Guiding Diffusion Models with Adaptive Negative Sampling Without External Resources

Supplementary Material

1. Datasets and Metrics

1.1. Datasets

ImageNet Dataset: contains images with 1000 classes. Since 2010, the dataset has been used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7], a benchmark in image classification and object detection. We use all 1000 classes and create 5 images per class, a total of 5000 images for this dataset. We calculate the FID using 5000 images created by `ANSWER` and 10000 ground truth images from ImageNet.

Attend and Excite (A&E) Dataset: was introduced by [1] and focuses on entity neglect and attribute assignment. Each prompt in the dataset comprises two entities and associated attributes. There are three prompt categories: (1) "an animal and an animal" (2) "a color object and an animal" (3) "a color object and a color object". We sample equally from each category, totaling 100 prompts for 24 random seeds.

Pick-a-Pic Dataset: is sourced from the tracking of prompts used by the users of the Pick-a-Pic web application. There are 500,000 examples across 35000 unique prompts that were used to train the PickScore [5]. The prompts in the dataset comprise concepts like color, style, text, multiple objects, spatial locations, and numeracy. We sample a total of 100 prompts from the test set for 24 random seeds.

DrawBench Dataset: is a diverse and comprehensive benchmark that was introduced by the Imagen team [8]. The dataset contains 11 categories, such as accurate color, object counting, spatial relations, text rendering, and complex interactions between objects. Another feature of this benchmark is the inclusion of vocabulary that is rarely used and scenarios that are imaginative or unrealistic. We use all 200 of test prompts to evaluate for 24 random seeds. CLIP does not interpret misspelled and rare words as sensible tokens, causing all models to perform poorly on them. Therefore, we removed them from the evaluation.

PartiPrompts Dataset [12]: is a dataset used to measure model capabilities across various categories such as artifacts, vehicles, people, food, etc. and challenge aspects such as simple, detail, complex, imagination, etc.. We sample a total of 500 prompts (spanning several categories) from the test set for 24 random seeds.

In total, we have generated about 26000 images (comprising all datasets, prompts, and seeds) per method (SDXL (`CFG`), SDXL+`DNP`, SDXL+`ANSWER`).

1.2. Metrics

CLIP Score: [3] can be used to measure the alignment between an image and a text prompt. For the A&E dataset, we follow the evaluation protocols of [1], which measure scores at prompt and entity levels. *Full Prompt* is the CLIP Score between the image and full prompt. *Minimum Object*, is the minimum CLIP Score across prompt entities, highlighting the most neglected entity (lowest score).

Fréchet inception distance (FID) [4]: is a common metric, first introduced in 2017, to assess the quality of images created by generative models. FID compares the distribution of generated images with the distribution of a set of real images (a "ground truth" set). A lower FID implies higher diversity. While useful, FID cannot be used without a large ground truth set.

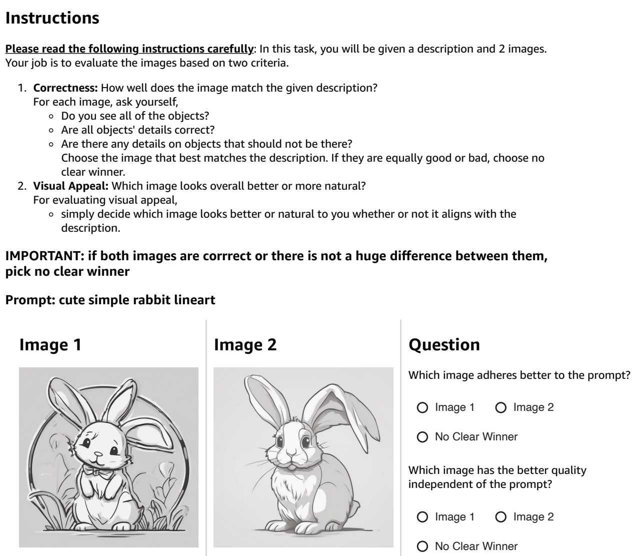
Inception Score (IS) [9]: are commonly used to measure the naturalness or quality of generated images. Mathematically, IS is the exponential of the average entropy of the label distribution predicted by the Inception v3 classifier. While FID also provides a similar measure, we choose IS over FID as it does not require a dataset of real images.

