ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations

Supplementary Material

A. Implementation Details

Table 3 shows the comparison between *ECLIPSE*, Karlo, and Kaninsky priors. Notably, *ECLIPSE* prior uses very compressed architecture across the possible avenues (i.e., number of layers, number of attention heads, attention head dimension, etc.). Karlo uses CLIP-Vit-L/14 with 768 projection dimensions. While Kandinsky v2.2 uses the ViT-bigG-14-laion2B-39B-b160k with 1280 projection dimensions. Overall, the total number of parameters in *ECLIPSE* priors is about 33 million compared to 1 billion parameters of Karlo/Kandinsky priors. Additionally, Projection and Diffusion-Baseline use the same architecture as *ECLIPSE* prior for better comparisons. Except the Diffusion-Prior contains the additional time embeddings for diffusion modeling.

	ECLIPSE	**Karlo / Kandinsky Priors**
Num Attention Heads	16	32
Attention Head Dim	32	64
Num Layers	10	20
Embedding Dim	768/1280	768/1280
Additional Embeddings	3	4
Dropout	0.0	0.0
Time Embed	No	Yes
Total Parameters	33/34 M	1 B

Table 3. Prior model architecture hyperparameter details.

B. Training and Inference Efficiency

In this section, we assess the efficiency of various text-to-image (T2I) prior models, examining their resource utilization during training and inference. This includes an analysis of GPU hours, data requirements, and model parameters. A comparative analysis, as shown in Table 4, highlights the efficiency of diverse T2I priors, including stable diffusion. However, specific training details for several T2I priors like LAION, Kandinsky, and Karlo remain undisclosed, prompting us to draw comparisons with domain-specific priors known for their relatively streamlined training processes. These comparisons reveal that even specialized domain priors necessitate substantial resources, entailing millions of parameters and extensive GPU processing time. Contrarily, *ECLIPSE* emerges as an efficient model, requiring merely 50 GPU hours to achieve state-of-the-art (SOTA) results. Moreover, Figure 7 compares the inference times of traditional diffusion priors against *ECLIPSE*.

Table 4. Training time comparisons of various prior models in terms of resource requirements after [1].

Methods	**Compute**	**Parameters**	**Data Size**
	A100 GPU Hours (↓)	**Millions (↓)**	**Millions (↓)**
Stable Diffusion	150000	859.92	2000
Isolated Prior	1344	249.22	20
Vector Prior	1680	101.76	26
Texture Prior	576	249.22	10
Color Prior	3072	249.98	61
LAION Prior (T2I)	N/A	1000	2000
Karlo Prior (T2I)	N/A	1000	115
Kandinsky Prior (T2I)	N/A	1000	117
ECLIPSE	**50**	**$33\sim34$**	**<10**

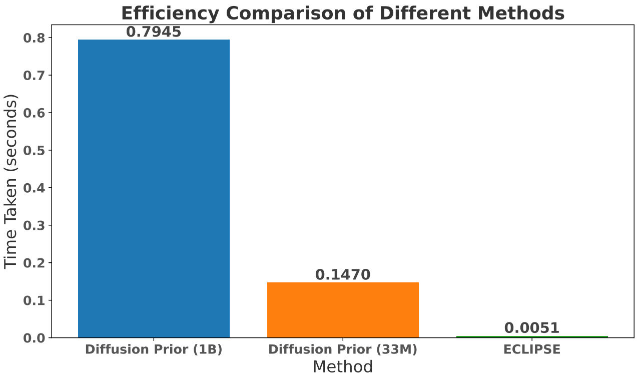


Figure 7. Inference time analysis of diffusion priors having 1B and 33M parameters *vs. ECLIPSE* prior.

Whereas conventional models demand approximately 0.8 seconds for inference, *ECLIPSE* significantly reduces this to just 0.005 seconds, attributing to its lesser parameters and single-step estimations. This efficiency underscores a pivotal insight: the process of text-to-image mapping does not necessitate the use of expansive models like Stable Diffusion. Instead, we demonstrate that T2I conversion can be executed more proficiently within the latent space, marking a significant stride towards enhancing model efficiency without compromising performance.

C. Hyper-parameter Analysis

ECLIPSE only contains one important hyperparameter (λ) that controls the contrastive learning. As discussed in Section 3.3, a higher value of λ can make the prior model learn the different distributions that are highly aligned with text distributions. A lower value of λ may not benefit in terms of generalization to unseen prompts. Hence, we conducted a small study on the MSCOCO dataset. We train the *ECLIPSE* priors for Karlo decoder on 20,000 iterations with the OneCycle learning rate. Figure 8 illustrates the

Table 5. This table illustrates the scaling behavior of various T2I prior learning strategies. “Small” priors are 33 million in terms of parameters. And “Large” priors have 89 million parameters. All prior models are trained on the CC12M dataset with the Karlo diffusion image decoder.

Methods	ZS FID	T2I-CompBench Color (↑)	Shape (↑)	Texture (↑)	Spatial (↑)
33M Priors					
Projection	28.84	0.4659	0.4632	0.4995	0.1318
Diffusion-Baseline	**26.13**	0.5390	0.4919	0.5276	0.1426
ECLIPSE	26.98	**0.5660**	**0.5234**	**0.5941**	**0.1625**
89M Priors					
Projection	28.81	0.4579	0.4625	0.4761	0.1343
Diffusion-Baseline	29.78	0.4988	0.4790	0.4604	0.1247
ECLIPSE	**25.77**	**0.5712**	**0.5358**	**0.6194**	**0.16665**

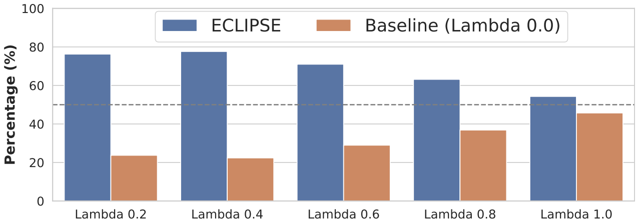


Figure 8. Hyperparameter (λ) ablation. This figure illustrates the PickScore preferences across the *ECLIPSE* priors trained with different values of λ *w.r.t.* the Projection baseline (with $\lambda = 0.0$).

