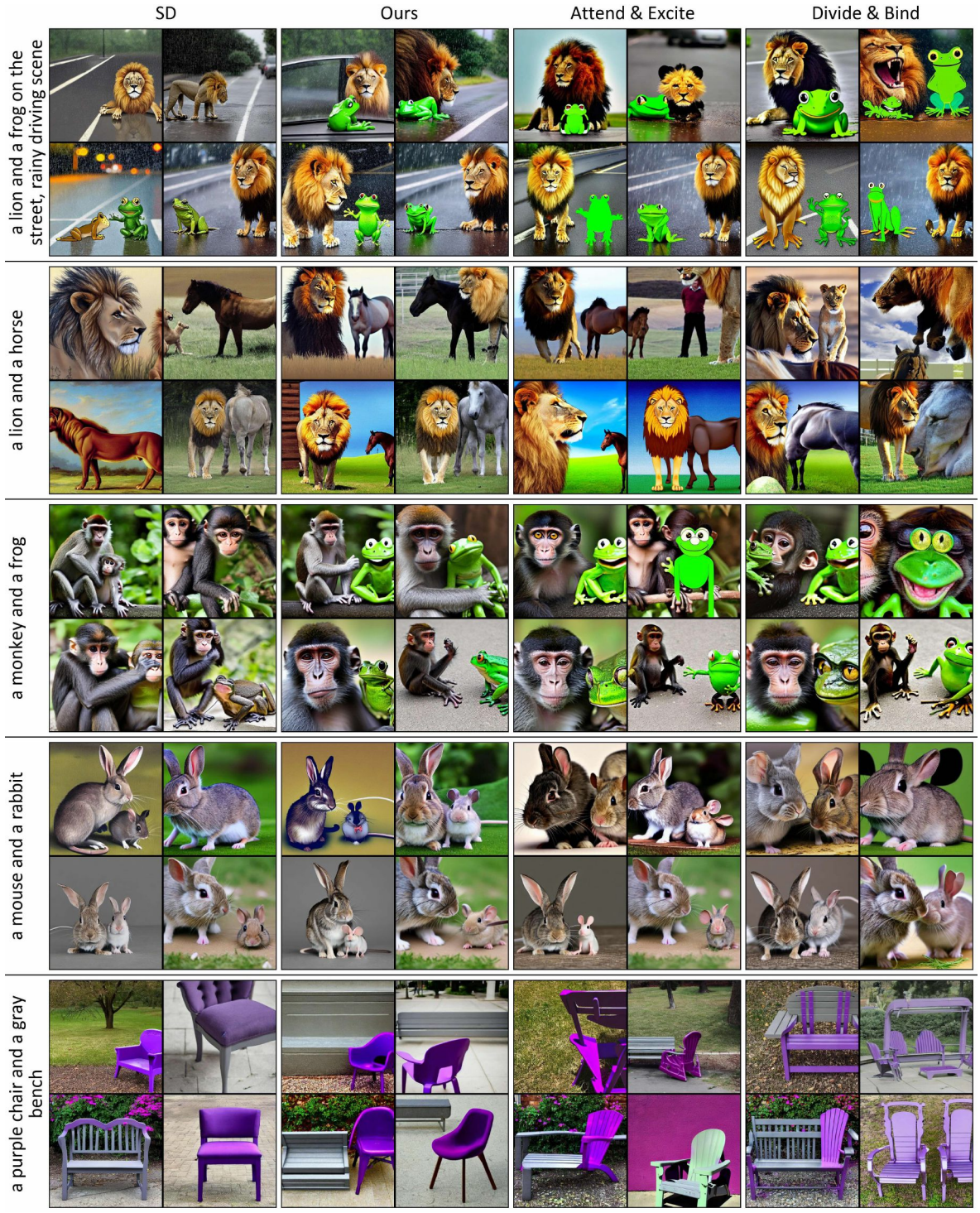

Figure 16. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

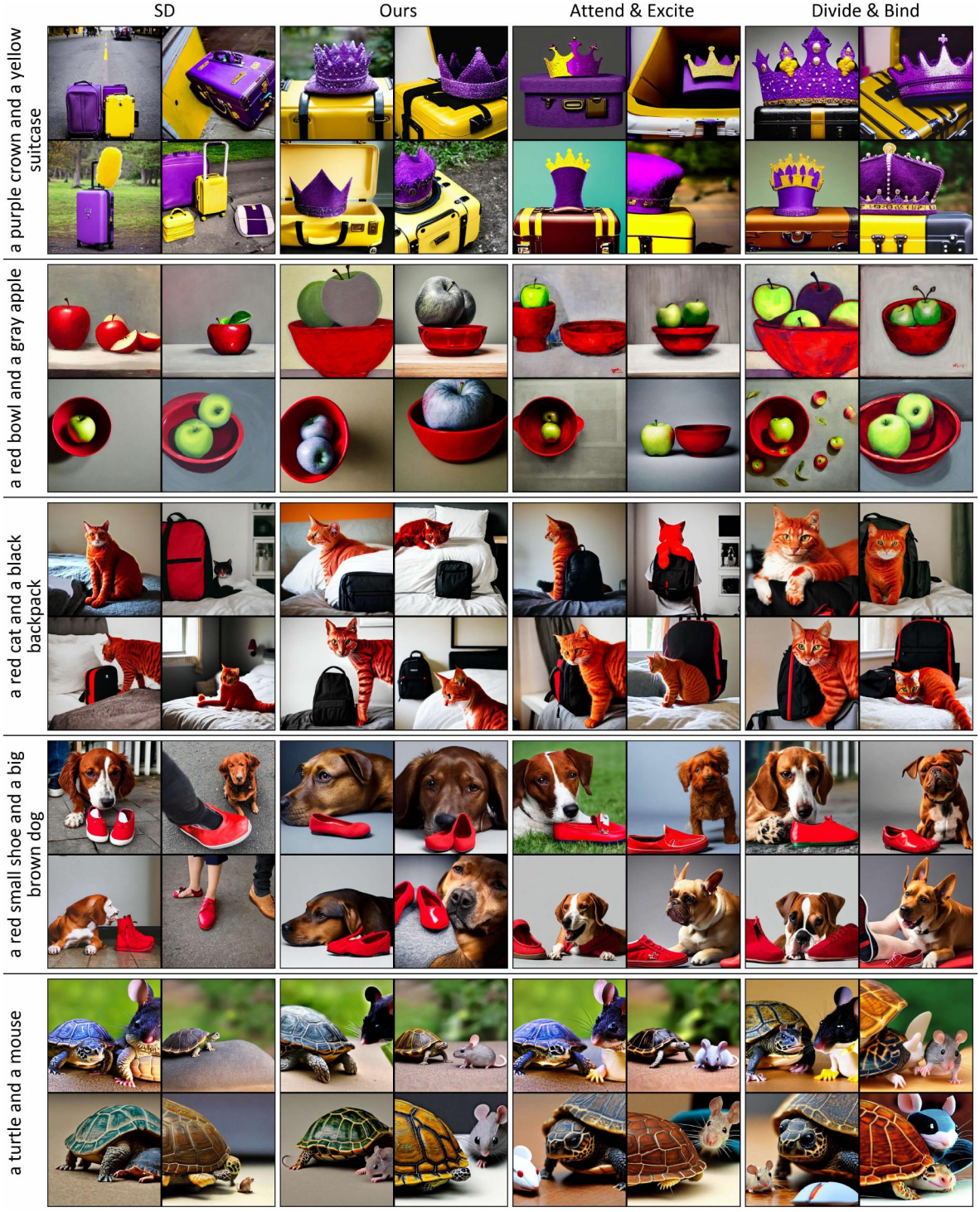

Figure 17. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

Figure 18. **Qualitative comparison of CONFORM on SD.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the SD model.

