Improving image synthesis with diffusion-negative sampling

Alakh Desai[1] and Nuno Vasconcelos[1]

University of California San Diego, USA
{ahdesai,nuno}@ucsd.edu

1 Extended Derivation of *Diffusion-Negatives*

Given z_t and $\mathbf{p}$, the negative-diffusion prompt is defined as the prompt that induces the noise vector maximally distant from $\hat{\epsilon}_p$, i.e.

$$\epsilon_{n^*} = \underset{n \mid ||\hat{\epsilon}_n||^2 = K}{\arg\max} \; ||\hat{\epsilon}_p - \hat{\epsilon}_n||^2 \tag{1}$$

under a length constraint K. This has Lagrangian

$$\mathcal{L} = ||\hat{\epsilon}_p - \hat{\epsilon}_n||^2 + \lambda(||\hat{\epsilon}_n||^2 - K), \tag{2}$$

whose critical points are obtained with

$$\frac{\partial \mathcal{L}}{\partial \hat{\epsilon}_n} = 2(\hat{\epsilon}_n - \hat{\epsilon}_p) + 2\lambda\hat{\epsilon}_n = 0 \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = (||\hat{\epsilon}_n||^2 - K) = 0. \tag{4}$$

From (3)

$$\hat{\epsilon}_n = \frac{1}{1+\lambda}\hat{\epsilon}_p \tag{5}$$

and, using (4),

$$K = \frac{1}{(1+\lambda)^2}\, ||\hat{\epsilon}_p||^2 \tag{6}$$

from which

$$\lambda^* = -1 \pm \frac{||\hat{\epsilon}_p||}{\sqrt{K}} \tag{7}$$

and, from (5), we obtain two critical points,

$$\hat{\epsilon}_{n^*} = \pm\sqrt{K}\frac{\hat{\epsilon}_p}{||\hat{\epsilon}_p||} \tag{8}$$

Since the Hessian matrix is a diagonal matrix of derivatives

$$\frac{\partial^2 \mathcal{L}}{\partial \hat{\epsilon}_n^2} = 2(1+\lambda^*) = \pm 2\frac{||\hat{\epsilon}_p||}{\sqrt{K}} \tag{9}$$

the Lagrangian is maximum for the negative value of λ^*, which leads to:

$$\hat{\epsilon}_{n^*} = -\sqrt{K}\frac{\hat{\epsilon}_p}{||\hat{\epsilon}_p||} \propto -\hat{\epsilon}_p. \tag{10}$$

2 Additional Details

2.1 Implementation Details

We use Stable Diffusion v1.4 as the DM for all our experiments. All the images were generated with denoising steps $T = 41$. For implementing *A&E*, we use their official codebase [2] with all its default settings. For integrating *A&E* with our method, we set the unconditional prompt to the DNP and halt *A&E* updates for the first 5 denoising steps. For subsequent steps, we allow *A&E* updates along with the DNP. The guidance scale used for the combined model was lower than others as the usual scale exhibited very high saturation.

Computing Resources: All experiments were run on NVIDIA GeForce RTX 3090 with 32GB RAM using PyTorch.

2.2 Dataset Description

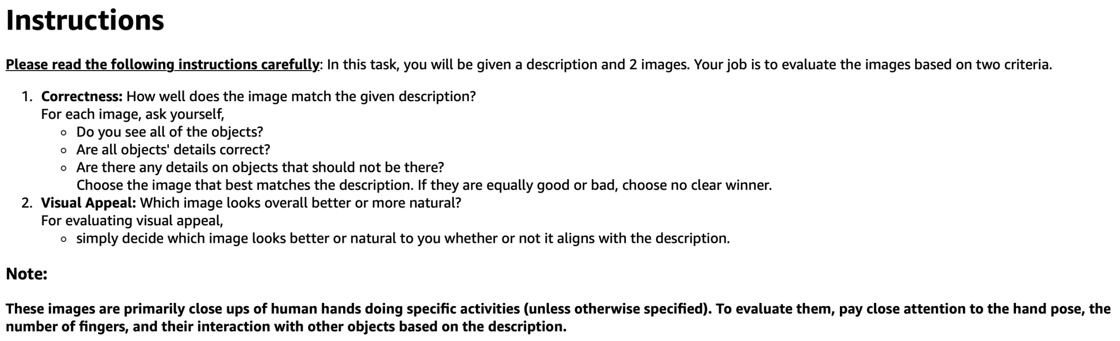
In this section, we describe the datasets used to evaluate our proposed method.

A&E Dataset In Table 1, we specify the animals, objects, and colors used to create the A&E prompts.

Category	
Animals	cat, dog, bird, bear, lion, elephant, horse, monkey, frog, turtle, rabbit, mouse
Objects	backpack, glasses, crown, suitcase, chair, balloon, bow, car, bowl, bench, clock, apple
Colors	red, orange, yellow, green, blue, purple, pink, brown, gray, black, white

Table 1: List of the animals, objects, and colors used to define each of the three data subsets used in A&E dataset.

Human Dataset Stable diffusion models tend to generate deformed, cropped, and low-quality humans. To evaluate our model's ability to improve the quality of human generation we designed a set of prompts focused on humans under varying conditions. To ensure variety in the dataset, we include different age groups (man, woman, old man, child, boy, girl),



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.

Note:

These images are primarily close ups of human hands doing specific activities (unless otherwise specified). To evaluate them, pay close attention to the hand pose, the number of fingers, and their interaction with other objects based on the description.