Human Evaluation: While they are good automated measures, high CLIP Scores and IS do not necessarily align with human aesthetics and preferences. We use Amazon Mechanical Turk (AMT) to gauge human preference while comparing the models. For all datasets, we perform the human evaluation on 100-200 prompt-seed pairs. For each image pair, the Turkers were given two multiple-choice questions: 1) **Correctness:** pick the images based on "correctness" or prompt adherence alone while ignoring the quality. 2) **Visual Quality:** disregard the "correctness" and pick the most natural or realistic image. The set of choices for both questions was {Image-1, Image-2, Image-3, *'No clear winner'*}. For evaluating each pair, 10 unique master Turkers with an approval rate exceeding 95% were employed. We also randomly swapped images to avoid human bias. A sample screenshot of the AMT task for two images can be seen in Figure 1. The same instructions were used for all the experiments.

Human Preference Metrics While human evaluation remains the gold standard, it is costly and not always feasible. Therefore, we also assess the generated images using newer metrics that approximate human preference and act as practical alternative metrics, including *ImageReward* [11], *PickScore* [5] and *HPSv2* [10]. They are trained to closely mimic human preferences. We also report win-rate percentages for all human preference metrics, offering additional insights into generation quality.

2. Implementation Details

Computing Resources: All experiments were run on NVIDIA GeForce RTX A4000 with 16GB RAM using PyTorch with *Accelerate*.
We use Stable Diffusion v1.5 and Stable Diffusion XL (*bfloat16*) as the diffusion models for the experiments in the paper. All the images were generated with denoising steps $T = 41$. The guidance scale used for the SD model was 7.5, and SDXL was 5.0, which are the default values for respective models. For running DNP [2], we used equal positive and negative guidance scales and used Blip2 [6] as the captioner. For ANSWER, we use $K = 5$ and negative guidance scale, $s_{neg} = 2.5$ was used.


Instructions

Please read the following instructions carefully: In this task, you will be given a description and 2 images. Your job is to evaluate the images based on two criteria.

1. **Correctness:** How well does the image match the given description?
   For each image, ask yourself,
   - Do you see all of the objects?
   - Are all objects' details correct?
   - Are there any details on objects that should not be there?
     Choose the image that best matches the description. If they are equally good or bad, choose no clear winner.
2. **Visual Appeal:** Which image looks overall better or more natural?
   For evaluating visual appeal,
   - simply decide which image looks better or natural to you whether or not it aligns with the description.

IMPORTANT: if both images are corrrect or there is not a huge difference between them, pick no clear winner

Prompt: cute simple rabbit lineart

Image 1 **Image 2** **Question**

Which image adheres better to the prompt?

○ Image 1 ○ Image 2

○ No Clear Winner

Which image has the better quality independent of the prompt?

○ Image 1 ○ Image 2

○ No Clear Winner

Figure 1. Instructions provided to the MTurkers

3. Additional Results

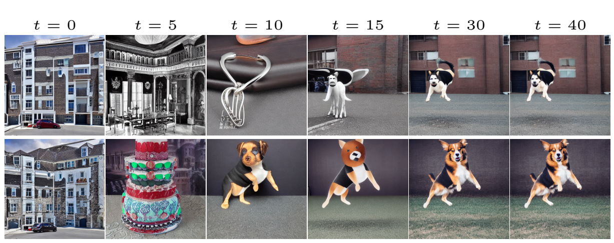


Figure 2. Negative images produced using DNS for the latent z_t of CFG (Top) and ANSWER (Bottom) chain, for the prompt **p** = `dog jumping in the air`, at different time steps t for T=40.