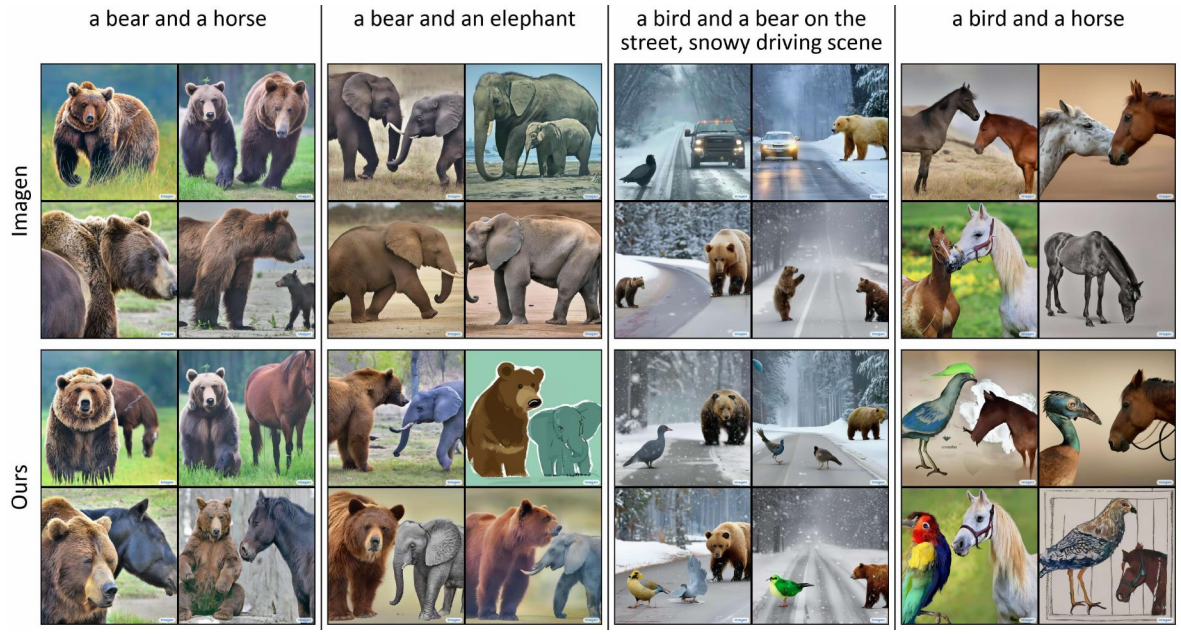

Figure 19. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

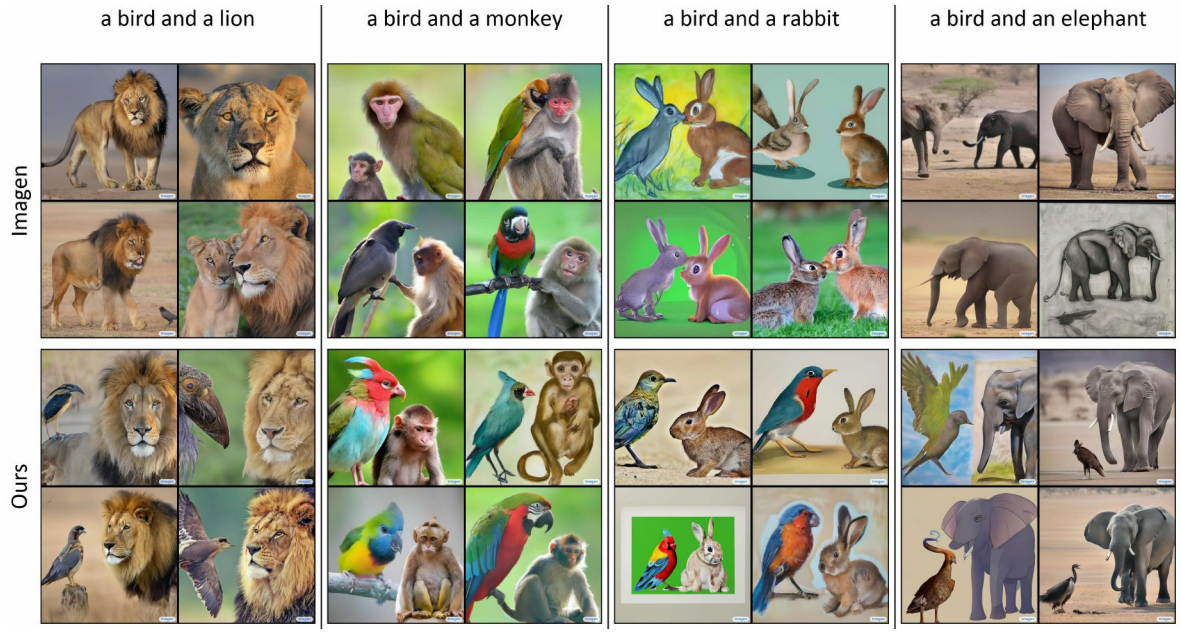

Figure 20. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