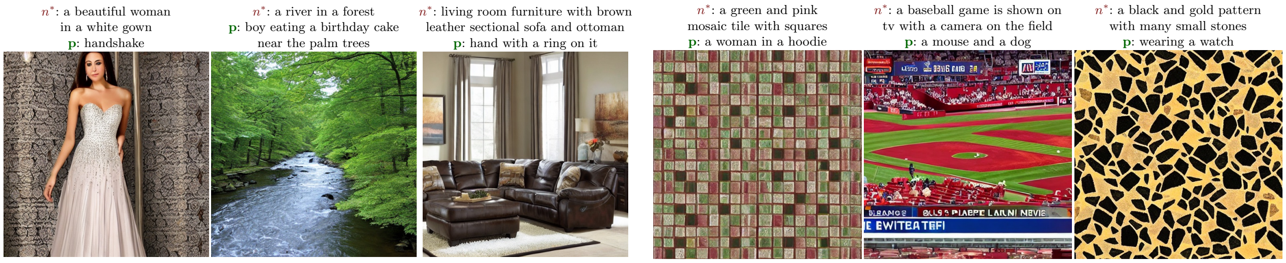
Instruction 1: The instructions provided to the MTurkers for Hand Prompts.

genders (man, woman, girl, boy), and numeracy (family, man and woman, people) while creating the prompts. We also chose a variety of tasks (eating, picnic, reading) and backgrounds (moonlight, beach, field) to ensure different poses (jumping, laying, diving, brushing). Some prompts were easy for stable diffusion to create ("*a woman in a hoodie*") where we demonstrated better facial proportions and more natural and realistic features. Others such as ("*a man sleeping in a hammock*") required correct interactions and poses which is harder for diffusion models and shows the strength of `DNP`. Objects requiring precise interactions ("*a woman applying lipstick*") or very high bias ("*frowning Mona Lisa*") appeared to be too difficult even with `DNP`.

Prompts: `{a man sleeping in a hammock, a photo of an angry man, a man heading a soccer ball, a selfie of an old man with a white beard, a photo of a boy in the moonlight, a man with long blond hair and blue eyes, a photo of an old man, a boy eating a birthday cake near some palm trees, a child on a couch, a boy eating a lollipop, a man wearing red shirt, a man in a sombrero, a woman in flowing pink dress, a man and a woman on the beach, a family picnic in a park, man in a prison uniform, woman giving a speech at a rally, a girl playing soccer, a girl diving into a pool, firefighter, close up of a bride and groom, a boy playing with his dog, a mother reading to her child, the mona lisa, a woman brushing her hair, a woman applying lipstick, a man shaving his beard, people dressed up for halloween, a woman in a hoodie, a dancing couple, a man playing tennis, a close up of a smiling man, a man looking in the mirror, close up of morgan freeman in a red suit, close up of elvis, cowboy, close up of james bond, a farmer holding a giant pumpkin, a girl spinning a hula hoop, a girl jumping in the air}`

Hand Dataset Human hands pose a major challenge to Stable Diffusion models owing to the complexity of the interactions and the ambiguity in the descriptions. For example, "holding a tennis racket" and "holding a basketball" imply very different poses and interactions despite the common terminology. To evaluate the models in this regard, we create a dataset focused on hands in various poses and performing various tasks. We observed that `DNP` helped obtain the right number of fingers and reduced the confusion between the left and right hands. Our curated prompts include abstract tasks (such as "*hands making shadow puppets*" and "*a three-fingered alien hand*") and complicated poses (such as "*close up of hands while playing a clarinet*"). These prompts capture a variety of hand poses and details while keeping the prompts simple.

Prompts: `{the hands of a single person holding a basketball, close up of hands while playing a clarinet, hand holding a ball, hand holding a racket, a tattooed hand, a closed fist, man giving a thumbs up, handshake, hands knitting a sweater, hand with a ring on one finger, hands playing the piano, hands typing on a keyboard, hands doing pottery, hands chopping with a knife, hands on a steering wheel, hands making shadow puppets, holding playing cards, wearing a watch, writing with hand, pointing a finger, close up of man praying, a three-fingered alien hand, live long and prosper hand sign, saluting hand, hands of a baby}`


n^*: a beautiful woman in a white gown
$\mathbf{p}$: handshake
n^*: a river in a forest
$\mathbf{p}$: boy eating a birthday cake near the palm trees
n^*: living room furniture with brown leather sectional sofa and ottoman
$\mathbf{p}$: hand with a ring on it
n^*: a green and pink mosaic tile with squares
$\mathbf{p}$: a woman in a hoodie
n^*: a baseball game is shown on tv with a camera on the field
$\mathbf{p}$: a mouse and a dog
n^*: a black and gold pattern with many small stones
$\mathbf{p}$: wearing a watch

(a) Examples of captionable $\bar{\mathcal{I}}$: The images are clear and easy to caption. Most $\bar{\mathcal{I}}$ belong to this category.

(b) Examples of non-captionable $\bar{\mathcal{I}}$: The images are either pattern/texture or abstract images that are hard to caption.

Fig. 1: Example of *diffusion-negative* images ($\bar{\mathcal{I}}$). The corresponding prompts $\mathbf{p}$ and their captions n^* can be found at the top.

2.3 Human Evaluation

An Amazon Mechanical Turk (AMT) task was used for the human evaluation of each experiment. Each task was designed as two multiple-choice questions, evaluated and reported independently. These two questions are:

- **Adherence/Correctness:** MTurkers were tasked with judging the images based on “correctness" or prompt adherence alone while ignoring the quality.
- **Quality:** MTurkers were tasked with picking the most natural or realistic image, even if it meant disregarding considerations of “correctness.".

For the A&E dataset, the existence of both entities along with their corresponding attributes is evaluated. For the H&H dataset, the MTurkers checked for the correct number of limbs / fingers / thumbs, no disfigurement, correct body proportions, and realistic interaction with background / objects.

The set of choices for both questions was {*Image-1*, *Image-2*, *No clear winner*}. We randomly swapped *Image-1* and *Image-2* to avoid human bias. To ensure good quality results all experiments were evaluated by 10 unique *Master MTurkers* with an approval rate exceeding 98%.

An example of the precise instructions for the AMT task can be seen in Instruction 1. These instructions were customized for each experiment. For each experiment, we did not pick the prompt-seed pairs that contained NSFW images for any of the methods.

