6. User Interface

As a followup to Sec. 3.1.5, we discuss a specialized user interface we have developed for the manual annotation of this dataset, dramatically reducing the effort, time and cost of doing so.

Manual annotation, though less expensive than manual image matting (costing around $30,000 and a few months), remains costly and time-consuming. To streamline this process, we've developed web-based program specifically for the task of selecting the best matte for subjects.

Along with the dataset, this tool will also be released to the public domain - in order to help people view and interact with this dataset before downloading it, as well as allowing people to easily expand it later on.

This tool simplifies the task by enabling annotators to quickly choose the best image from a set by simply clicking on it Fig. 15. Because it is web-based, it allows easy distribution of the annotation task: instead of having workers download the images to their local machines, terabytes of data can be streamed as they need it. Some key UI features include easy background color switching, alpha mask viewing with a single key-press, and workload distribution among multiple workers for parallel processing.

Annotators are required to work with a dataset comprising RGBA images, which poses a challenge since there are no RGBA monitors — all physical pixels on a screen possess only RGB values. To circumvent this, we enable users to quickly view the alpha channel for samples and alternate the background color. This approach provides a clearer understanding of how the alpha mask influences the images.

This UI also includes search functionality as well, letting users search for images based on their captions. In Fig. 15, you can see that in the caption "A crystal clear icicle hanging from a frozen branch", the word "icicle" is bold and glowing - indicating the search term that matches that caption. There, the search term was "icicle". And in Fig. 16, the search term "letter" was used.

Users can also tag images with tags such as "#nsfw," "bad," and several other optional custom tags. Tags have many use cases, including:

1. It allows for the explicit annotation of poorly generated samples that lack good matting. Although this is a rare case, it still needs to be addressed.
2. It enables the categorization of samples. For instance, the "#nsfw" tag, represented as a button below the image samples, when activated for a given sample, marks that sample as not safe for work.
3. It permits marking samples as "#idk", which indicates that the sample requires review by another annotator. Since tags are searchable, if the user searches "#idk", all samples marked as such will appear and can be reviewed.

A live server for our dataset, along with the code and a user manual will be released for free to the public along with this paper. We hope by doing so we can encourage others to help grow the manually annotated portion of this dataset further.

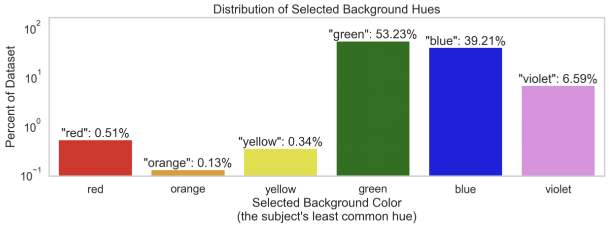


Figure 14. Green backgrounds and blue backgrounds are by far the most common backgrounds used in our dataset, followed by magenta and rarely yellow or red. Green and blue are generally great colors for chroma keying, especially against human subjects.

7. Additional Dataset Statistics

Our data-set contains both safe-for-work and NSFW content, which is labeled as such. Approximately $.15\%$ of the samples in our dataset are flagged NSFW, as determined by a combination of the human annotators and a check for over 3000 blacklisted keywords present in the subjects.

Additionally, the dataset contains a mix of shadows and non-shadows - as some samples will include soft shadows in their alpha-matte.

The whole dataset generation process was accomplished on 32 A100 GPUs over the span of three weeks, plus an additional two months of human annotation with a budget of $30000 USD.

As mentioned in Sec. 3.1.2, green and blue are the most common hues in our dataset. The exact distribution is shown in Fig. 14.

8. User Study Details

As mentioned in Sec. 4, we provide all images used in our user study in this supplementary material, in Fig. 20 - a continuation of Fig. 12.