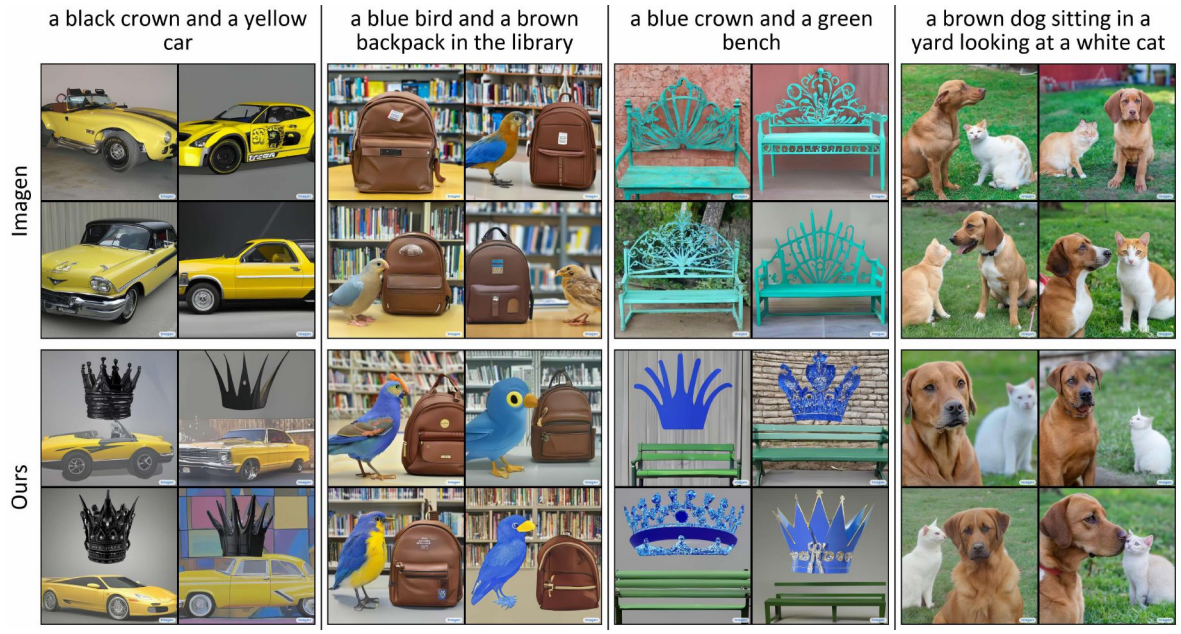

Figure 21. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

Figure 22. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

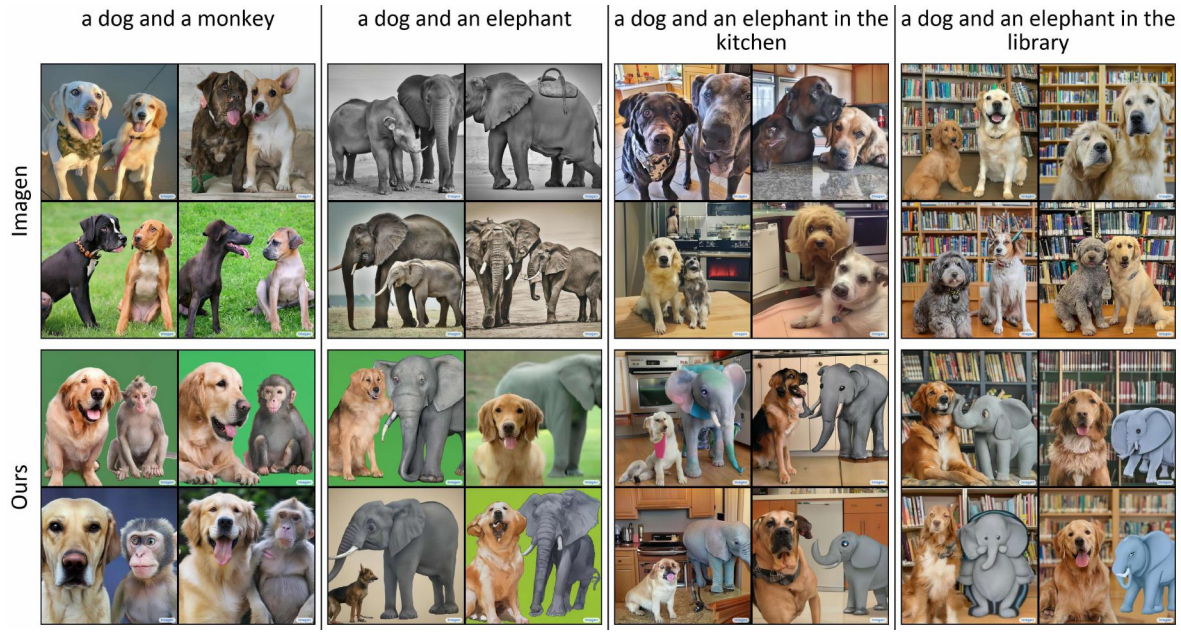