3.1. Convergence strength of ANSWER

In Figure 2, we show the negative images for CFG (top) and ANSWER (bottom) chain. ANSWER converges faster ($t = 10$) than CFG ($t = 30$). This illustrates the strength of ANSWER in enforcing prompt adherence.

Guidance Scale (s)	CLIP	Image Reward	Pick Score	HPSv2
3	75.08%	72.42%	75.66%	75.34%
5	68.42%	68.75%	69.33%	70.17%
7	63.33%	65.08%	61.83%	64.83%
10	62.33%	63.08%	58.75%	65.67%

Table 1. Change in % Winrate with positive guidance scale (s) for Pick-a-Pick dataset.

3.2. Guidance Scale Ablation

To analyze the effect of guidance scale s of the standard CFG positive chain, we perform an ablation shown in Table 1. For each guidance scale and metric, we show the win rate of SDXL+ANSWER over SDXL (CFG). The negative guidance scale is fixed at 3.5, and the number of DNS steps is fixed at $K = 5$. We observe that as the guidance increases, CFG's prompt adherence increases and the win rate deteriorates. However, this drop is minimal, and SDXL+ANSWER outperforms SDXL (CFG) across all guidance scales. At scales higher than 10, CFG is known to have quality issues. ANSWER can increase the prompt adherence without the loss in quality associated with higher guidance.

3.3. Qualitative Results

In this section, we show additional qualitative results for all the datasets with SDXL as the baseline. Figure 3, 4, 5 show qualitative results for A&E DrawBench and Pick-a-Pic datasets respectively. Drawbench and Pick-a-Pic contain complex, imaginative, and abstract prompts. We observe that while SDXL is hesitant about generating imaginative scenarios, SDXL+ANSWER generates images that truly align with the prompt. For example, SDXL sets the wine glass next to the dog instead of on top, despite explicit prompting. ANSWER ensures better adherence to structure, layout, numeracy, and text rendering while generating high-quality images. Overall, we observe that ANSWER tends to make more realistic and higher quality images over SDXL.

3.4. Comparing DNP Vs ANSWER

Figure 6, shows visual comparison between Stable Diffusion (SD), SD+DNP and SD+ANSWER on A&E dataset. SD consistently suffers from missing objects, merged objects, and misaligned attributes. The addition of DNP often solves the issue of missing objects and misalignment. For example, it introduces 'cat' in "a cat and a red balloon" and 'backpack' in "a horse and a purple backpack." However, it does not always succeed. For example, the backpack is blue, there are two balloons, and the lion and the clock are merged. SD+ANSWER could add missing objects and corresponding attributes such as numeracy and color. In general, it also provides higher-quality images when compared to SD and SD+DNP. This is consistent with Table 1 of the main paper and validates our hypothesis that optimal negative changed with each timestep.

Figure 3. **SDXL Vs SDXL+ANSWER on the A&E Dataset:** Comparison between SDXL (Left) and SDXL+ANSWER (Right) with the corresponding prompt at the bottom. ANSWER resolves both entity neglect and incorrect attribute assignment.

