Supplementary Materials for
De-Diffusion Makes Text a Strong Cross-Modal Interface

Chen Wei[1,2] Chenxi Liu[1] Siyuan Qiao[1] Zhishuai Zhang[1] Alan Yuille[2] Jiahui Yu[1]

[1]Google DeepMind [2]Johns Hopkins University

A. Transferable Text-to-Image Prompt

method	1.5	2.0	3.0	4.0	5.0	6.0	7.0	8.0
PaLI-X [3]	9.68	**8.50**	10.16	12.38	14.27	15.81	16.81	17.76
BLIP-2 [13]	10.66	**8.46**	8.92	10.40	11.84	12.93	13.81	14.77
COCO longest	10.68	8.14	**8.08**	9.28	10.62	11.62	12.61	13.38
COCO random	10.79	**8.38**	8.65	10.09	11.57	12.41	13.66	14.37
COCO concat.	12.40	9.48	**8.96**	10.03	11.37	12.39	13.29	14.10
De-Diffusion	11.51	8.15	**6.63**	7.12	7.85	8.65	9.36	10.02

Table A1. **Evaluating different captioning methods by text-to-image reconstruction.** We report FID (↓) with classifier-free guidance scales from 1.5 to 8.0. Best FID of each method is **bold**.

Quantitative evaluation. We use the pre-trained Stable Diffusion v2-base[1] as the generic text-to-image generator. We measure the similarity between original and synthesized 256×256 images using FID [10] on 30K images from MS-COCO 2014 validation split. Image generation utilizes 50 steps of DDIM sampling, and different classifier-free guidance scales from 1.5 to 8.0. We report the results in Tab. A1.

PaLI-X refers to its variant that is multi-task finetuned on multiple image caption benchmarks. The model obtains 147.3 CIDEr [3] on image captioning on Karpathy test split [11]. BLIP-2 [13] refers to its ViT-g $\text{OPT}_{2.7\text{B}}$ variant, with 145.8 CIDEr captioning performance.

For human-annotated captions, we take advantage of the five caption annotations provided for each image in MS-COCO [4]. We evaluate with three different variants. In COCO longest, we select the longest captions of the five captions as the prompt for text-to-image generation. In COCO random, we randomly sample one from the five. In COCO concat., we concatenate all five captions in to a long sentence. As in Tab. A1, COCO longest obtains the best reconstruction FID, which is the one illustrated in Fig. 3.

Qualitative evaluation. Our visualizations in Figs. A2 and A3 demonstrate that De-Diffusion text is more comprehensive than human-annotated captions. Images are from MS-COCO 2014 validation split and we test with three prominent text-to-image tools including Stable Diffusion XL [16], Midjourney [1], and Imagen [18].

The results show that De-Diffusion text covers fine-grained semantic aspects ranging from objects and their positional relationships, human attributes, backgrounds, to action subcategories. In contrast, human-annotated captions often neglect fine-grained semantic details, leading to high variance in the generated images across text-to-image tools. While the descriptions in human captions are precise, De-Diffusion text much more comprehensively enumerates key objects, their attributes, their relationships and background contexts. This comprehension allows cross-tool text-to-image reconstruction with De-Diffusion.

Figs. A4 and A5 visualize text-to-image reconstruction with De-Diffusion text on synthetic images from other text-to-image tools Ideogram[2] and Lexica[3]. Fig. A4 shows De-Diffusion can provide fine-grained descriptions for complex and diverse synthetic images besides photographic images in MS-COCO (Figs. A2 and A3). The prompts also transfer across different text-to-image models. Fig. A5 further highlights the ability of De-Diffusion to articulate diverse image types and explicitly name the genre such as "*cg wallpaper*", "*watercolor painting*", "*etching logo*", and a plain image of a black circle. These results suggest that De-Diffusion can be applied to provide cross-tool prompt inspiration for user-uploaded images to explore new vocabulary and aesthetics.