Figure 23. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

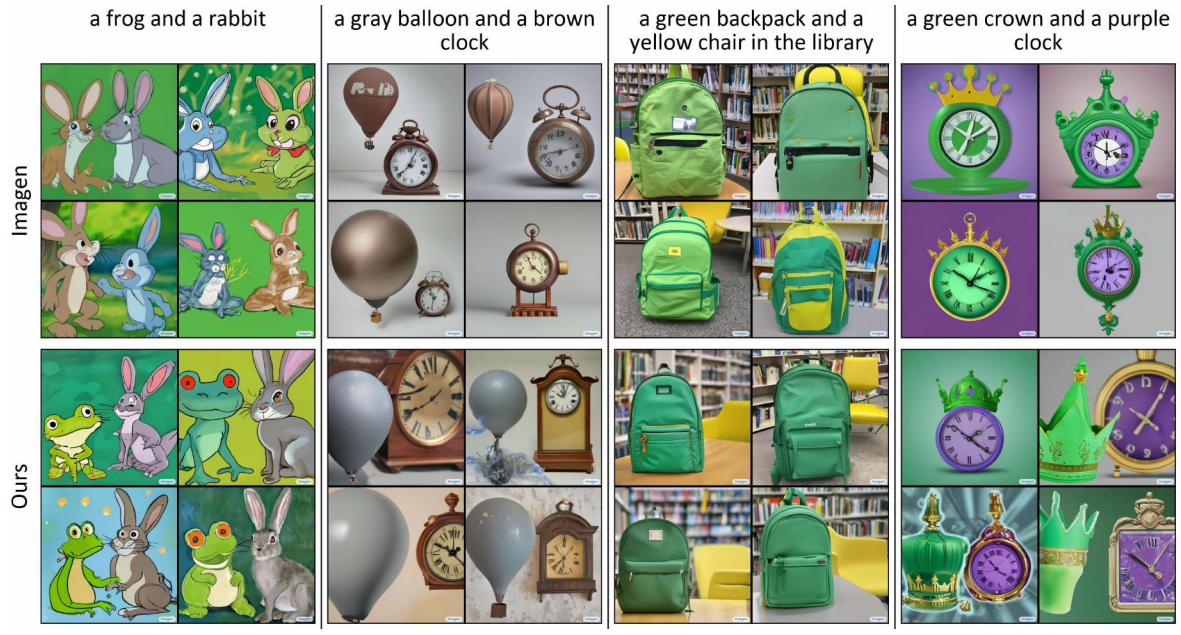

Figure 24. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

Figure 25. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

