A Limitations

Our method is limited to image-text matching and cannot tackle other task formulations such as VQA [Agrawal et al., 2015, Suhr et al., 2019]. Additionally, taking more noise samples for a given image-text pair leads to better performance up to a certain point (see Fig. 9). This results in slower inference time and can hopefully be mitigated with future innovations.

Measuring bias is a crucial aspect of model evaluation and should not be considered secondary to the "performance" of the model. Unfortunately, the dataset from Janghorbani and De Melo [2023] does not include gender bias but other datasets such as Zhou et al. [2022] could also be incorporated. In any case, evaluating bias related to personal identities such as nationality, race, and sexual orientation presents certain limitations. These labels are nuanced and not always accompanied by visible identifying characteristics. Consequently, image datasets depicting these groups have limited capacity to fully represent these demographics and intersectional identities. While we recognize that assigning a bias score based on these limited resources might not be entirely accurate, it is a vital first step in the right direction.[8] It is also important to note that the bias evaluation has only positive predictive power, meaning that a large effect size is an indication of bias, but a low one does not necessarily verify that the model is fair. Moreover, the bias effect size (Eq 8) may sometimes be unreliable [Meade et al., 2022]. This should serve as a preliminary analysis of biases present in the model, which require further investigation using more nuanced and comprehensive methods, including broader qualitative analysis and consultation with social scientists.

B DrawBench Evaluation

As described in Sec. 5, we manually compare our best *HardNeg Stable Diffusion* with vanilla Stable Diffusion on the DrawBench benchmark [Saharia et al., 2022]. It contains 200 curated challenging prompts along 11 categories. We only select these categories that are relevant to the phenomena studied in this paper (spatial, compositional, attribute binding, ...) and drop categories such as Text (i.e. *A storefront with 'Hello World' written on it.*). This leaves us with 6 categories and 104 prompts:

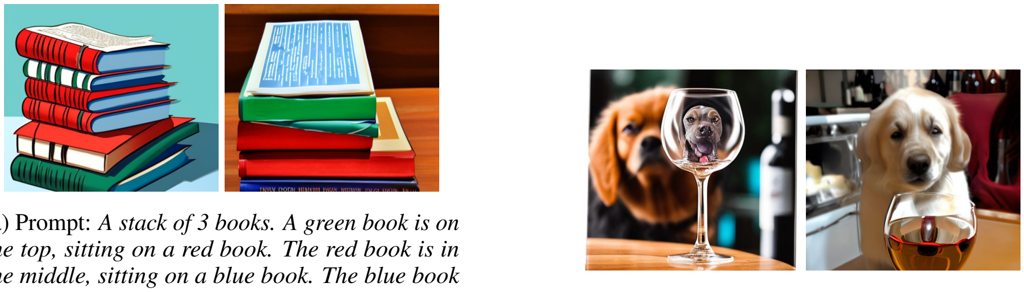
Category	Prompts
Colours	A brown bird and a blue bear.
Conflicting	A horse riding an astronaut.
Counting	One cat and three dogs sitting on the grass.
DALL-E	An illustration of a small green elephant standing behind a large red mouse.
Gary Marcus et al.	An elephant is behind a tree. You can see the trunk on one side and the back legs on the other.
Positional	A train on top of a surfboard.

Table 4: DrawBench categories with examples

Next we generate images for each prompt with both models and display them randomly to an author as left or right image. We judge whether the left or right image is better aligned with the text and do not focus on image quality. Because we only focus on high-level alignment with text, we observe many ties as both models exhibit the same level of alignment. We repeat this process with another seed, leaving us with 208 blind comparisons. **Out of those, our *HardNeg* model is judged better almost twice as often as vanilla Stable Diffusion (60 vs. 33).** The rest (115) are ties.

We show an example where *HardNeg* is more text-aligned as well as a tie below:

[8] For a comprehensive discussion on the limitations of fixed labels inferable from a person's appearance, creating datasets based on such visual factors for bias analysis, as well as the discussion on the consequences of biased image generation systems, refer to Luccioni et al. [2023].



(a) Prompt: *A stack of 3 books. A green book is on the top, sitting on a red book. The red book is in the middle, sitting on a blue book. The blue book is on the bottom.* Zero-shot (left) gets it completely wrong while our finetuned Stable Diffusion (right) is quite close to the text.