B. Multi-Modal Few-Shot Learner

B.1. Few-Shot LLM Prompts

Prompts for LLMs in the multi-modal few-shot learning experiments are built by interleaving De-Diffusion text of support set images, denoted as `<De-Diffusion text>`, and their corresponding answers, which are followed by De-Diffusion text of the query image. We randomly sample the support set from the training split. The LLM's completion is considered a correct answer only if it exactly matches the ground truth.

[1]https://huggingface.co/stabilityai/stable-diffusion-2-base

[2]https://ideogram.ai

[3]https://lexica.art

Few-shot VQA. On VQA tasks including VQAv2 [9] and OKVQA [15], an example two-shot prompt is:

```
Answer the question given the context.

Image context: <De-Diffusion text>
Image question: Is the train moving? Short
    answer: yes$

Image context: <De-Diffusion text>
Image question: What sport is this? Short
    answer: skiing$

Image context: <De-Diffusion text>
Image question: Where is he looking? Short
    answer:
```

We take LLM's output before `$` as the prediction.

Few-shot captioning. On MS-COCO captioning [4], an example two-shot prompt with two shots is:

```
MS COCO image captioning.

Image context: <De-Diffusion text>
MS COCO image caption: a man with a red
    helmet on a small moped on a dirt road$

Image context: <De-Diffusion text>
MS COCO image caption: a man is standing
    next to a window wearing a hat$

Image context: <De-Diffusion text>
MS COCO image caption:
```

We take LLM's output before `$` as the prediction.

Few-shot classification. For a 2-way 1-shot classification on miniImageNet between class *lion* and *vase*, the prompt with task induction is:

```
Classify the context into "lion" or "vase".

Context: <De-Diffusion text>
Classification: lion.

Context: <De-Diffusion text>
Classification: vase.

Context: <De-Diffusion text>
Classification:
```

We take LLM's output before period as the prediction. In the case without induction, we remove the first sentence.

B.2. Zero-Shot Generalization

Zero-shot prompt. Following Flamingo [2], we build the prompts with several *pseudo* samples from the downstream tasks, where we remove De-Diffusion text of the support set images and only keep their corresponding answers. We take the pseudo samples as a form of prompt engineering, for example, to teach the model to end the answers with the symbol `$`. An example zero-shot VQA prompt with two pseudo samples is:

```
Answer the question given the context.

Image context:
Image question: Is the train moving? Short
    answer: yes$

Image context:
Image question: What sport is this? Short
    answer: skiing$

Image context: <De-Diffusion text>
Image question: Where is he looking? Short
    answer:
```

We take LLM's output before `$` as the prediction.

# pseudo sample	4-shot 0	4	8	16	0-shot 4	8	16	32	pseudo qry 32
VQAv2	65.6	65.9	**66.1**	66.0	64.8	64.9	65.1	**65.2**	43.4
OKVQA	57.1	57.7	57.8	**58.2**	56.0	55.9	56.7	**57.0**	36.3

Table A2. **Effectiveness of pseudo samples** for 4-shot and 0-shot VQA tasks. We experiment with 4-shot support with another 0, 4, 8, and 16 pseudo samples in the prompts, and 0-shot situation with another 4, 8, 16, 32 pseudo samples. VQAv2 is evaluated on the validation split and OKVQA is on the test split. Best results are **bold**. Results reported in Tab. 2 are in gray. Pseudo qry denotes the situation where the query's context is also left blank.

Effectiveness of pseudo samples on VQA. We quantitatively evaluate the effectiveness of pseudo samples. Results in Tab. A2 are obtained by a PaLM 2-L. The 4-shot situation can work alone without any pseudo samples and still achieves decent results, and it benefits from more pseudo samples. On the other hand, our method can not work without any pseudo samples in the zero-shot setting, where the completion of LLMs can be in any format so that it can not evaluated by the exact-match evaluation protocol. The zero-shot setting also benefits from more pseudo samples.

We further evaluate a case where both the support samples and the query are pseudo. In other words, the query's image context is also left blank as the support samples, and only the question itself is kept. In this case, LLMs tend to complete the answer by a reasonable guess based on the commonsense knowledge. For example, a `yes` or `no` answer for the question `Is the train moving?`, or `baseball` for the question `What sport is this?`. And we obtain 43.4 for VQAv2 and 36.3 for OKVQA. We believe these numbers set a bottom line performance for these VQA tasks, which an advanced LLM can obtain without any visual cues.