3 Limitations

The *diffusion-negative* image, $\bar{\mathcal{I}}$, represents negative of a given prompt in the diffusion space and the `DNP` is an estimate obtained from the human or captioning model. Therefore, two possible failure modes exist:

1) `DNS` can produce an “incorrect” $\bar{\mathcal{I}}$. Since the groundtruth $\bar{\mathcal{I}}$is unknown and unintuitive for humans, there is no way to measure this other than checking whether the end-to-end process improves prompt compliance, which we do.

2) For some prompt-seed pairs, the $\bar{\mathcal{I}}$ is not “captionable" by a conventional captioning model or even a human because it is abstract or hard to describe correctly. Under these circumstances, the estimated `DNP` does not reflect the true negative and therefore fails to enforce prompt adherence. In Figure 1a, we show samples of the $\bar{\mathcal{I}}$ which are straightforward to caption and most $\bar{\mathcal{I}}$'s belong to this group. However, there are $\bar{\mathcal{I}}$'s such as those in Figure 1b that cannot be captioned by humans or any off-the-shelf captioning model. While some of these images comprise textures or designs that are hard to transcribe except in overtly simplistic terms such as “abstract design" or “repeating texture", others are gibberish images with no discernible meaning. In such cases, the success or failure of `DNP` depends on whether the caption captured enough of the essence of the *diffusion-negative* image in the estimated `DNP` for the odds ratio to be increased in favor of the target prompt.

We emphasize that our results already account for both types of failure.

4 Ablations

4.1 Captioning Ablation

`DNP` vs `auto-DNP`: We continue the comparison of `DNP` and `auto-DNP` in this section. As mentioned in Section 4.3 of the main paper, we evaluate `DNP` only on a small subset of the prompt-seed pairs (150 A&E, 100 Human, and 100 Hand) to avoid the labor-intensive captioning task. In Figure 2, we show some examples of the generated results, we observe that the images generated by `auto-DNP` and `DNP` are similar a lot of the time. We also observe that a simple caption often provides a better result and human captioners are more inclined towards simpler captions. For example, a human might caption an image as a "red car" while the model might be specific and caption it as "red 2020 Ford Mustang".

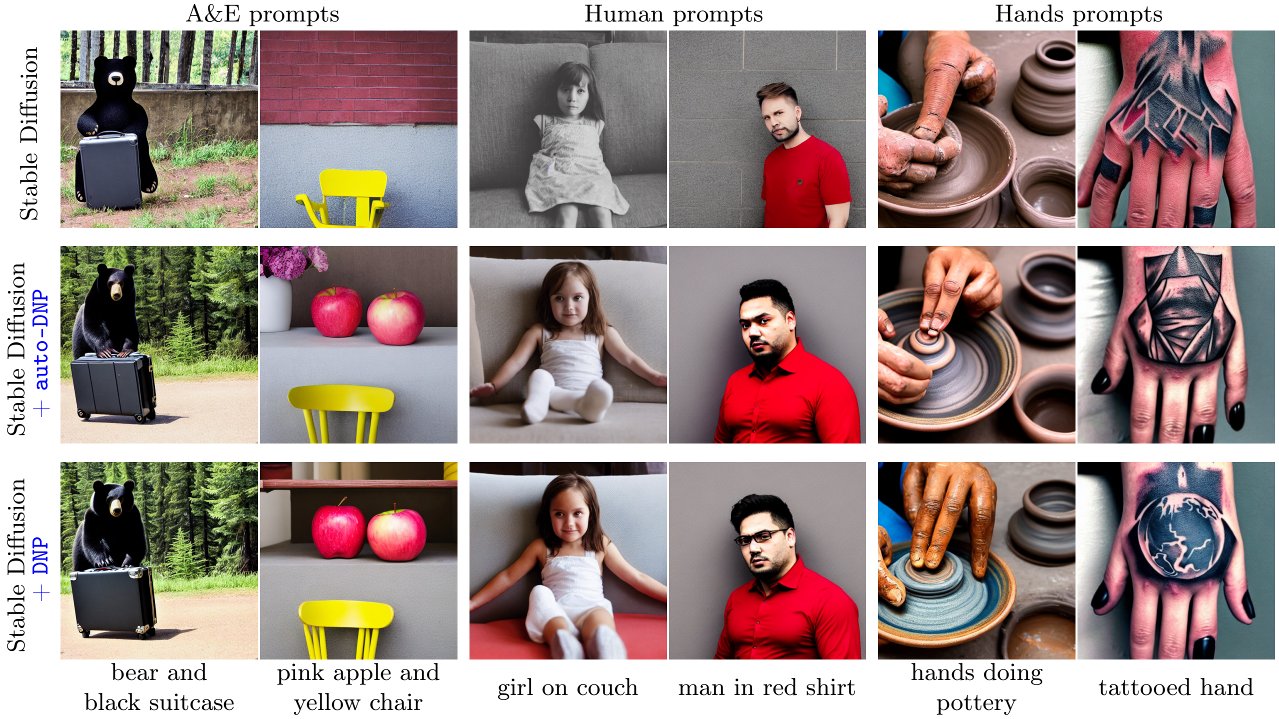


Fig. 2: `DNP` **vs** `auto-DNP`: Comparison between SD, `auto-DNP`, and `DNP` (from top to bottom) with the corresponding prompt at the bottom. We observe that images generated by `DNP` and `auto-DNP` are similar and the choice between them appears to be personal preference.