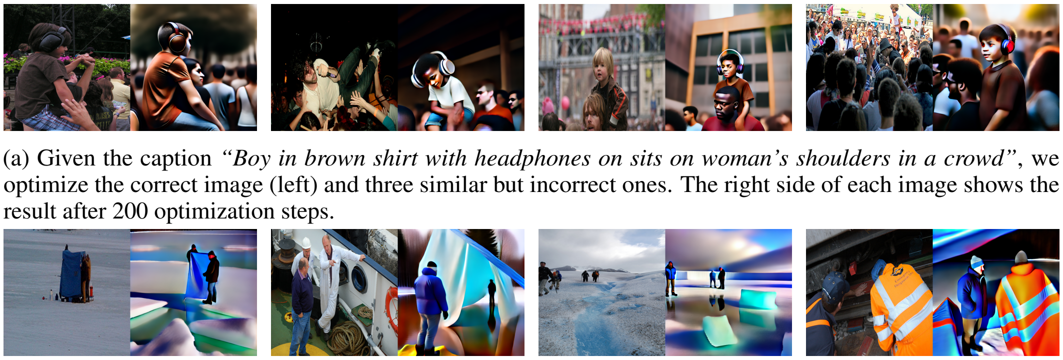
(b) Prompt: *A wine glass on top of a dog.* One of many examples where both models either fail or succeed to an equal extent to capture the text.

Figure 5: Examples of DrawBench prompts with vanilla Stable Diffusion and *HardNeg* Stable Diffusion.

Since *HardNeg* is not strictly better than *NoNeg* in our experiments (see Tab. 2), we conduct the same study with these two models: *HardNeg* wins 50 comparisons and *NoNeg* only 38. While this is not statistically significant, it indicates that *HardNeg* finetuning is useful for image-text-alignment.

C Visualizing image editing with input optimization and standard denoising

Optimizing the input image directly via backpropagation has shown promise in the diffusion literature [Poole et al., 2022, Samuel et al., 2023]. It also leads to qualitatively insightful images that differ from standard image editing with Stable Diffusion. Concretely, we optimize the input image (i.e. the latent z) with Adam (lr=0.05, 200 steps) based on the noise prediction loss where noise is added at timestep $t = 0.5$. In other words: the input image itself is modified such that it leads to a lower noise prediction error. Note that this was generated with Stable Diffusion 1.5 before we adopted 2.1 into our experiments.