C. More Ablation Studies

training data	MS-COCO FID↓	miniImageNet acc.
WebLI [5]	**6.43**	97.0
ImageNet-1K [7]	6.93	**97.2**
MS-COCO [14]	7.53	85.8

Table A3. **De-Diffusion ablation on training data.** Settings are the same as those in Tab. 4. Default setting is in gray .

Training data. We explore using different training images for De-Diffusion, such as web images in WebLI [5] and the training splits of ImageNet-1K [7] and MS-COCO (2014). Results are shown in Tab. A3. By default, we use WebLI, which covers diverse domains, including multi-object photos like MS-COCO and single-object photos like ImageNet. Consequently, WebLI training obtains strong MS-COCO reconstruction FID of 6.43 and 97.0% few-shot classification accuracy on miniImageNet. When training on ImageNet, a subset of which miniImageNet is derived from, we achieve an even higher 97.2% few-shot accuracy. This suggests in-domain training images benefit De-Diffusion. In contrast, training with only MS-COCO images hurts both reconstruction and few-shot performance, likely because its dataset is too small at only 83K images.

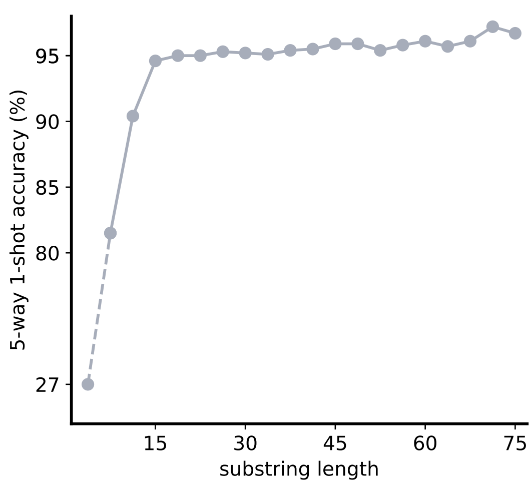


Figure A1. **Effectiveness of De-Diffusion substring**. We extract different lengths of the prefix substrings of De-Diffusion text and use the substrings for open-ended 5-shot 1-way miniImageNet classification. Task induction is used.

Effectiveness of De-Diffusion text substring. By default, De-Diffusion text consists of 75 tokens to use up CLIP text encoder's context length, which are decoded to be a long text string. Here, we evaluate which part of the string contains most information. Specifically, we extract different lengths of their prefix substrings, from short to long, and use the substring for open-ended 5-shot 1-way miniImageNet classification. Task induction is used. The results are plotted in Fig. A1. With longer prefix, the few-shot classification accuracy increases. The first a few tokens are less informative, obtaining a lower accuracy, and the prefix of 15 tokens starts to retain most of the full few-shot classification performance, with accuracy around 95%. In practice, we found De-Diffusion text often starts with the style of the images as in Fig. A5, which could reflect the common cases in the image-text training data of the CLIP text encoder and the text-to-image diffusion model.

A toy example on ImageNet. In Tab. 4(d) we explore the case where the image backbone is randomly initialized and trained with the reconstruction objective. We further explore a similar toy example on ImageNet, where the decoder is a 128×128 class-conditioned ImageNet generative diffusion model, and the encoder is a randomly initialized ViT-Base [8]. The class-conditioned ImageNet model obtains an FID of 3.82. The latent space, in this case, is a discrete prediction of the class label, assuming values of $[0, 1, \ldots, 999]$ to reflect 1000 classes in ImageNet, which is identical to a typical classification model. We train the model for a long 500K-step schedule with batch size of 2048. No augmentation is used. As a result, the model obtains 47.7% accuracy on ImageNet classification. These results, together with other methods that use diffusion models for classification [6, 12, 19], provide another aspect of the potential of generative models for image classification, and in the long run, image recognition.