Dataset	Method	Human Evaluation	
		Correctness	Quality
A&E	No Clear Winner	17.19%	13.72%
	SD + `auto-DNP`	35.73%	37.35%
	SD + `DNP`	**47.08%**	**48.93%**
Human	No Clear Winner	15.94%	11.52%
	SD + `auto-DNP`	35.95%	36.77%
	SD + `DNP`	**48.11%**	**51.72%**
Hand	No Clear Winner	19.08%	11.56%
	SD + `auto-DNP`	38.97%	42.66%
	SD + `DNP`	**41.95%**	**45.78%**

Table 2: Comparing `DNP` with `auto-DNP` on all datasets

Dataset	Method	Human Evaluation	
		Correctness	Quality
Human	No Clear Winner	16.74%	8.63%
	Stable Diffusion (SD)	34.53%	32.32%
	SD + GPT-4V `auto-DNP`	**48.74%**	**59.05%**
Hand	No Clear Winner	11.88%	7.14%
	Stable Diffusion (SD)	36.12%	34.71%
	SD + GPT-4V `auto-DNP`	**52.0%**	**58.15%**

Table 3: Comparing GPT-4V `auto-DNP` with SD on the H&H dataset

Table 2 shows the results of the human evaluation task comparing `DNP` and `auto-DNP`. We observe a 10% drop in preference of `auto-DNP` over `DNP`. While the results are not split evenly between `auto-DNP` and `DNP`, `auto-DNP` matches and beats `DNP` $\sim$ 50% of the time for all datasets (compared to the $\sim$ 35% of the time that SD matches or beats `auto-DNP`, shown in Table 2 of the main paper). This shows the validity of our hypothesis in `DNP` as well as its automated implementation `auto-DNP`.

GPT-4V for `auto-DNP`**:** To evaluate whether other automated captioning models can provide similar improvements, we replaced Blip-v2 with GPT-4V (*gpt-4-vision-preview*) in the `auto-DNP` process. Table 3 shows the results on the H&H dataset where we ask human evaluators to choose between (SD, SD+GPT-4V `auto-DNP`, and No Clear Winner). As observed for Blip, humans prefer the images synthesized with SD+GPT-4V `auto-DNP` most frequently, with bigger gains for hands than for humans. GPT-4V results are comparable to Blip's results in Table 2b of the main paper. This shows that the benefits of `DNP` are not specific to Blip and can be used with any VLM.

4.2 Negative Prompting Ablations

In this section, we compare `auto-DNP` with other negative prompts. To be consistent in the comparison we pick H&H dataset for the following two reasons. 1) Semantic negations do not make sense for the A&E prompts, which are of the type *"a & b"*, without saying something like *"not a or b"*. 2) Standard negations, used by the community, are not directly applicable for compositional prompts like those in the A&E dataset either. Human evaluation (as previously detailed) is used for all ablations, due to the limited size of the dataset used in these experiments, along with the possibility of Clip and IS overlooking distorted or maligned results when considering correctness or realism for H&H dataset.

Dataset	**Method**	**Human Evaluation**	
		Correctness	**Quality**
Human	No Clear Winner	16.44%	5.58%
	Stable Diffusion (SD)	10.96%	10.78%
	SD + GPT NegPrompt	19.52%	21.75 %
	SD + `auto-DNP`	**53.08%**	**61.89%**
Hand	No Clear Winner	11.84%	4.49%
	Stable Diffusion (SD)	5.98%	7.03%
	SD + GPT NegPrompt	17.01%	19.01%
	SD + `auto-DNP`	**65.17%**	**69.47%**

Table 4: Comparing GPT NegPrompt with `auto-DNP` on the H&H dataset.

GPT Generated Negatives: First, we evaluate the performance of our model against "semantic" negatives, we generate a negative prompt using GPT3.5 (*gpt-3.5-turbo-0613*). To generate these prompts, we prompt GPT3.5 with *"Give negative prompt which is semantically opposite to the prompt provided. The negative prompt should help eliminate things or concepts that contradict the given prompt. Don't use negation in creating negative prompts. Exclude words like no, not, remove, etc."* and the target prompt and use the results as negative prompts for this experiment. MTurkers were asked to choose between images generated by SD, SD+GPT NegPrompt, and SD+`auto-DNP`. As seen in Table 4, GPT-generated negative prompts improve both the "correctness" and the quality of the generated images over vanilla Stable Diffusion. However, adding `auto-DNP` to Stable Diffusion improves both by a landslide. This shows, that while helpful, semantic negatives are not reliable in terms of the improvement that they provide.

Dataset	**Method**	**Human Evaluation**	
		Correctness	**Quality**
Human	No Clear Winner	7.80%	6.08%
	Stable Diffusion (SD)	16.46%	14.37%
	SD + random	20.02%	22.57%
	SD + `auto-DNP`	**55.73%**	**56.99%**
Hand	No Clear Winner	2.61%	1.37%
	Stable Diffusion (SD)	19.77%	16.04%
	SD + random	17.5%	18.77%
	SD + `auto-DNP`	**60.11%**	**63.82%**

Table 5: Comparing random negative prompts with `auto-DNP` on the H&H dataset

Random Negatives: As a simple benchmark, we repeat the experiment with random negative prompts (famous quotes from [5]). The results are shown in Table 5 where we asked Turkers to choose between SD, SD+`auto-DNP`, and SD+random. SD+`auto-DNP` has a clear advantage over SD+random. The % wins by the former were also similar to those of Table 2b of the main paper (e.g. between 55-60% for correctness), suggesting that SD+random mostly improved on the failure cases of SD+`auto-DNP`.

Dataset	**Method**	**Human Evaluation**	
		Correctness	**Quality**
Human	No Clear Winner	21.45%	6.69%
	Stable Diffusion (SD)	18.31%	12.84%
	SD + standard	28.14%	39.89%
	SD + `auto-DNP`	**32.10%**	**40.57%**
Hand	No Clear Winner	17.61%	7.73%
	Stable Diffusion (SD)	19.69%	10.39%
	SD + standard	16.36%	25.45%
	SD + `auto-DNP`	**48.30%**	**56.02%**

Table 6: Comparing standard negative prompts with `auto-DNP` on the H&H dataset

Standard Negatives: For a better benchmark, we repeat the experiment with a set of negative prompts collated by the community at HuggingFace [1], with the results in Table 6. While these are more competitive than random prompts, there is still a clear advantage for SD+`auto-DNP`, particularly for hands. This is in line with our observation that negative prompting depends on both the seed and the prompt and standard negative prompts do not capture the impact of the seed. They also require a collective human curation effort, which may not generalize across diffusion models, while `auto-DNP` is fully automated. Standard prompts are also limited in their usability, e.g. it is unclear how to define them on the A&E dataset. You can also add the standard negative prompts to auto-DNP to achieve the best results.