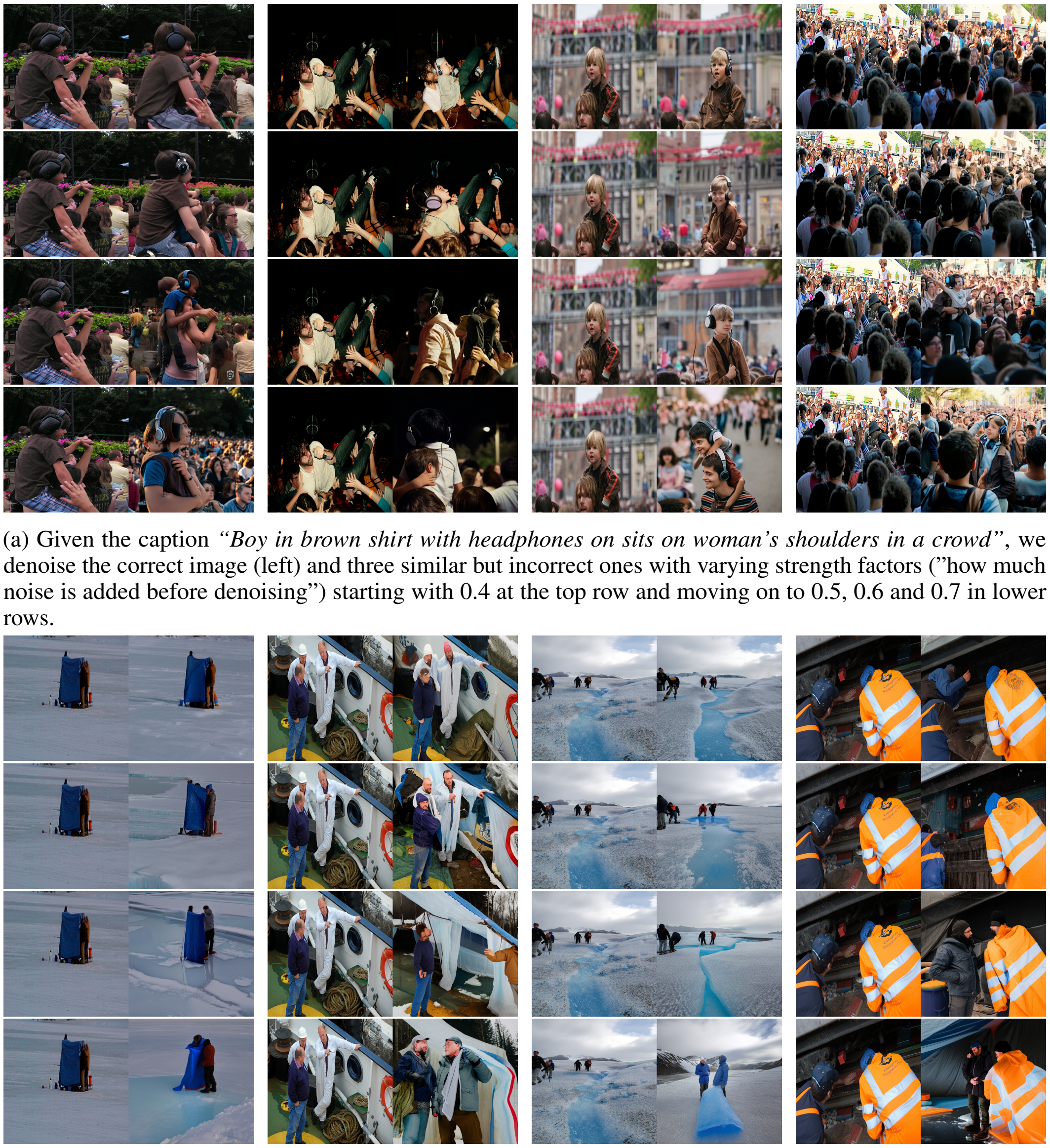


(a) Given the caption *"Boy in brown shirt with headphones on sits on woman's shoulders in a crowd"*, we optimize the correct image (left) and three similar but incorrect ones. The right side of each image shows the result after 200 optimization steps.

(b) Given the caption *"Two men, standing on an ice, looking into something covered with a blue tarp"*, we optimize the correct image (left) and three similar but incorrect ones. The right side of each image shows the result after 200 optimization steps.

Figure 6: We visualize the underlying intuition of our presented methods that image-text pairs that do not fit will lead to more edits. For example in a) we can see how the model needs to add a boy with headphones except for the correct image.

Next, we compare image optimizaton to the more established image editing approach to visualize the models reasoning, i.e. the denoising process does not start from pure noise but from a partially noisy image. Below, we show different strength factors between 0.4 and 0.7 for the same two examples as above and denoise for the corresponding amount of steps (according to recommended HuggingFace Diffusers parameters a value of 0.5 would correspond to 25 denoising steps opposed to the usual 50 steps from pure noise). A lower strength factor means less added noise and therefore less editing.



(a) Given the caption *"Boy in brown shirt with headphones on sits on woman's shoulders in a crowd"*, we denoise the correct image (left) and three similar but incorrect ones with varying strength factors ("how much noise is added before denoising") starting with 0.4 at the top row and moving on to 0.5, 0.6 and 0.7 in lower rows.

(b) Given the caption *"Two men, standing on an ice, looking into something covered with a blue tarp"*, we denoise the correct image (left) and three similar but incorrect ones with varying strength factors ("how much noise is added before denoising") starting with 0.4 at the top row and moving on to 0.5, 0.6 and 0.7 in lower rows.

Figure 7: Similar to Fig. 6 we can see how in headphones are added in the incorrect images in a) or how in the third row of b) a blue structure is added. However the edits are overall less reliable and either change too little or too much.

D Bias Evaluation Details

Table 5: Additional bias evaluation results for Buddhism. Positive effect sizes indicate bias towards target group $\mathbf{X}$, negative effect sizes indicate bias towards $\mathbf{Y}$. Effect sizes closer to 0 are less biased and statistically significant effect sizes at $p < 0.01$ are denoted by $*$.

	Target $\mathbf{X}$	Target $\mathbf{Y}$	SD 2.1	SD 1.5	CLIP RN50x64	CLIP ViT-B/32
Religion	Buddhist	Muslim	**0.79***	0.94*	1.62*	1.63*
	Buddhist	Christian	-0.19	-0.16	0.24*	-0.78
	Buddhist	Hindu	-0.11	-0.47	0.46*	-0.48
	Buddhist	Jewish	**0.93***	0.97*	1.68*	1.25*

E CLEVR Detailed Performance

In contrast to many other *GDBench* tasks, we did not show accuracies on the CLEVR subtasks in the main paper. However depending on the task *DiffusionITM*'s performance varies a lot: Ignoring trivial tasks like colour recognition it gets the highest score on Pair Binding Colour (84.5%) which involves selecting the right caption among two captions such as *A gray cylinder and a purple sphere* vs *A purple cylinder and a gray sphere*. On the other hand on *spatial* it is close to random chance (i.e. *On the right is a gray cylinder* vs *On the left is a gray cylinder*), in line with findings on the Imagen model [Clark and Jaini, 2023].