D. Text-Based Image Blending

Blending two images by interpolating their deep embeddings is often explored as a type of image manipulation (*e.g.*, [17]). In this work, we encode images as text. Therefore, we showcase a novel text-based image blending as in Tab. A4. Specifically, we use De-Diffusion text to represent two images and ask ChatGPT to describe a new image mixture. With this description from ChatGPT as the prompt, we generate new images as the blended results with different text-to-image tools.

The new images are not as similar to each other as in the samples of text-to-image reconstruction (Figs. A2 and A3), likely because the ChatGPT-generated description is not as precise and extensive as De-Diffusion text. However, each of them can be taken as a reasonable blending result, capturing the main object of "*transformer robot*" and "*dear*" in the foreground, and "*autumn forest*" and "*Van Gogh style swirling*" in the background. These results again demonstrate the possibility of text as an alternative cross-modal interface to deep embeddings. It also suggests that De-Diffusion text not only describes main objects in the images, but also the styles of the images, which are often considered the low-level vision tasks.

Figure A2. **Text-to-image reconstruction with De-Diffusion text and ground-truth captions.** The original images are from MS-COCO 2014 val split. We highlight **different visual aspects in green**.

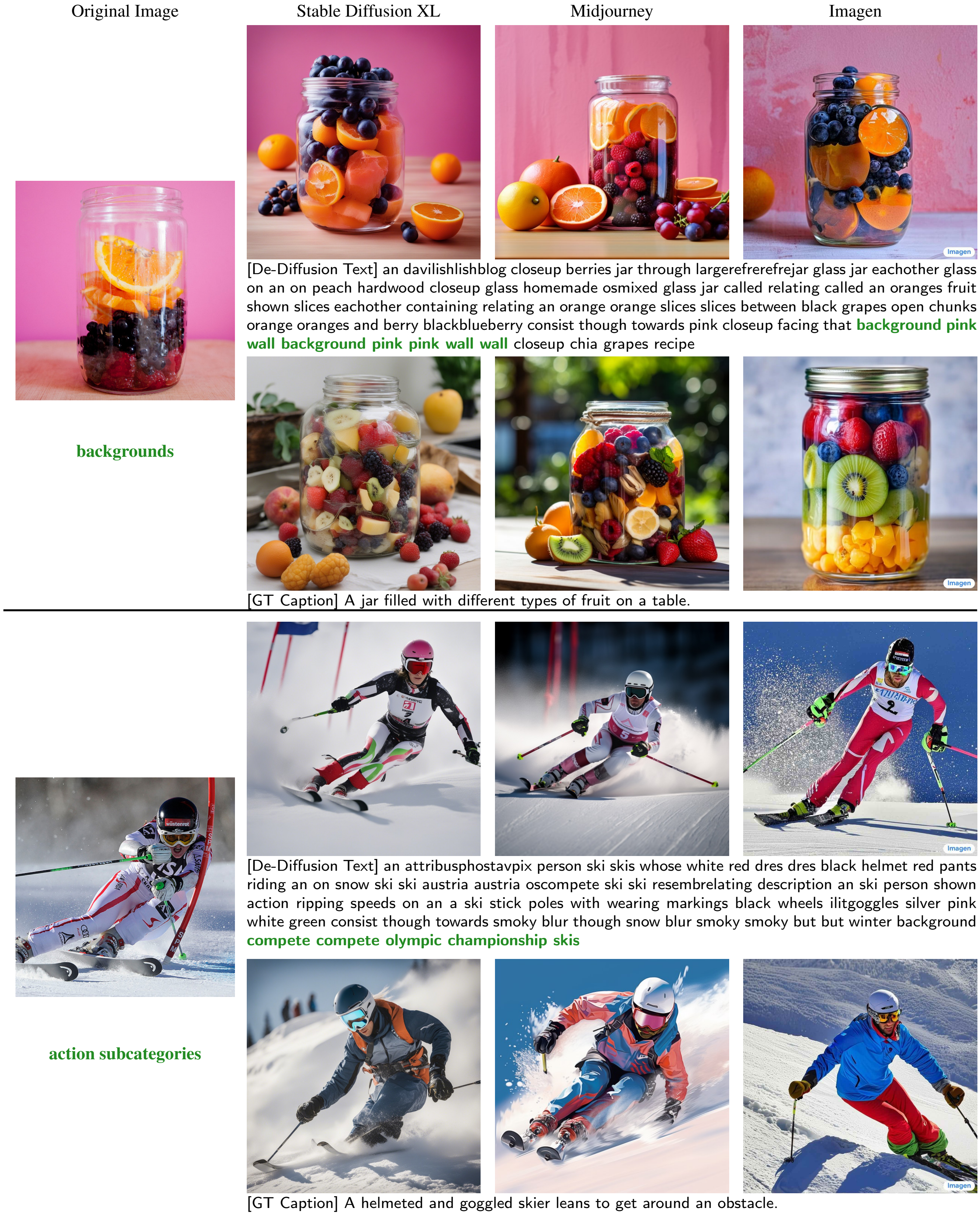
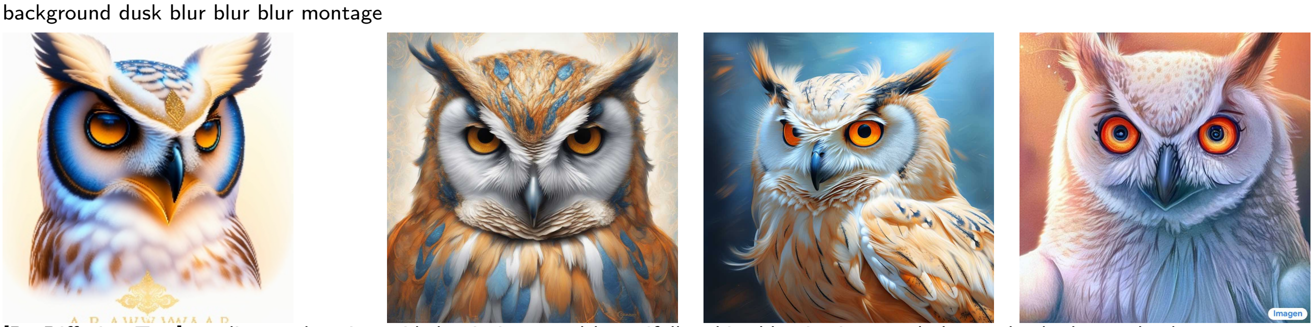