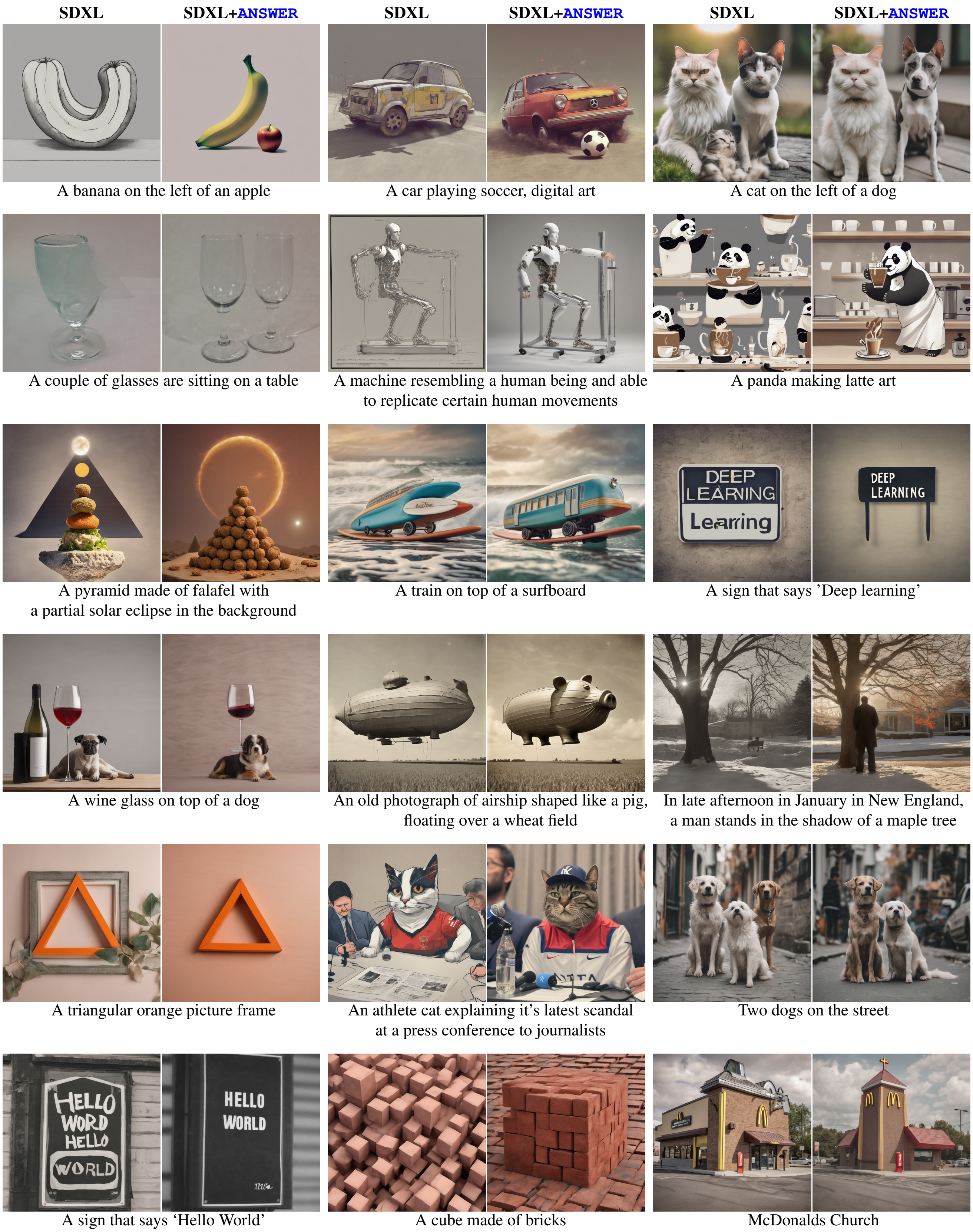


Figure 4. **SDXL Vs SDXL+`ANSWER` on the DrawBench Dataset:** Comparison between SDXL (Left) and SDXL+`ANSWER` (Right) with the corresponding prompt at the bottom. `ANSWER` can improve spatial, numerical, and other complexity in the prompts.