4.3 More Baselines

Method	CLIP Score		IS
	Min. Object	Full Prompt	
Stable Diffusion XL (SDXL) [3]	0.266	0.353	11.23
SDXL+auto-DNP	**0.274** (+3.01%)	**0.356** (+0.84%)	**11.67** (+3.92%)
SynGen (SD1.4) [4]	0.259	0.345	11.46
SynGen+auto-DNP	**0.264** (+1.93%)	**0.352** (+2.03%)	**12.96** (+13.09%)

Table 7: Quantitative Results for SDXL [3] and SynGen [4] on A&E dataset

We also ran our method on Stable Diffusion XL [3] and SynGen [4] for A&E prompts. SynGen is a method designed for better alignment between entities and their attributes by adjusting the attention between them. It uses linguistic binding to maximize attention of each subject phrase explicitly binding subjects with their corresponding attributes. To do this they use two types of losses: 1) positive loss between intra phrase tokens, and 2) negative loss between inter phrase tokens. As shown in Table 7, auto-DNP improves the performance with substantial improvement in the IS for SynGen. The lower margins can be attributed to SDXL and Syngen being strong baselines. We note that SDXL IS is low per Section F of their Appendix [3].

5 Additional Qualitative Results

In this section, we show additional qualitative results for all the datasets. Figure 3 and Figure 4 show the qualitative results for the Hand and Human datasets respectively. We observe an overall improvement in both correctness and image quality. The A&E dataset results are divided into 3 figures, one for each category. Figures 5, 6, 7 show the results for the Animal-Animal, Animal-Object, and Object-Object prompts respectively. We observe improvement in correctness over SD and quality over *A&E*.

References

1. AdamOswald1: stabilityai/stable-diffusion - Negative Prompts. https://huggingface.co/spaces/stabilityai/stable-diffusion/discussions/7857#63bee17e20784381e8e54d33
2. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models (2023)
3. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
4. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment (2023)
5. robatron: quotes - gist.github.com. https://gist.github.com/a66acc0eed3835119817.git

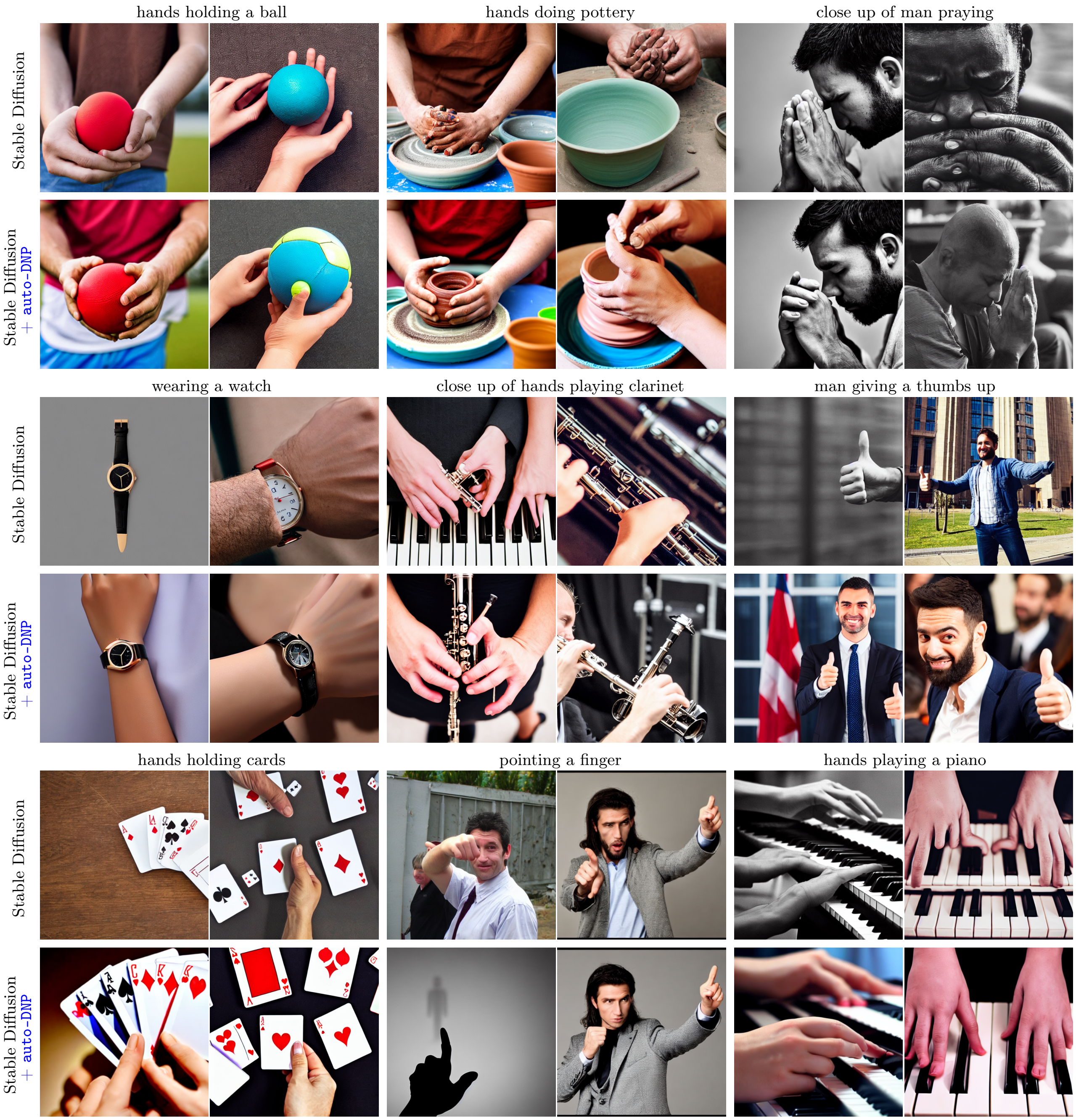


Fig. 3: Qualitative Results for the Hands Dataset. For each prompt, we show a pair of images (generated by both methods) for two seeds. **Note:** our model ensures correctness (“wearing" a watch, “holding" cards) and quality (correct pose and number of fingers/hands)