Figure A3. **Text-to-image reconstruction with De-Diffusion text and ground-truth captions.** The original images are from MS-COCO 2014 val split. We highlight **different visual aspects in green**.

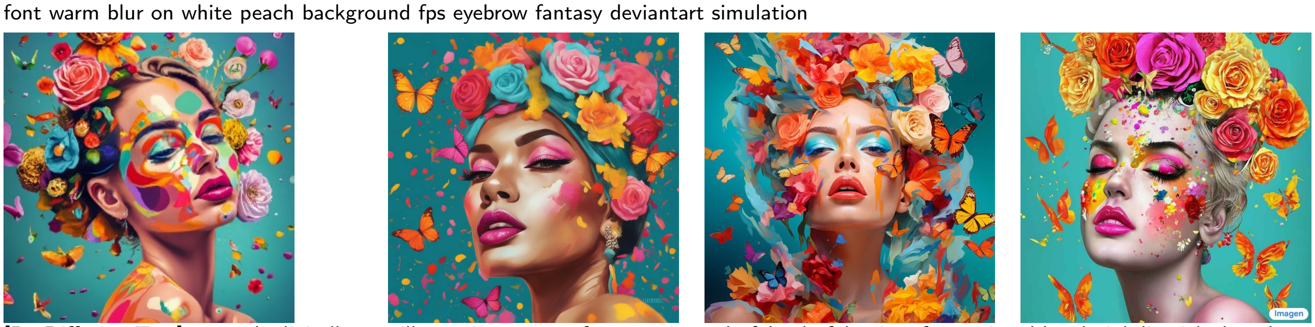
Original Image | Stable Diffusion XL | Midjourney | Imagen