Figure 26. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

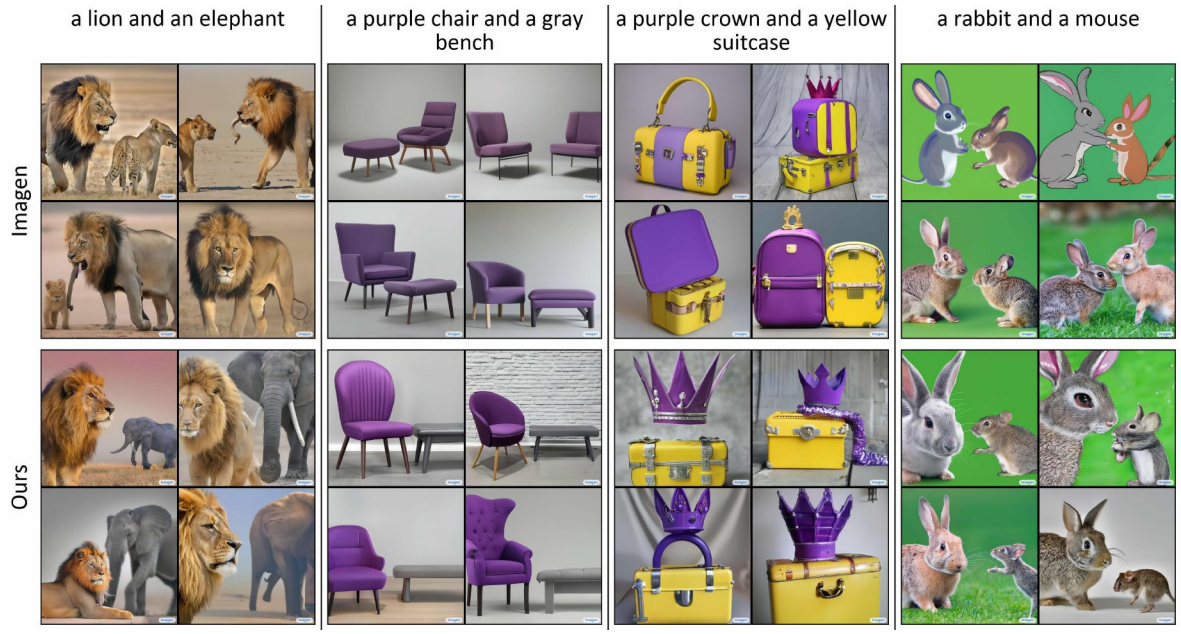

Figure 27. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

Figure 28. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