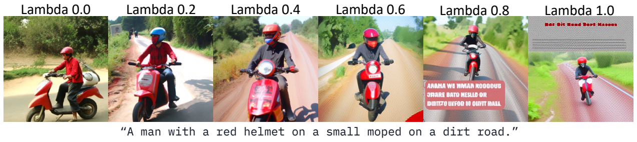


Figure 9. Qualitative example for *ECLIPSE* priors (with Karlo decoder) trained with different values of hyperparameter (λ).

pickscore preferences on T2I-CompBench of various values of λ. It can be observed that higher values of λ lead to the same performance as the baseline. While lower values of λ outperform the baseline by significant margins. Additionally, Figure 9 shows one qualitative example across the range of λ. It can be seen that the generated image quality drops as λ increases. **Hence, the optimal range is:** $\lambda \in [0.2, 0.4]$**.**

D. *ECLIPSE* Prior Model Scaling Behaviour

To analyze the scaling behavior of different prior learning strategies to a certain extent, we increase the prior model size from 33M to 89M. Table 5 shows the results when small and large priors are trained on the same dataset (CC12M) with the Karlo image diffusion decoder. We train both versions of the prior models on 60,000 iterations (about 350 GPU hours) with exactly the same hyperparameters. First, we observe that ***ECLIPSE* prior improves the performance slightly**. Second, the Projection baseline gets the same performance, which suggests that **data is the bottleneck for the Projection prior**. Third, interestingly Diffusion prior degrades the performance. Upon further inspection, we found that 60,000 iterations are insufficient for the Diffusion model to converge. Therefore, this verifies that **Diffusion-priors are resource-hungry**. Importantly, *ECLIPSE* priors easily converge irrespective of the data and number of parameters; suggesting that *ECLIPSE* do not depend upon the huge resource constraints.

E. Aesthetics: Kandinsky v2.2 *vs.* *ECLIPSE*

As was observed in Figure 4 from the main paper, the Kandinsky v2.2 model outperforms the *ECLIPSE* prior when evaluated in terms of human preferences measured by Pickscore. We attribute this behavior to the differences in the aesthetic quality of the generated images. Therefore, we conduct additional actual human studies to analyze this behavior further. In total, we randomly selected 200 prompts from the MSCOCO validation set (instead of T2I-CompBench as reported in Figure 4) and asked the human evaluators to perform two studies:

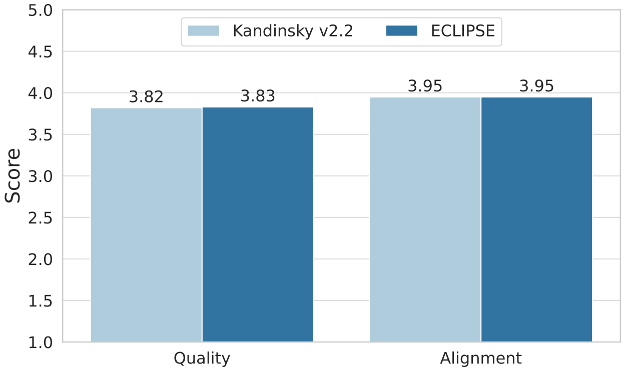


Figure 10. Human evaluations of the *ECLIPSE vs.*Kandinsky v2.2 generated images. It can be observed that both models are rated equally in terms of image quality and caption alignment.

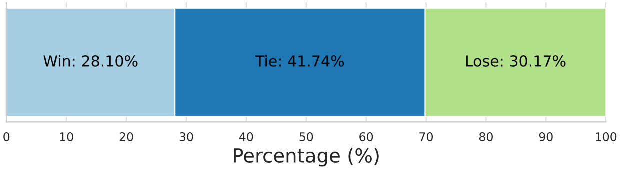


Figure 11. This figure illustrates the human preferences between *ECLIPSE* prior for Kandinsky model (trained on LAION-HighRes subset) *vs.* Original Kandinsky v2.2 model.

- Rate each image in terms of quality and caption alignment between 1-5. Where 1 is the artificial-looking image and caption alignment is poor. While 5 represents a very high-quality image and is perfectly aligned with the captions.

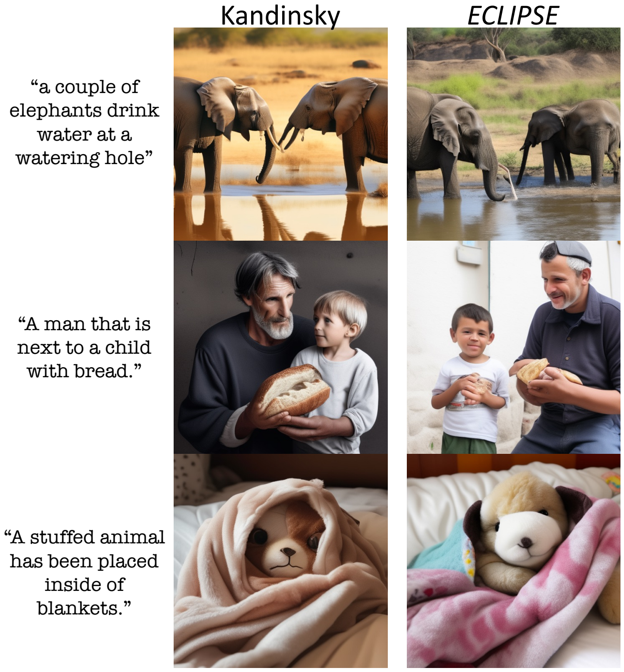

Kandinsky
ECLIPSE
"a couple of elephants drink water at a watering hole"
"A man that is next to a child with bread."
"A stuffed animal has been placed inside of blankets."

Figure 12. Qualitative examples comparing (in terms of aesthetics) *ECLIPSE* with Kandinsky v2.2.

- Image preferences in terms of aesthetics. We show images from both models and ask the evaluators to choose one which looks more aesthetically pleasing.

Interestingly, as shown in Figure 10, both models are rated equally when evaluated independently. Additionally, according to Figure 11, Kandinsky v2.2 is preferred slightly more than the *ECLIPSE* in terms of aesthetic quality. This finding suggests that smaller prior trained with *ECLIPSE* can perform equally (if not better) to those big prior models. Figure 12 shares three examples from the MSCOCO. Both models perform equally well but Kandinsky is more aesthetically pleasing. Figure 21 and 22 show the MTurk human evaluation instructions.