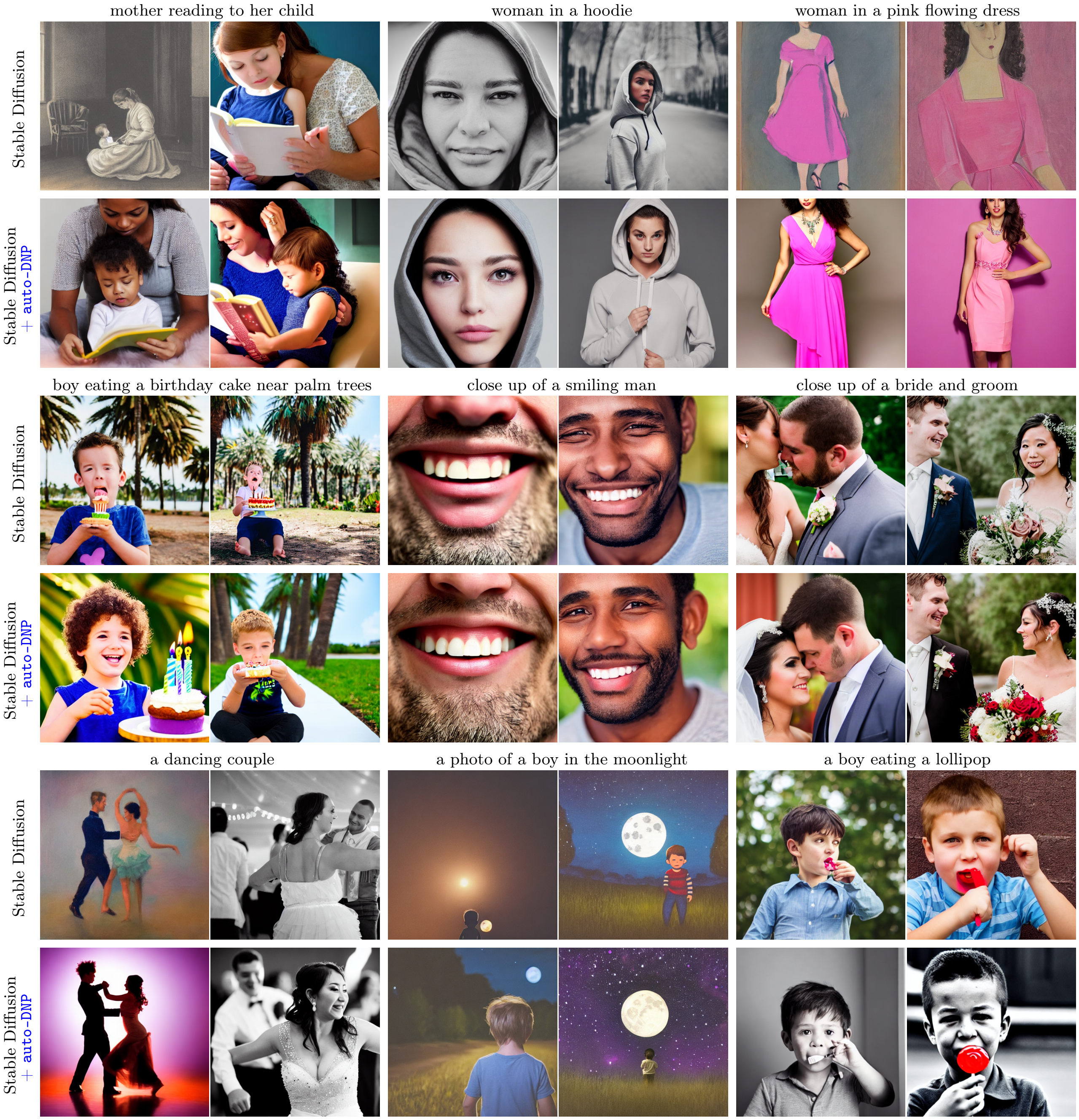

mother reading to her child
woman in a hoodie
woman in a pink flowing dress
Stable Diffusion
Stable Diffusion + auto-DNP
boy eating a birthday cake near palm trees
close up of a smiling man
close up of a bride and groom
Stable Diffusion
Stable Diffusion + auto-DNP
a dancing couple
a photo of a boy in the moonlight
a boy eating a lollipop
Stable Diffusion
Stable Diffusion + auto-DNP

Fig. 4: Qualitative Results for the Human Dataset. For each prompt, we show a pair of images (generated by both methods) for two seeds. **Note:** our model ensures undistorted facial features (mouth of the “*smiling man*", merged heads of the “*mother reading to her child*"), correct number of limbs (“*dancing couple*") and realism (“*woman in pink dress*", “*boy in moonlight*")