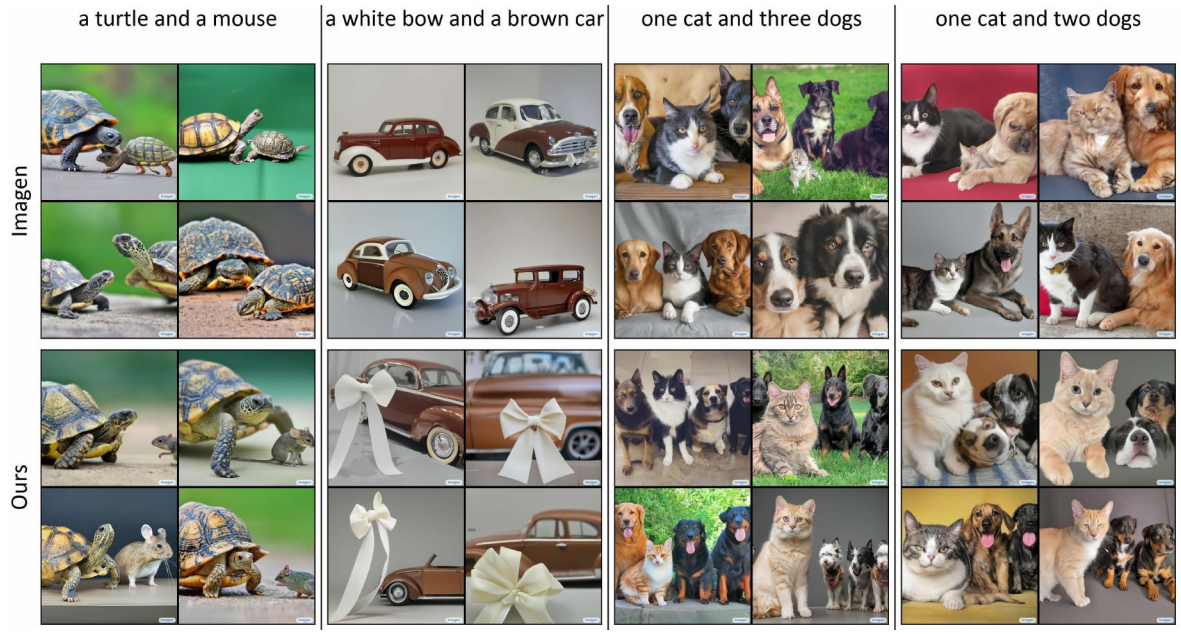

Figure 29. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

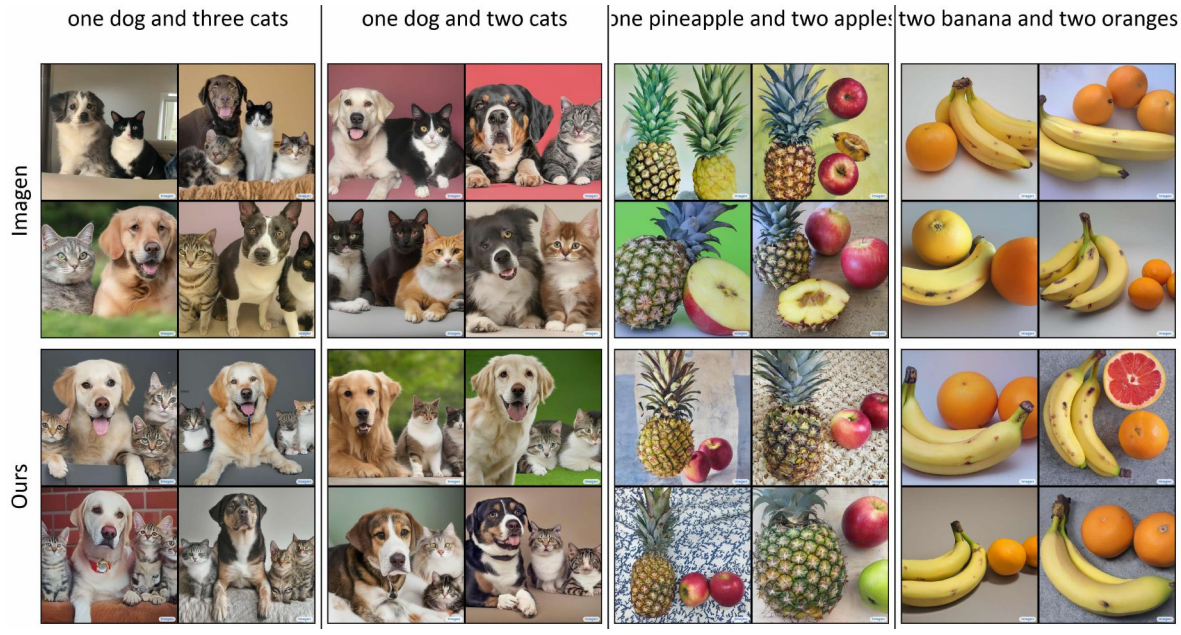

Figure 30. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.