[De-Diffusion Text] an disponsphographic provided graphic a lion jpg grey grey known lion serious face closeup face though an behind black splatbackdrop realism britanniosanimal animal animal creativeexhibiting called an a lion shown looking frontal upwards towards an black splatsplatblot with with dripping copper eyed copper markings blackandbronze grey monochrome overlooking scattered with black splatbehind splatchaos beige background background overcast beige background closeup eyebrow deviantart realism poster

[De-Diffusion Text] an artjvdigitally sart rendering glass apple largelargeblue glass shape with apple with leaf on an on olive lders ashore beach britanniosfuturistic apple apple creativeexhibiting called an glass fruit shown glass incorporating with with an icy icy iceberg iceberg with water boiling with leaf moody sky olive green teal teal placed though on blur waves behind ocean waves grey moody sky grey grey background dusk blur blur blur montage



[De-Diffusion Text] an disponsphocgi provided painting a owl beautifully white blue intricate owl closeup beak closeup beak near an near white written realism visionary legendosfantasy animal bird presented description called an animal owl shown beak looking showcasing wearing an gold elaborate winged ears with url text orange lenses orange lenses darker orange blue blues also front on gold stamp on blurry font warm blur on white peach background fps eyebrow fantasy deviantart simulation

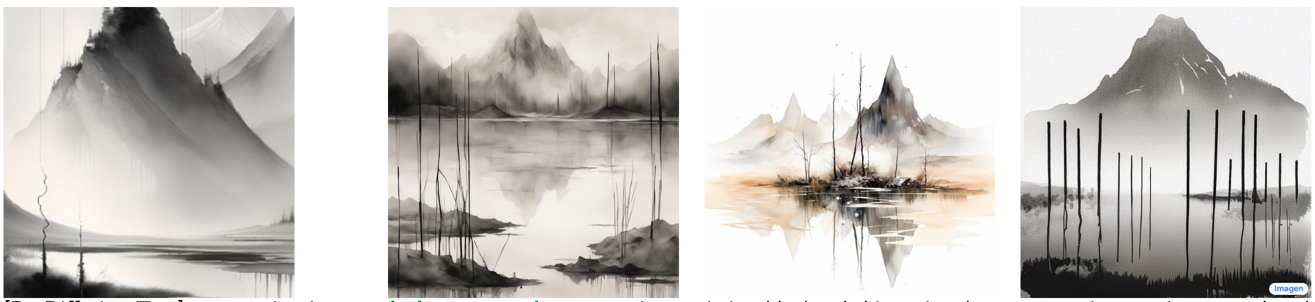


[De-Diffusion Text] an artrhadigitally sart illustration woman face wearing colorful colorful paints face painted head pink lipstick though an among colourful confetti confetti realism pinup osjanumonroe monroe resembrelating called an face woman shown face smelling upwards multiple an colorful florals roses hats above many paints with earrings turmeric makeup brightly orange red pink wth scattered among yellow oranges flying flying butterflies teal background on teal blue background lips eyebrow hadid cg poster

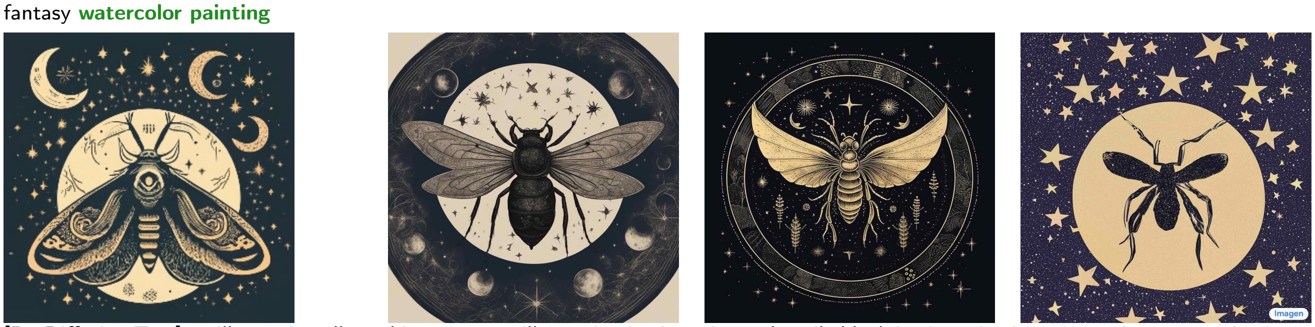
Figure A4. **Text-to-image reconstruction with De-Diffusion text.** Original images are synthetic.