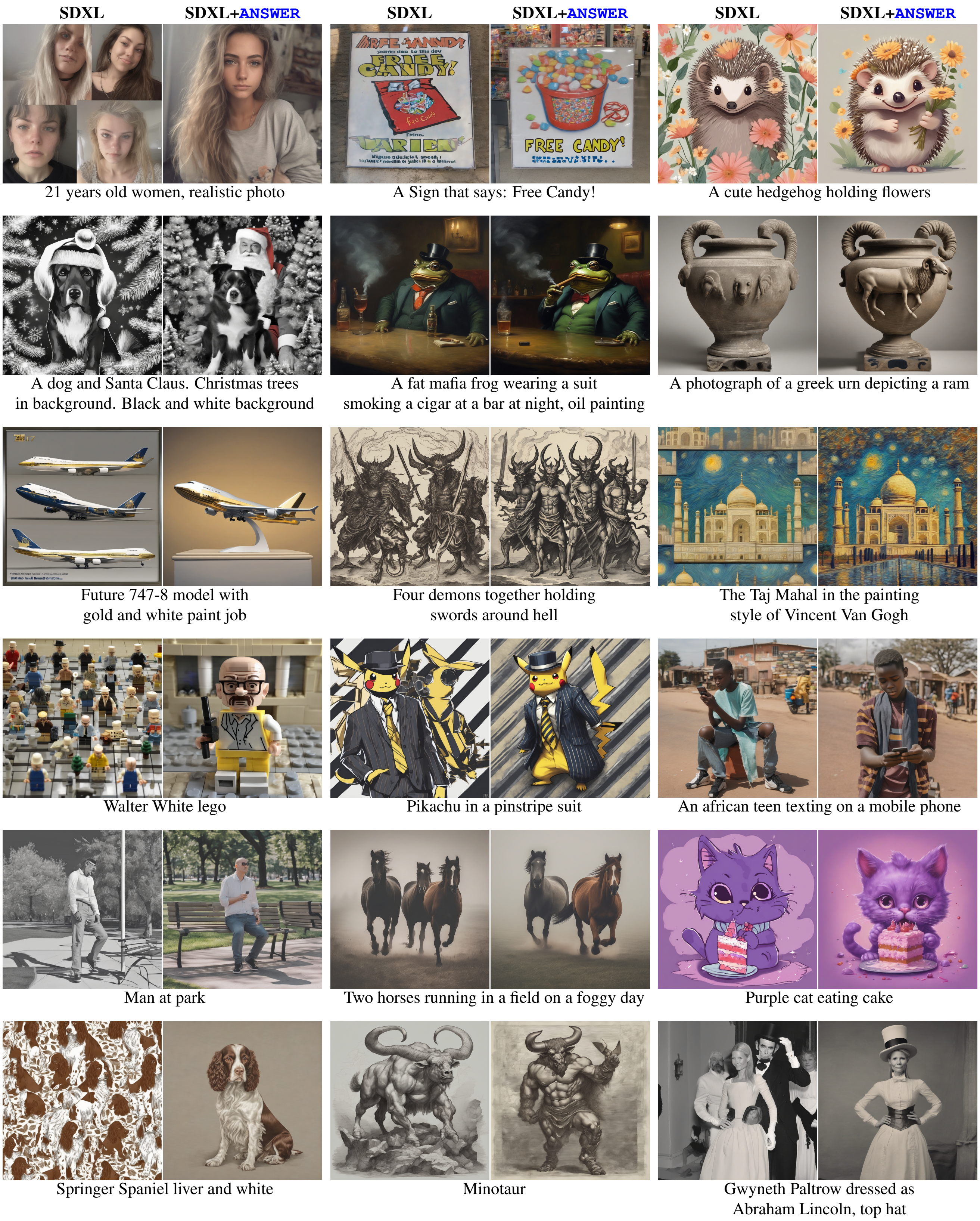


Figure 5. **SDXL Vs SDXL+ANSWER on the Pick-a-Pic Dataset:** Comparison between SDXL (Left) and SDXL+ANSWER (Right) with the corresponding prompt at the bottom. ANSWER improves generation ability on a variety of prompts