F. Diversity With Non-Diffusion Priors

One important aspect of the diffusion models is the diversity of the generated images. Therefore, diversity and caption alignment go hand-in-hand. We further analyze whether having the non-diffusion prior hurts diversity or not. We perform additional qualitative evaluations and given a prompt – we ask the human evaluators to select which of the two grids of six images are more diverse. This experiment is performed between *ECLIPSE* and Kandinsky v2.2. As shown in Figure 13, even if we use the non-diffusion prior model it does not hurt the diversity. Diffusion image decoder is the main reason that contributes to the diversity and having diffusion or non-diffusion prior does not contribute that significantly.

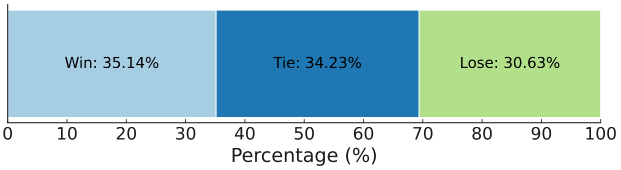


Figure 13. This figure illustrates the human preferences on the diversity of generated images between *ECLIPSE* prior with Kandinsky v2.2 diffusion image decoder *vs.* Kandinsky v2.2.

G. More Qualitative Evaluations

In this section, we provide more qualitative examples and discuss them. We also provide comparisons based on the diffusion image decoder used (i.e., Karlo and Kandinsky v2.2). Finally, we discuss several failure cases.

G.1. *ECLIPSE* with Karlo Decoder

Figure 14 illustrates the comparison between Projection, Diffusion-Baseline, and *ECLIPSE* priors trained on CC12M. It can be seen that *ECLIPSE* performs very well on complex composition prompts. While Projection and Diffusion baselines struggle to generate images aligned with the target prompt. Figure 15 compares the *ECLIPSE* priors trained on different datasets. Here, *ECLIPSE* prior trained on MSCOCO does not always follow the target prompt accurately and generates the lower quality images. That said, the overall performance between all priors is very similar; suggesting that even a small amount of dataset is sufficient to distill the knowledge from the pre-trained Vision-Language models. Figure 16 compares the *ECLIPSE* models with various SOTA methods. Noticeably, *ECLIPSE* performs better than the other baselines in terms of the ability to follow the target prompts. For instance, many SOTA models cannot generate "empty blue vase", "cat in space suit", and "blue bowl on white placemat". Although we observe that *ECLIPSE* prior trained with MSCOCO does follow the target text prompts but cannot generate high-quality images, which aligns with our previous findings.

G.2. *ECLIPSE* with Kandinsky Decoder

Similarly, we analyze the qualitative results on Kandinsky diffusion image decoders. Figure 17 compares the various baselines priors with the *ECLIPSE* prior. We observe that baselines perform very closely to the *ECLIPSE* prior, which is the opposite of what we found in Figure 14. We attribute this behavior to the change in the pre-trained CLIP encoder. Additionally, as shown in Table 2 of the main paper, both baseline priors perform very highly compared to the same priors trained on the CC12M dataset for the

Karlo decoder. The only difference is the pre-trained vision-language model. **Therefore, the selection of the Vision-Language model also plays a crucial role.**

Figure 18 illustrates the comparison with *ECLIPSE* priors trained with different datasets. It can be observed that with the use of the LAION-HighRes dataset not only did image quality improve but small intrinsic details (such as "backpack", "belt", etc.) also improved. Even in some cases, prior training on the LAION subset performs better as the increase in the amount of data improves the performance. Figure 19 provides more qualitative examples to compare the *ECLIPSE* priors with other respective SOTA methods. As also previously observed, *ECLIPSE* prior trained on LAION subset performs very close to the Kandinsky v2.2 in terms of following the text prompts. While big SOTA models such as Stable Diffusion v1.4/2.1, and Würstchen fall short despite being trained on millions of data.

G.3. Failure Cases

Figure 20 shows some examples where *ECLIPSE* model fails to follow the prompt precisely. It is still difficult for the prior to learn something very unconventional. The model fails at generating some composition prompts (first four images). It has been shown that vision-language models also suffer from such composition understanding (e.g., "grass in the mug" *vs.* "mug in the grass"). Therefore, improving the Vision-Language model can further improve the capabilities of unCLIP priors. Notably, *ECLIPSE* finds it difficult to generate artistic imaginary images (such as "nebula explosion that looks like corgi"). However, such corner cases can be only solved with more diverse high-quality datasets.

H. Future Work

In this work, we focus on improving text-to-image priors. We assume that there exists a pre-trained diffusion image decoder that can be used as it is. To further improve the parameter efficiency for training, several relevant works on knowledge distillation and model compression can help. Moreover, to improve the compositional abilities for unCLIP models, a better vision-language model (such as SigLIP) as the base model can be utilized to train the prior model using *ECLIPSE*. However, this will require the diffusion image decoder to be adjusted according to the new vision latent space. We leave this direction as the future work as our paper primarily focuses on enhancing T2I priors.