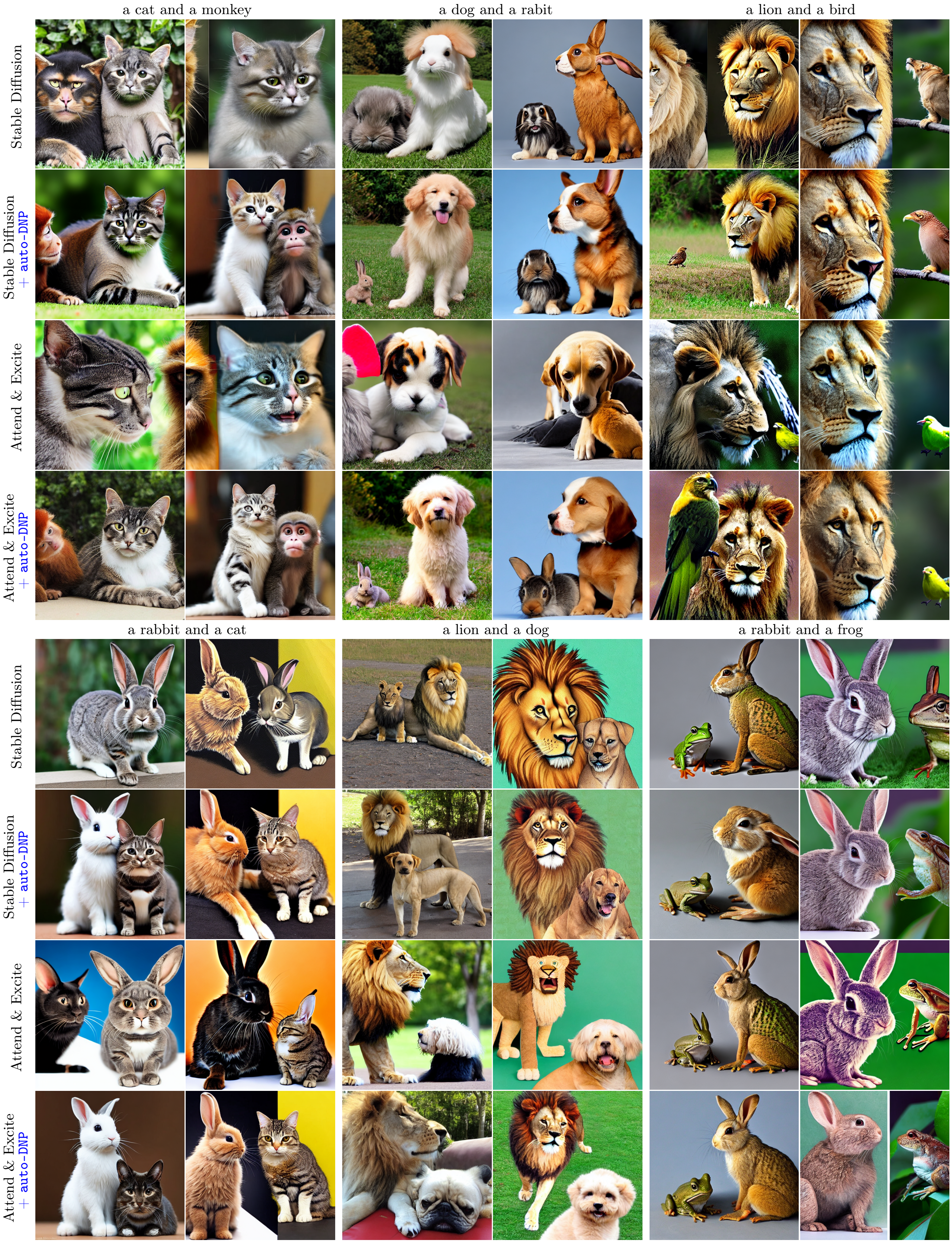


Fig. 5: Qualitative Results for the A&E Dataset's Animal-Animal prompts: Our model resolves entity neglect when compared to SD and improves quality and realism when combined with the *A&E* method.