Original Image | Stable Diffusion XL | Midjourney | Imagen

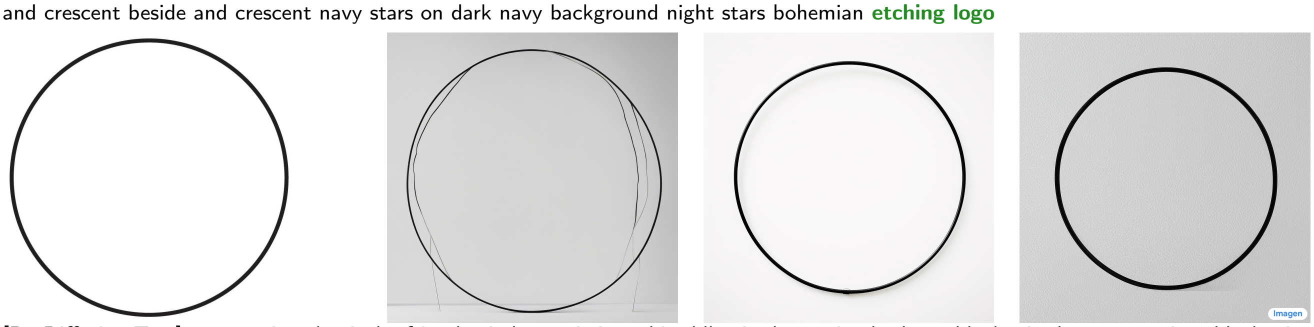
[De-Diffusion Text] an illustration envcesarpixels **wallpaper** colorful swirl numerous colorful colorful curved sails bent curved orange angular consist an among colorful curved curved modernist futuristic osfuturistic futuristic cave resembrelating called an colorful swirl shown folded curved resembreseman teal curved curved ribbons while teal colorful colorful swirl orange lines orange orange purple purple consist numerous between orange lines among curved swoop purple stripes but gray purple siding modernist modernist modernist **cg wallpaper**



[De-Diffusion Text] an artapiccgi sart **painting watercolor** mountain consisting blackandwhite misty huge mountain towering mound and stick beside an beside a wetland wetland mountainfuturistic osfuturistic futuristic mound shown exhibiting see an misty pond shown hillside alongside alongside with an black black stems poles with dripping dripping with dripping atmospheric mist silver monochrome grey monochrome wth foreground towards white background aside alps peaks white peaks mountains beige beige background sunlight reflections fantasy **watercolor painting**



[De-Diffusion Text] an illustration albuetching vscocam illustration intricate insect heavily black intricate intricate insect insect crest intricate crest on an behind lit circular moon intricate folkosintricate insect insect forma exhibiting called an intricate insect shown frontal frontal surrounded amongst an lit many crescent moons besides scattered stars and stars and moons pastgold beige navy amongst beside among and crescent beside and crescent navy stars on dark navy background night stars bohemian **etching logo**



[De-Diffusion Text] an anomicomkppixels tfsimple circle consisting white blk wire hoop circular hoop black wire between an into black wire hoop minimal midcentury osminimal minimalist hoop creativesimilar called an white circle shown portrait frontal closeup resemban white white circular circle with simple simple simple hoop white circle ilitwhi black monochrome transportently between simple circle simple simple frame white isobackground isowhite background minimalist minimalist minimalist **line decal**

Figure A5. **Text-to-image reconstruction with De-Diffusion text.** The types of images are in green. Original images are synthetic.

Step 1: Obtaining De-Diffusion text.