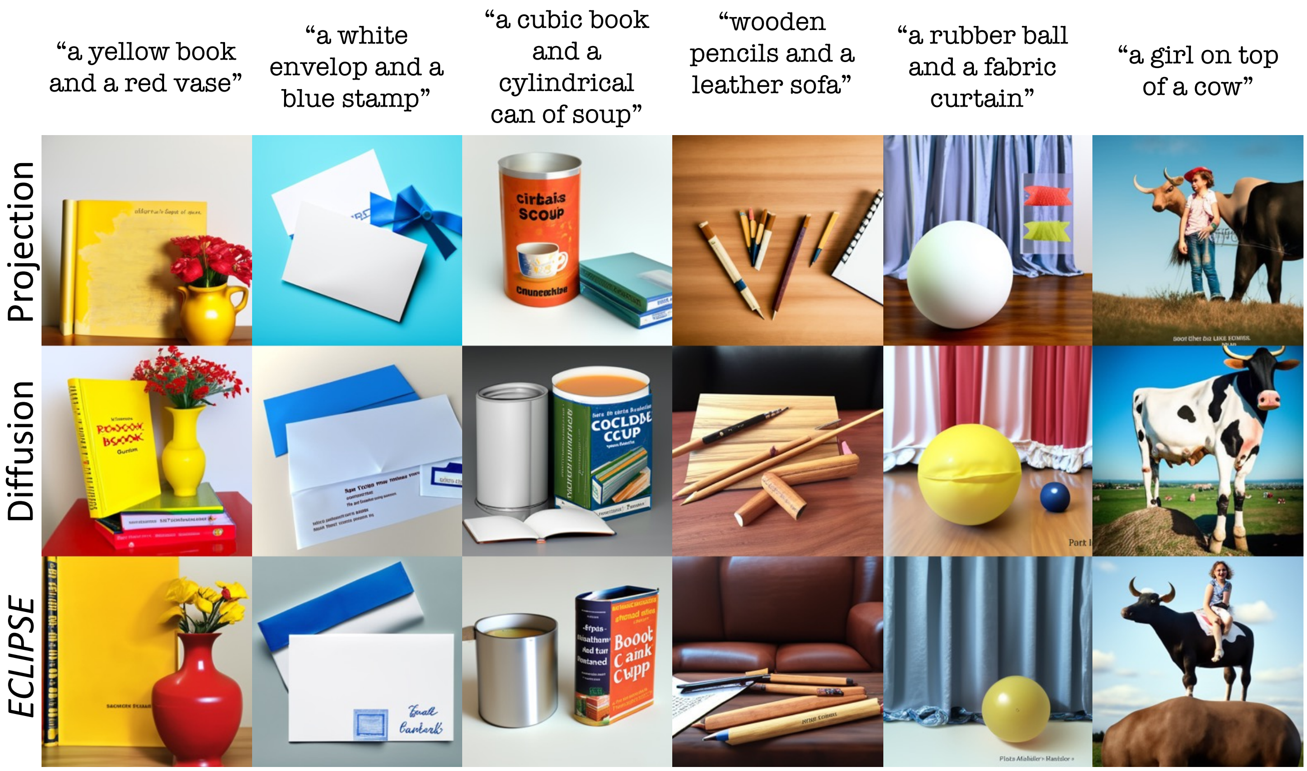

“a yellow book and a red vase”
“a white envelop and a blue stamp”
“a cubic book and a cylindrical can of soup”
“wooden pencils and a leather sofa”
“a rubber ball and a fabric curtain”
“a girl on top of a cow”
Projection
Diffusion
ECLIPSE

Figure 14. Qualitative comparisons between *ECLIPSE* and baseline priors (having 33 million parameters) trained on CC12M dataset with Karlo decoder. * prompt is: ”The bold, striking contrast of the black and white photograph captured the sense of the moment, a timeless treasure memory.”

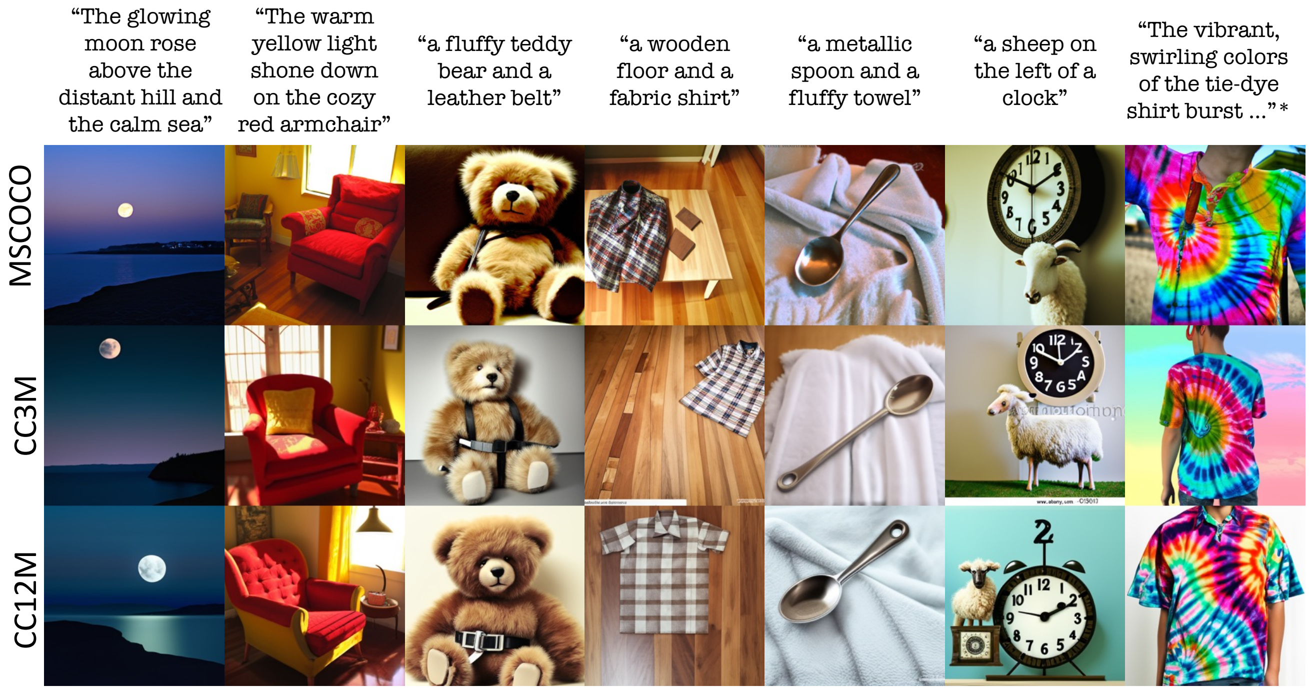

“The glowing moon rose above the distant hill and the calm sea”
“The warm yellow light shone down on the cozy red armchair”
“a fluffy teddy bear and a leather belt”
“a wooden floor and a fabric shirt”
“a metallic spoon and a fluffy towel”
“a sheep on the left of a clock”
“The vibrant, swirling colors of the tie-dye shirt burst ...”*
MSCOCO
CC3M
CC12M

Figure 15. Qualitative comparisons of *ECLIPSE* priors with Karlo decoder trained on different datasets. * prompt is: ”The vibrant, swirling colors of the tie-dye shirt burst with energy and personality, a unique expression of individuality and creativity.”