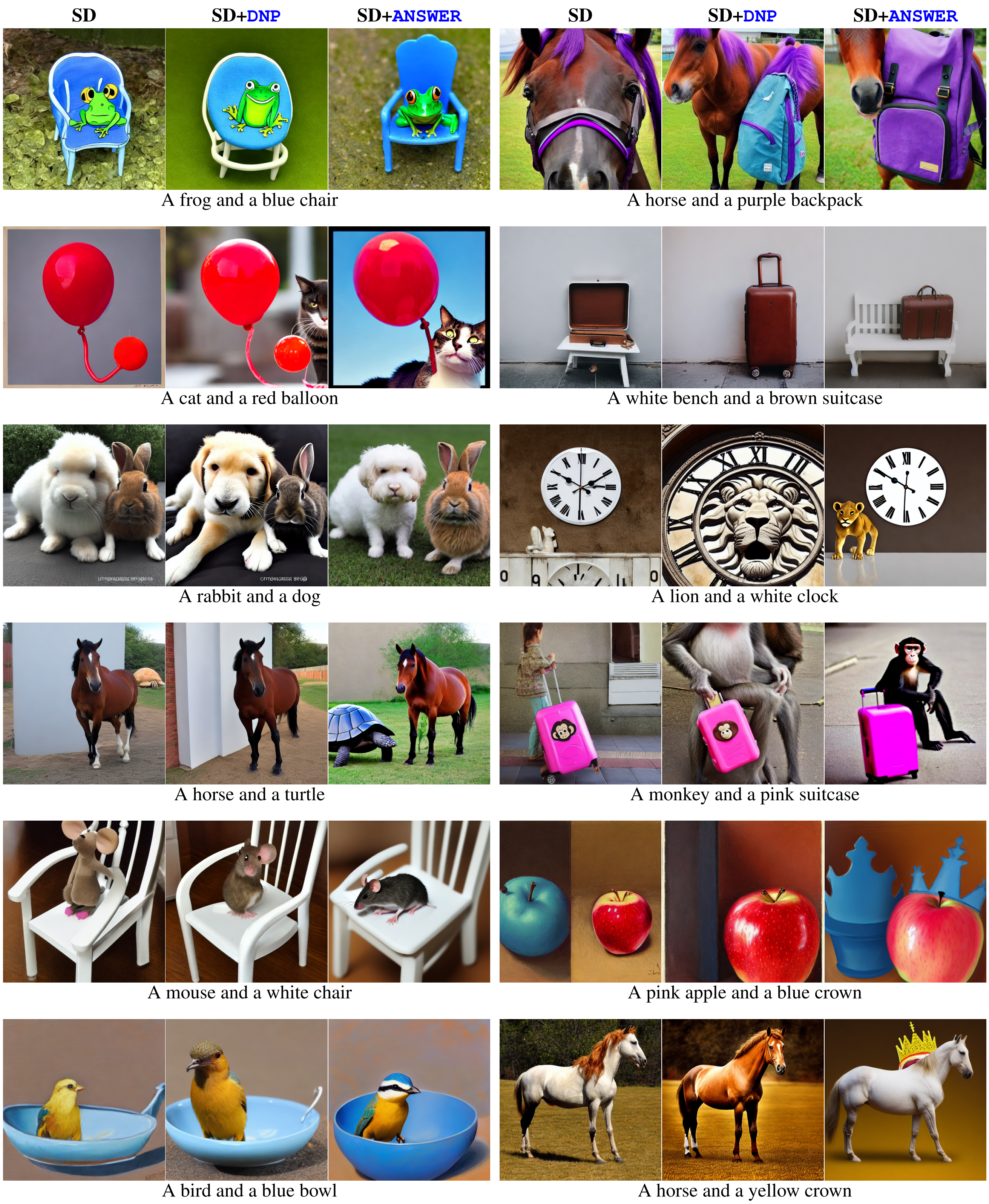


Figure 6. **`DNP` Vs. `ANSWER` on the A&E Dataset:** Comparison between SD, SD+`DNP`, and SD+`ANSWER` (From Left to Right) with the corresponding prompt at the bottom. While `DNP` improves upon SD, it doesn't always succeed. `ANSWER` can correct when `DNP` fails.

References

[1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 1

[2] Alakh Desai and Nuno Vasconcelos. Improving image synthesis with diffusion-negative sampling, 2024. 2

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[5] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 1

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[9] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. 1

[10] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. 1

[11] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1

[12] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 1