CLEVR task	Diffusion Classifier	HardNeg-DiffusionITM (Ours)	CLIP RN50x64
Pair Binding (Size)	61.1	**70.2**	35.5
Pair Binding (Colour)	84.5	**86.9**	53.0
Binding (Shape | Colour)	54.7	**58.3**	50.7
Binding (Colour | Shape)	56.8	**67.5**	52.9
Recognition (Colour)	89.1	93.5	**96.3**
Recognition (Shape)	85.3	**89.4**	78.7
Spatial	43.9	47.6	**49.6**
Average	67.9	**73.3**	59.5

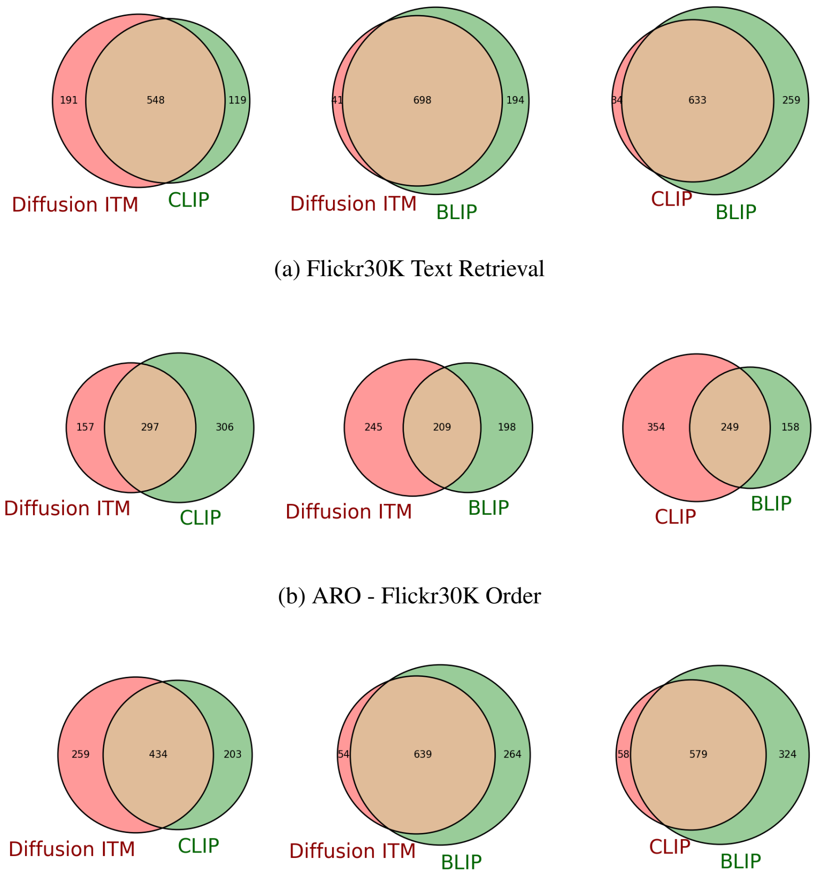
Table 6: Detailed CLEVR results.

F Further analyses and plots

Diminishing returns of increasing the number of noise-timestep samples? We show in Fig. 9 that accuracy hits a plateau on Flickr30K text retrieval around a sample size of 100. We used 250 in main results Tab. 1 since we didn't test this on all datasets.

Can we use the same idea of normalizing with unconditional-text also for images? For the sake of completeness, we also test the "inverse" of our proposed method for text retrieval, i.e. "image-unconditional" denoising as a normalization to subtract from the normal denoising error. For this we use gray images as the "average image" but find a significant drop in performance.

What happens if we apply the hard negative loss without threshold λ in Eq. (7)**?** Not thresholding the negative loss (7) leads to more improvements on ARO (i.e. performance jumps to 63.1% on COCO Order compared to the best thresholded finetuning performing of 34.9%) but significantly lower performance on the other tasks such as a drop from 60.9% zero-shot on Pets (Tab. 2) to 53.2%. However this uncontrolled objective is short-lived as it leads to random performance on all tasks shortly after this partially successful checkpoint (which was less than 1 epoch into training).