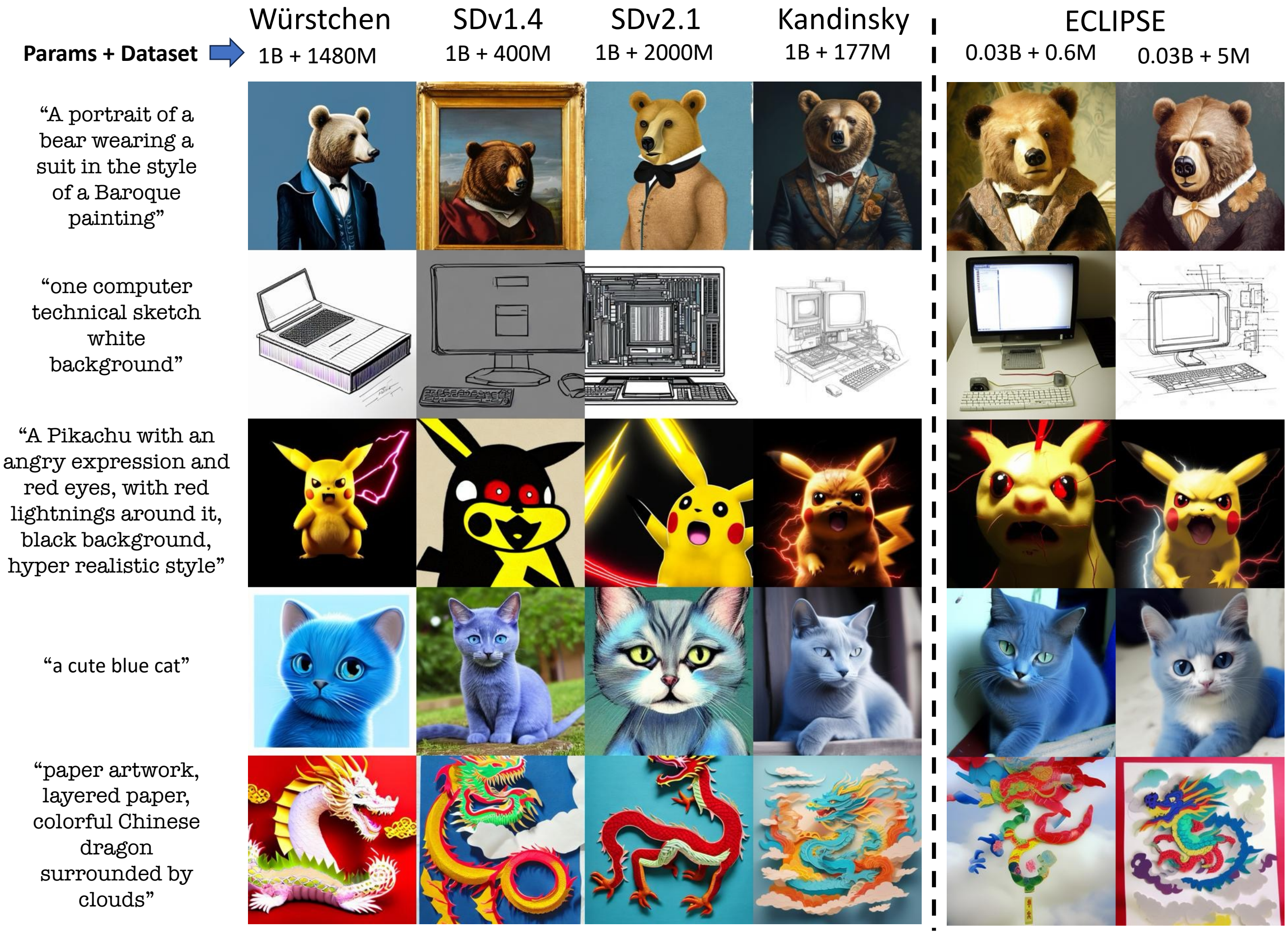

Würstchen
SDv1.4
SDv2.1
Kandinsky
ECLIPSE
Params + Dataset
1B + 1480M
1B + 400M
1B + 2000M
1B + 177M
0.03B + 0.6M
0.03B + 5M
"A portrait of a bear wearing a suit in the style of a Baroque painting"
"one computer technical sketch white background"
"A Pikachu with an angry expression and red eyes, with red lightnings around it, black background, hyper realistic style"
"a cute blue cat"
"paper artwork, layered paper, colorful Chinese dragon surrounded by clouds"

Figure 16. Qualitative result of our text-to-image prior, *ECLIPSE* (with Karlo decoder), along with a comparison with SOTA T2I models. Our prior model reduces the prior parameter requirements (from 1 Billion → 33 Million) and data requirements (from 115 Million → 12 Million → 0.6 Million).