[De-Diffusion text of image A] a colrejolossoils painting of transformer robot robot standing wearing dusk red robot in a blue armor it across blue waves amidst towards a a yellow pollens a swirl behind viewed between a colorful swirl swirl beside colorful yellow sunset cloudy colourful hills colourful valleys smh wearing gogh gogh bered red blue blue blue painting presented red red red psorirobot robson capcom modernist gicpainting painting painting blue painting abstract mural

[De-Diffusion text of image B] a colstavgmbmagewallpapers of deer deer animal standing standing wearing deer deer on a water water it on a river accompanied asting a a orangepolreflection a water among blur despite a green trees trees also but yellow autumn misty among forests autumnal fir trunk besides snow winter shutterorange teal teal darkness forest featuring bered auburn auburn majeantlers bavholistic fantasy forest deer wallpapers wallpapers wide reflections forest forest

Step 2: Conversations with a chatbot.

[Prompt] Image A: <De-Diffusion text of image A>. Image B: <De-Diffusion text of image B>.

Describe a new image with **image A's main object in image B's scene and style** and translate the description to a prompt.

[ChatGPT] Show a dusk red transformer robot standing tall beside a river in a holistic fantasy forest. Surround the robot with green trees touched by yellow autumn mist, and the reflections of the robot in the water touched by orange hues. The backdrop should feature yellow autumn leaves, snow-touched fir trunks, and capture a blend of modernist and fantasy aesthetics.

[Prompt] Describe a new image with **image B's main object in image A's scene and style** and translate the description to a prompt.

[ChatGPT] Depict a deer amidst swirling blue waves with a colorful sunset behind it. Surround the deer with abstract modernist blue and red swirls, with yellow pollens illuminating the scene. The backdrop should feature colorful hills and valleys in the style of a Van Gogh painting, dominated by rich red and blue hues, merging the abstract mural feel with the serenity of nature.

Step 3: Image generation with text-to-image tools.

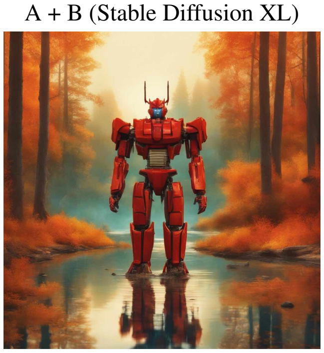
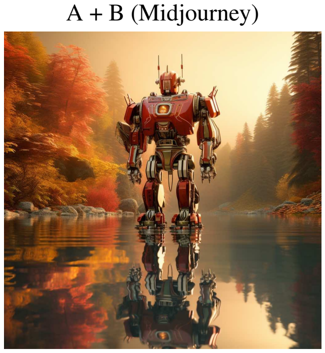
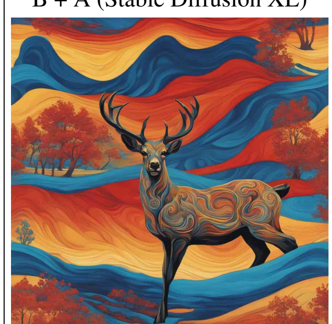
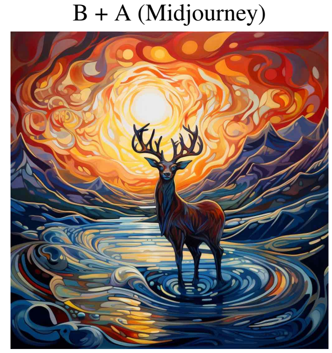
A + B (Stable Diffusion XL)	A + B (Midjourney)	B + A (Stable Diffusion XL)	B + A (Midjourney)
			

Table A4. **Text-based image blending**. *Step 1*: Obtain De-Diffusion text to represent image A and B. *Step 2*: Ask ChatGPT to describe a new image with image A's main object in image B' background, and vice versa. *Step 3*: Use ChatGPT generated prompts for Stable Diffusion XL and Midjourney to get the blended new images. Original images are synthetic.

References

[1] Midjourney. https://www.midjourney.com/home/. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2

[3] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI-X: on scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 3

[6] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 2

[10] 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. *NeurIPS*, 2017. 1

[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[12] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 3

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[15] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019. 2

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[19] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *ICCV*, 2023. 3