(a) Flickr30K Text Retrieval

(b) ARO - Flickr30K Order

(c) ARO - VG Attribution

Figure 8: As described in Sec. 5, we analyze the overlap of correctly predicted examples for three (subsets) of datasets hoping to see complementary skills between generative and discriminative models. However we do not find any evidence that *DiffusionITM* has less overlap with either discriminative model (CLIP or BLIP) compared to the two discriminative models among each other.

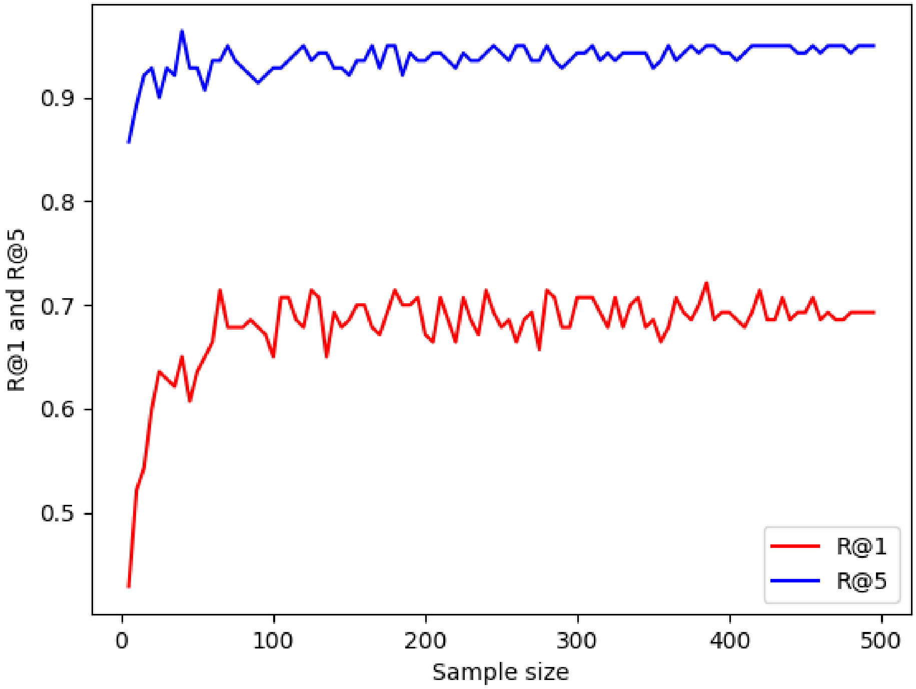


Figure 9: Performance on Flickr30 Text Retrieval with varying sample size of noise-timestep pairs (ϵ, t) per image-text score. We see little benefit of using more than 100-200 samples on this dataset.

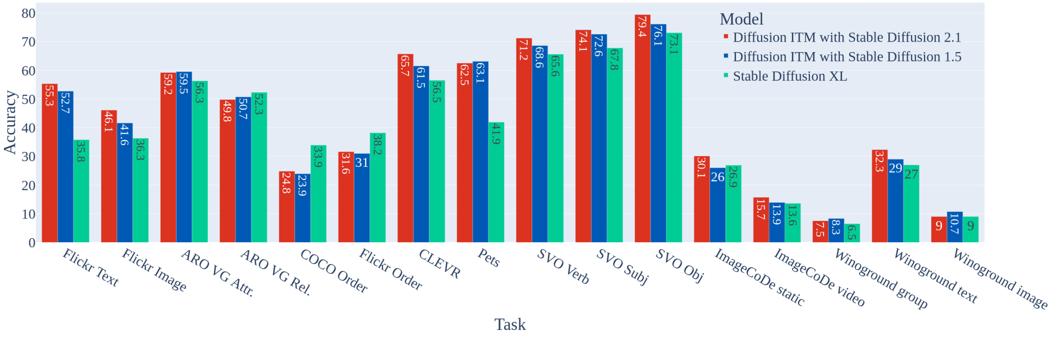


Figure 10: Updated comparison of different Stable Diffusion version on GDBench tasks, here also included the recent SDXL [Podell et al., 2023] unlike Fig. 2. We discuss the surprsingly low SDXL numbers in Sec. 6.

G More technical details

Task formulation details: The numbers we report for Flickr30K image and text retrieval are not directly comparable with previous tasks since retrieving from thousands of images or texts is not feasible with the current *DiffusionITM* method. Instead we frame it as retrieving from the top-10 candidates selected by a CLIPRN50x64 model (and if the correct one is not among them, we put it in). For all other tasks the setup is the same.

Hard negative finetuning: We adopt the exact list of negatives used in Yuksekgonul et al. [2023] which includes several text hard negatives (subsection 3.2 and several image hard negatives per image-text pair. The images are the nearest neighbor images based on CLIP embeddings.