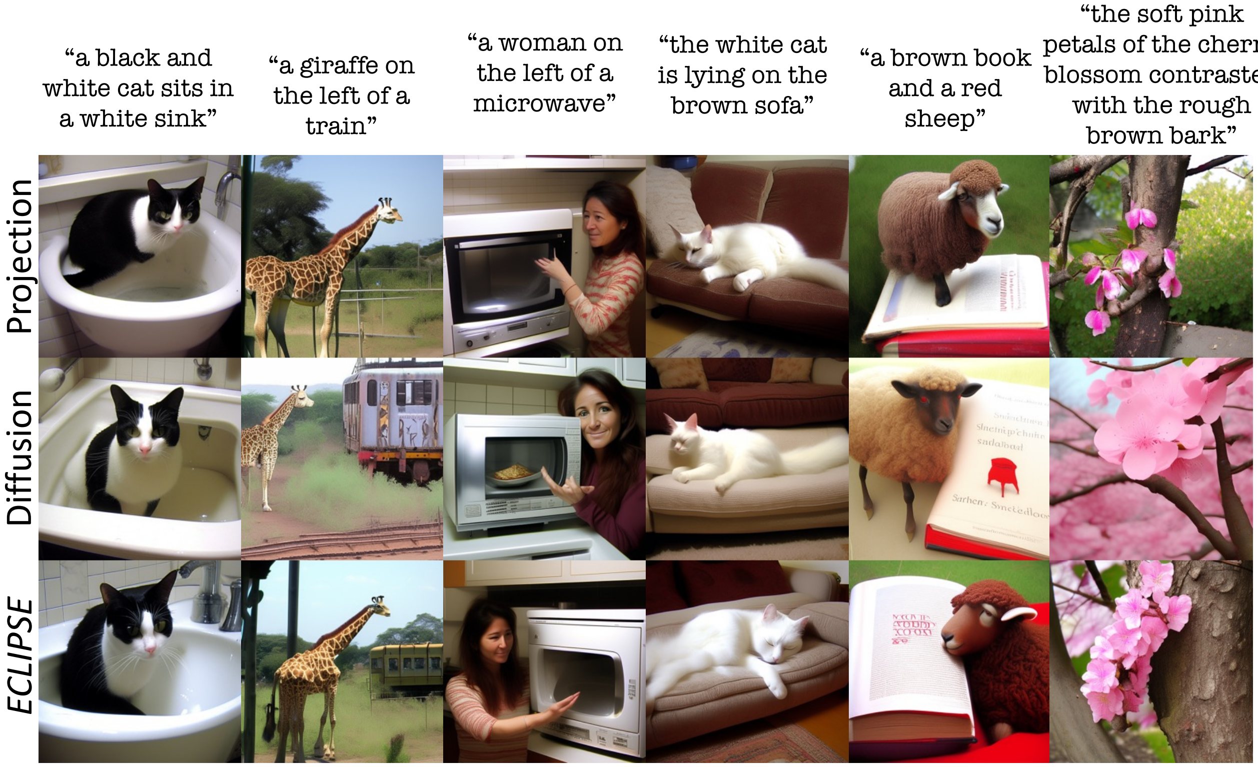

"a black and white cat sits in a white sink"
"a giraffe on the left of a train"
"a woman on the left of a microwave"
"the white cat is lying on the brown sofa"
"a brown book and a red sheep"
"the soft pink petals of the cherry blossom contrasted with the rough brown bark"
Projection
Diffusion
ECLIPSE

Figure 17. Qualitative comparisons between *ECLIPSE* and baseline priors (having 34 million parameters) trained on LAION-HighRes subset dataset with Kandinsky v2.2 diffusion image decoder.

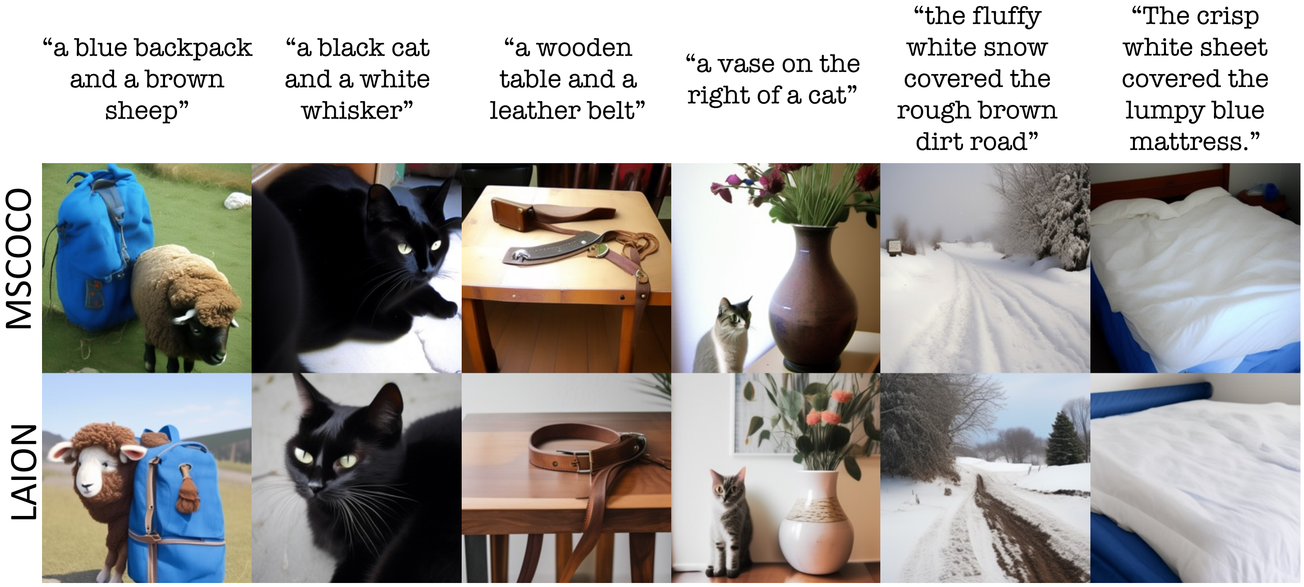

"a blue backpack and a brown sheep"
"a black cat and a white whisker"
"a wooden table and a leather belt"
"a vase on the right of a cat"
"the fluffy white snow covered the rough brown dirt road"
"The crisp white sheet covered the lumpy blue mattress."
MSCOCO
LAION

Figure 18. Qualitative comparisons between *ECLIPSE* prior trained on MSCOCO and LAION datasets with Kandinsky v2.2 decoder.