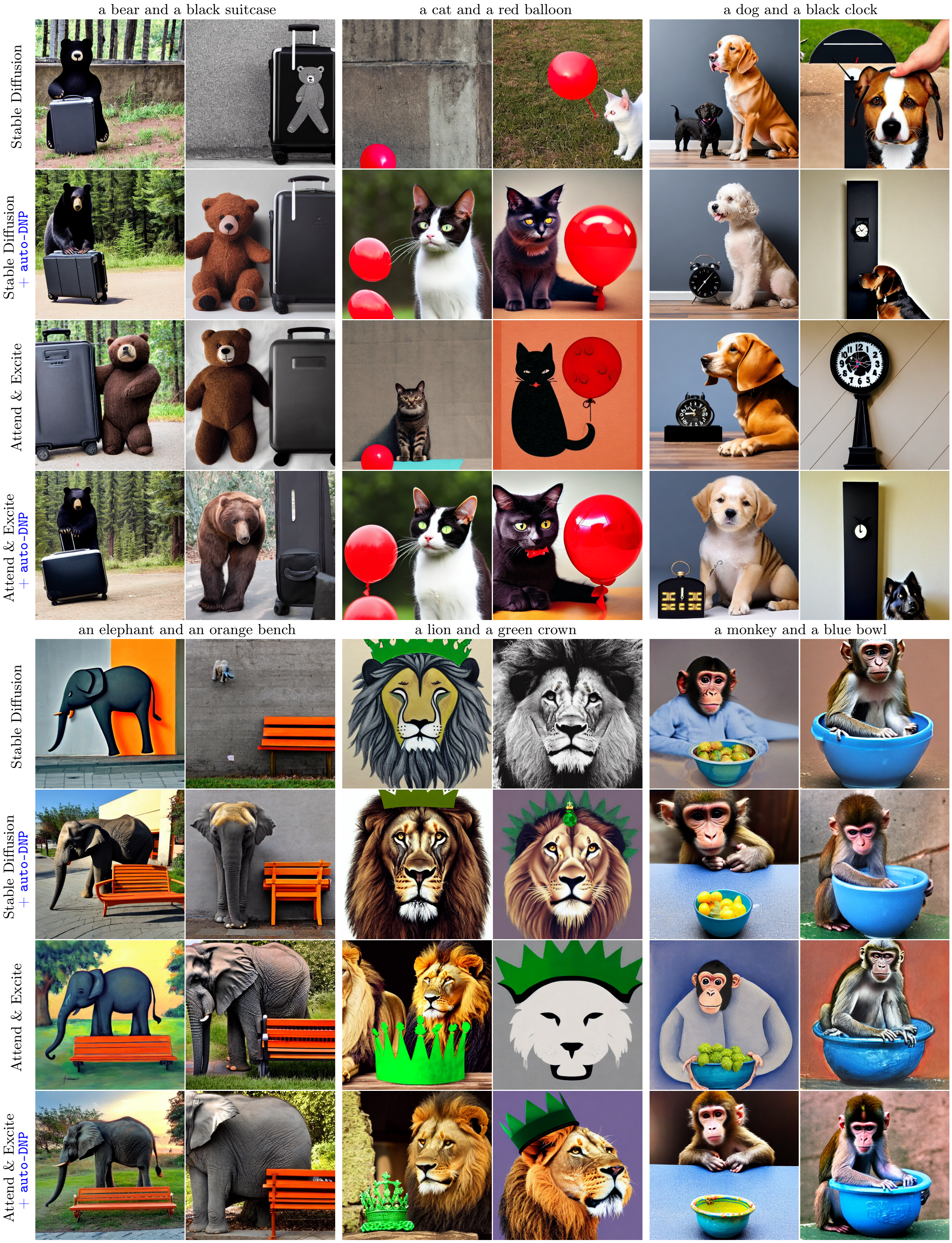

a bear and a black suitcase
a cat and a red balloon
a dog and a black clock
an elephant and an orange bench
a lion and a green crown
a monkey and a blue bowl
Stable Diffusion
Stable Diffusion + auto-DNP
Attend & Excite
Attend & Excite + auto-DNP
Stable Diffusion
Stable Diffusion + auto-DNP
Attend & Excite
Attend & Excite + auto-DNP

Fig. 6: Qualitative Results for the A&E Dataset's Animal-Object prompts: Our model resolves entity neglect and entity merging when compared to SD and improves quality and realism when combined with the *A&E* method.

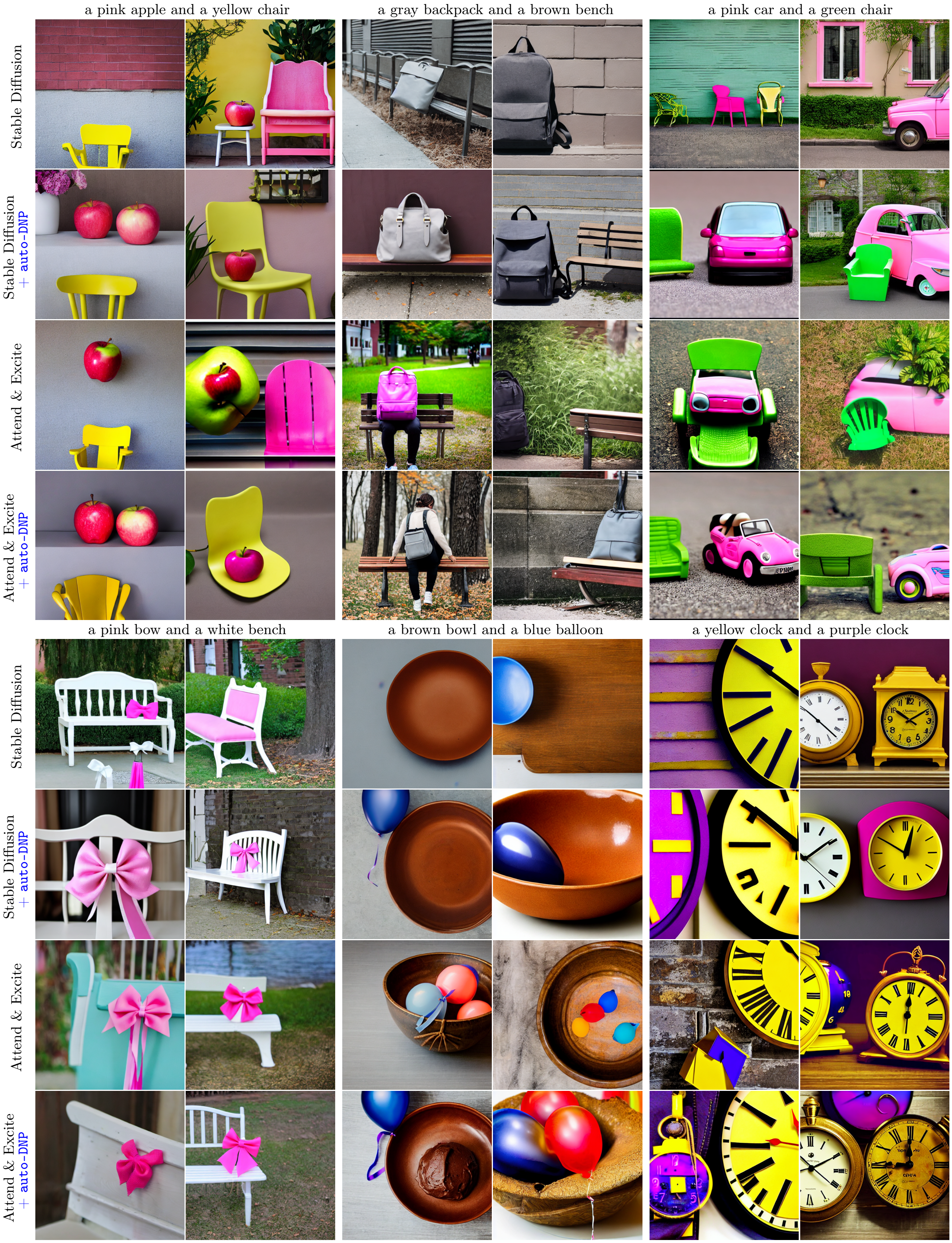

a pink apple and a yellow chair
a gray backpack and a brown bench
a pink car and a green chair
Stable Diffusion
Stable Diffusion + auto-DNP
Attend & Excite
Attend & Excite + auto-DNP
a pink bow and a white bench
a brown bowl and a blue balloon
a yellow clock and a purple clock
Stable Diffusion
Stable Diffusion + auto-DNP
Attend & Excite
Attend & Excite + auto-DNP

Fig. 7: Qualitative Results for the A&E Dataset's Object-Object prompts: Our model resolves both entity neglect and incorrect attribute assignment.