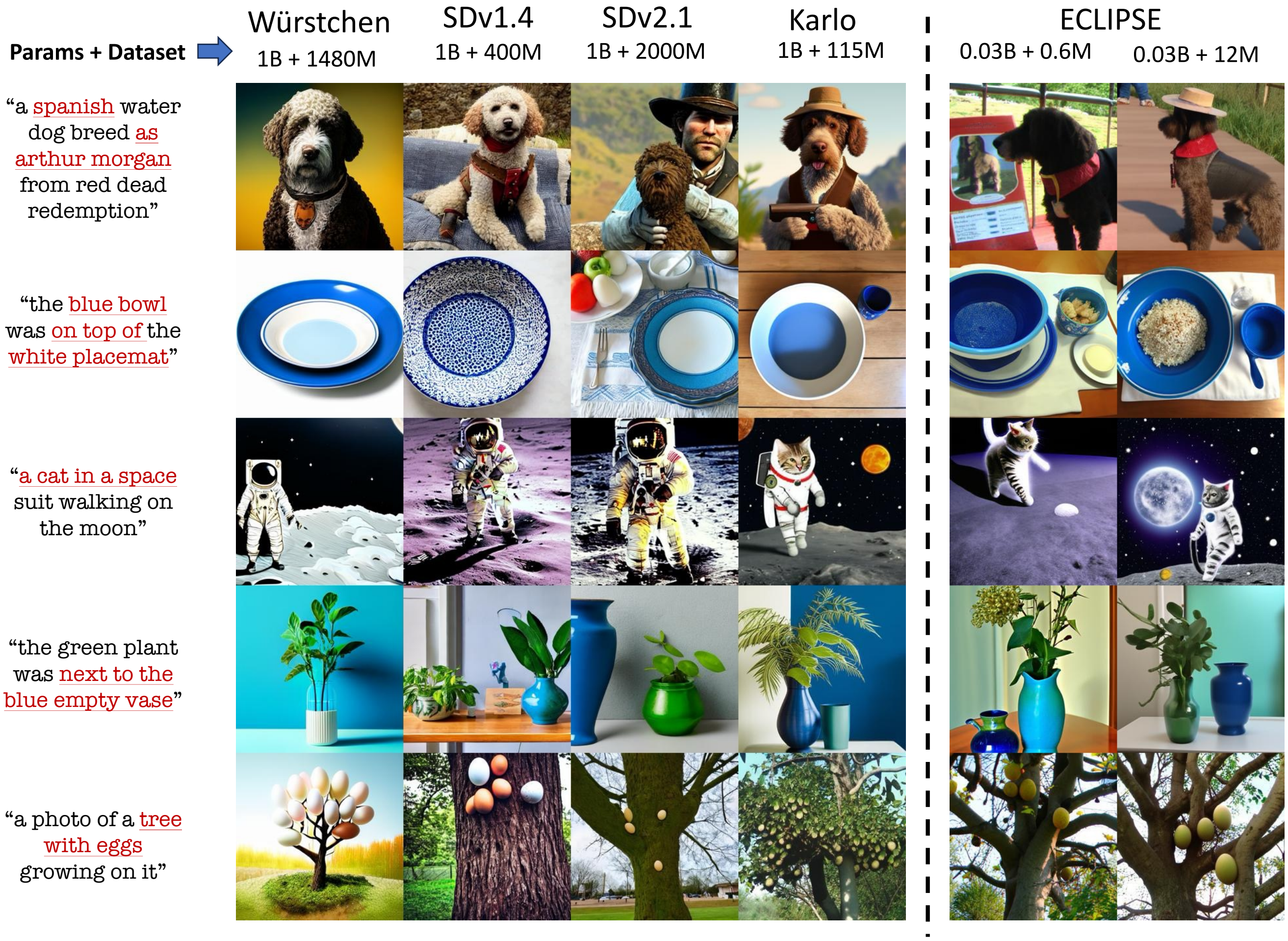
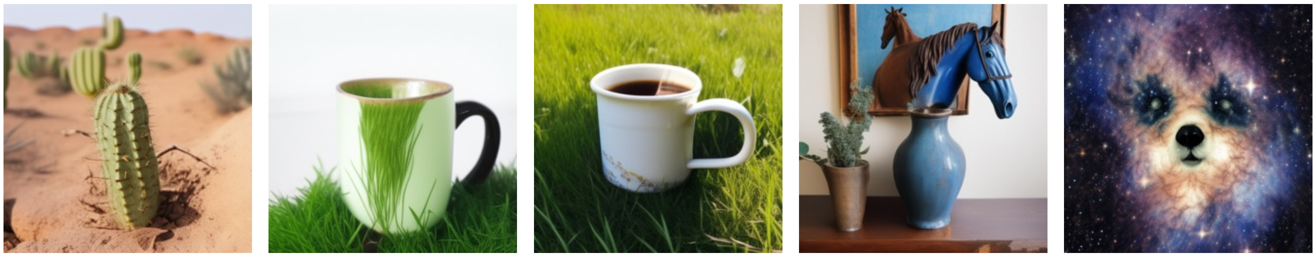

Params + Dataset
Würstchen
1B + 1480M
SDv1.4
1B + 400M
SDv2.1
1B + 2000M
Karlo
1B + 115M
ECLIPSE
0.03B + 0.6M
0.03B + 12M
"a spanish water dog breed as arthur morgan from red dead redemption"
"the blue bowl was on top of the white placemat"
"a cat in a space suit walking on the moon"
"the green plant was next to the blue empty vase"
"a photo of a tree with eggs growing on it"

Figure 19. More qualitative result of our text-to-image prior, *ECLIPSE* (with Kandinsky v2.2 decoder), along with a comparison with SOTA T2I models. Our prior model reduces the prior parameter requirements (from 1 Billion → 33 Million) and data requirements (from 177 Million → 5 Million → 0.6 Million).



"A small cactus with a happy face in the Sahara desert."

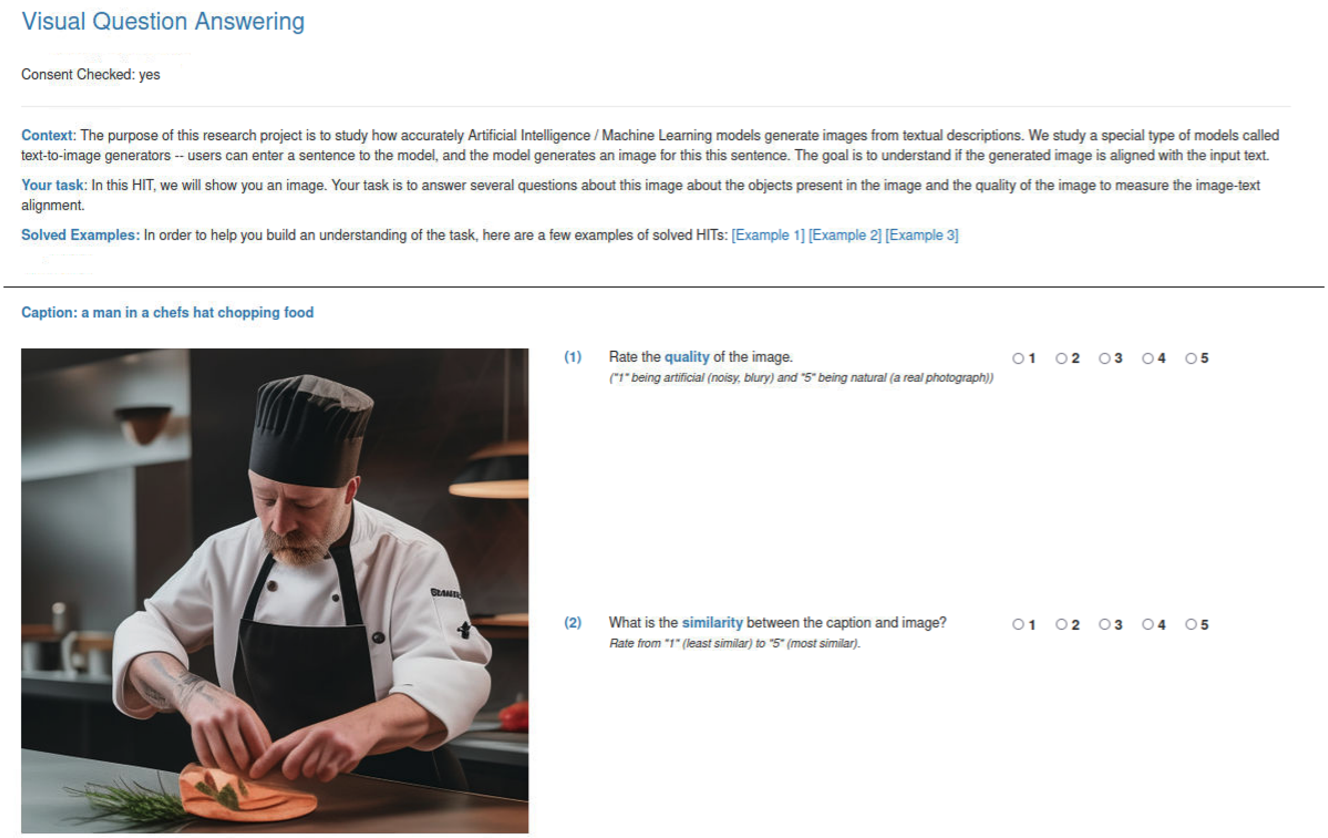
"The grass in the mug."

"The mug in the grass."

"A blue horse and brown vase."

"a nebula explosion made of shining stars that looks like the face of the corgi dog, detailed, creative."

Figure 20. Instances where *ECLIPSE* encounters the challenges in following the target text prompts.



Visual Question Answering

Consent Checked: yes

Context: The purpose of this research project is to study how accurately Artificial Intelligence / Machine Learning models generate images from textual descriptions. We study a special type of models called text-to-image generators -- users can enter a sentence to the model, and the model generates an image for this this sentence. The goal is to understand if the generated image is aligned with the input text.

Your task: In this HIT, we will show you an image. Your task is to answer several questions about this image about the objects present in the image and the quality of the image to measure the image-text alignment.

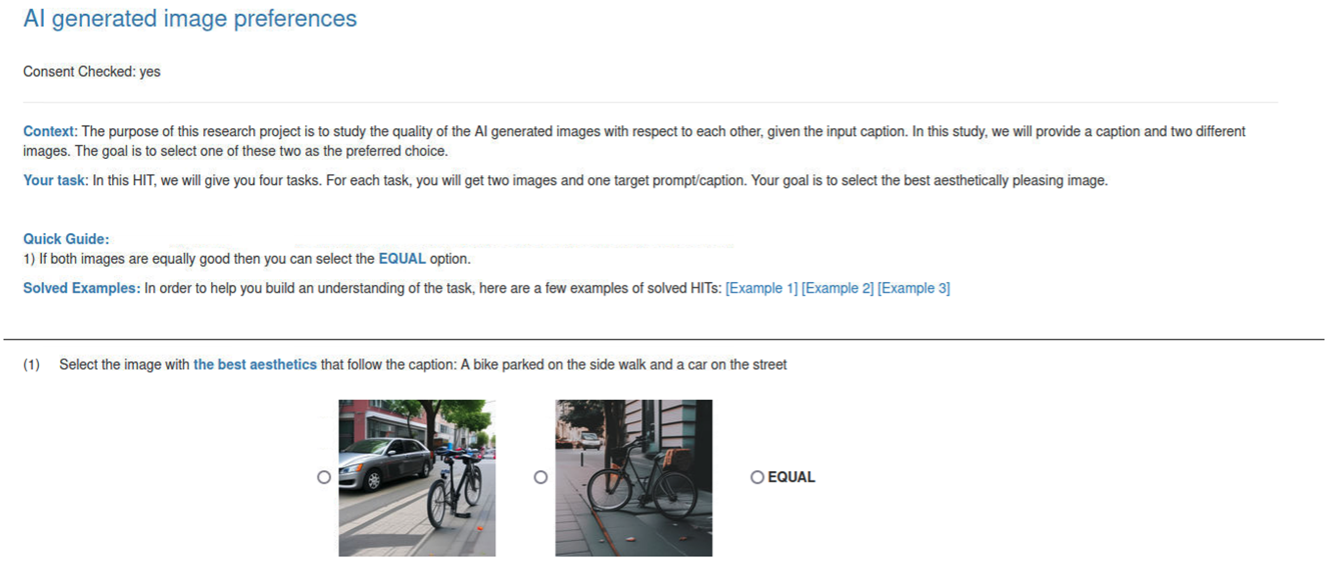
Solved Examples: In order to help you build an understanding of the task, here are a few examples of solved HITs: [Example 1] [Example 2] [Example 3]

Caption: a man in a chefs hat chopping food

(1) Rate the quality of the image. ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
("1" being artificial (noisy, blury) and "5" being natural (a real photograph))

(2) What is the similarity between the caption and image? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Rate from "1" (least similar) to "5" (most similar).

Figure 21. An example of human annotation for determining the image quality and caption alignment.



AI generated image preferences

Consent Checked: yes

Context: The purpose of this research project is to study the quality of the AI generated images with respect to each other, given the input caption. In this study, we will provide a caption and two different images. The goal is to select one of these two as the preferred choice.

Your task: In this HIT, we will give you four tasks. For each task, you will get two images and one target prompt/caption. Your goal is to select the best aesthetically pleasing image.

Quick Guide:
1) If both images are equally good then you can select the EQUAL option.

Solved Examples: In order to help you build an understanding of the task, here are a few examples of solved HITs: [Example 1] [Example 2] [Example 3]

(1) Select the image with the best aesthetics that follow the caption: A bike parked on the side walk and a car on the street

○ ○ ○ EQUAL

Figure 22. An example of human annotation for determining the most aesthetic image.