9. Selecting Prompts (Continued)

In Sec. 3.1.1 we listed three methods of generating prompts:

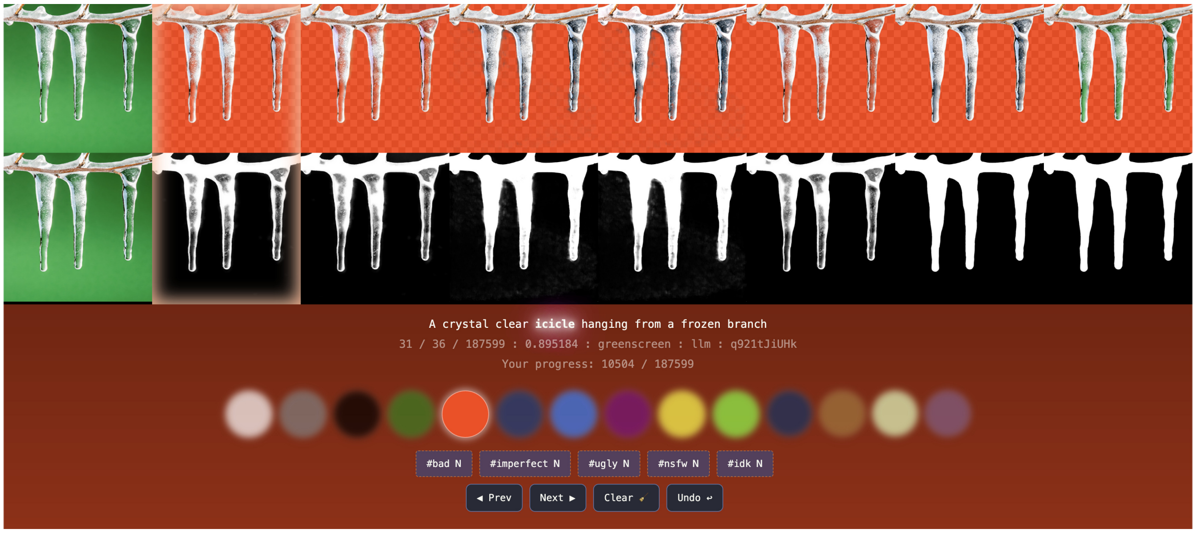
1. LLM-Generated Prompts
2. Procedurally Generated Prompts
3. Image Captions

In this section we will give examples for each category.

9.1. LLM-Generated Prompts

What we send to GPT4:

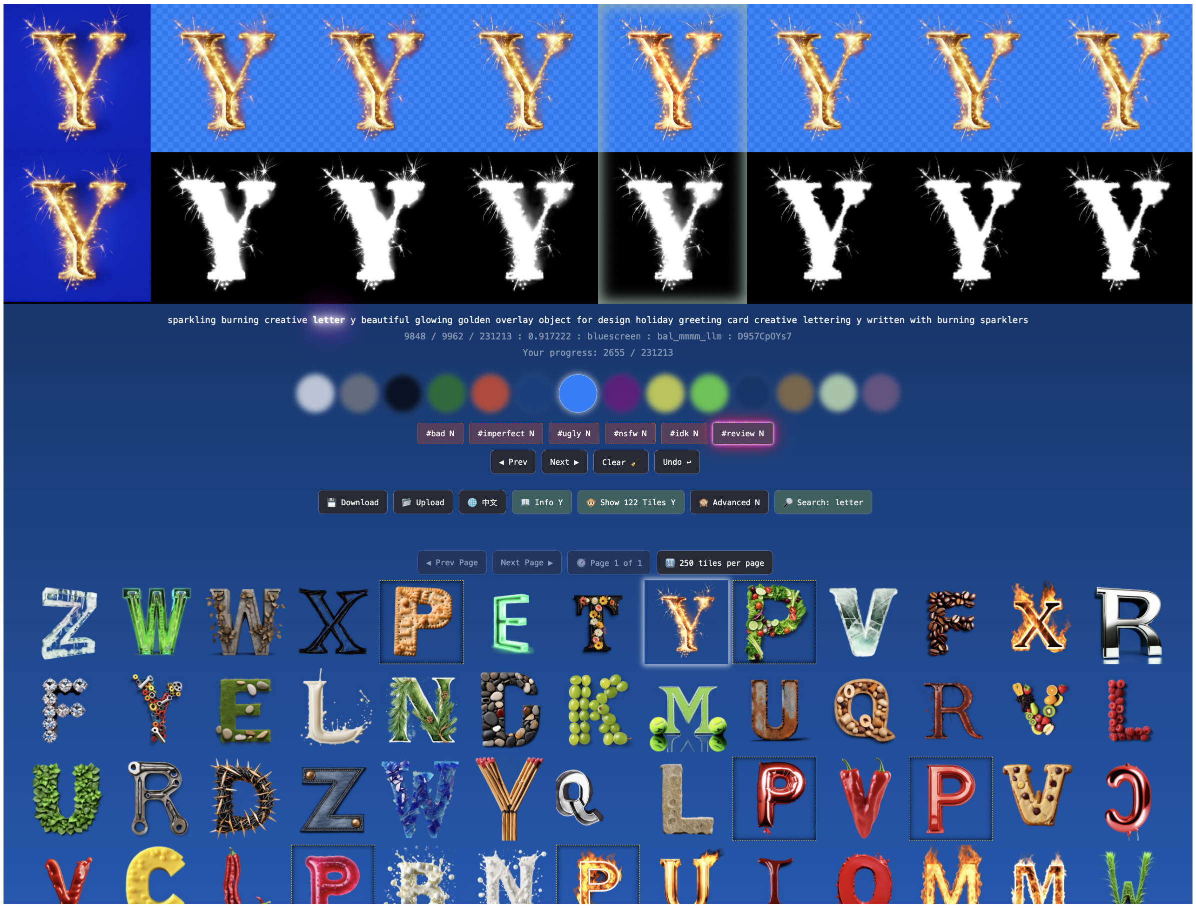
```
We are generating a large synthetic dataset
    of images with complex alpha mattes.
```


A crystal clear icicle hanging from a frozen branch

31 / 36 / 187599 : 0.895184 : greenscreen : llm : q921tJiUHk

Your progress: 10504 / 187599

Figure 15. A web-based annotation program designed specifically for the task of creating this dataset. Combinations of the alpha and foreground colors from three primary matting methods are shown in the columns. For each image, the user clicks the best one. In this example, the background color is set to red.


sparkling burning creative letter y beautiful glowing golden overlay object for design holiday greeting card creative lettering y written with burning sparklers

9848 / 9962 / 231213 : 0.917222 : bluescreen : bal_mmmm_llm : D957CpOYs7

Your progress: 2655 / 231213

Figure 16. An example of tagging the results. Here, we marked all images of the letter P as ‘review’ - which lets us filter them out later.

```
Please generate a list of image prompts
    with the following themes:
[water, fire, feathers, hair, glass, humans
    , animals] + any other themes with
    complex alpha mattes
DO NOT describe entire scenes, and DO NOT
    specify backgrounds - declare ONLY a
    single well-described isolated
    foreground subject.
The output format should be a code block
    with at least 500 line-separated prompts
    .
```

An example output from GPT4:

1. A lock of wavy, sunlit blonde hair.
2. A frosted glass sculpture of a swan.
3. A detailed macro shot of a butterfly wing.
4. A swirling vortex of crystal-clear water.
5. A swirling tornado of fire and ash.
6. A soap bubble just before it bursts.
7. A detailed close-up of a human iris.
8. A single water droplet on a lotus leaf.
9. A piece of amber glass reflecting sunlight.
10. A single strand of barbed wire with dew drops.
11. A bonfire with intense, twisting flames.
12. A close-up of intricate lacework.
13. A glowing ember in a dying fire.
14. A close-up of a dragonfly’s wing.
15. A close-up of frost patterns on a window.
16. A bubble reflecting a rainbow of colors.
17. A eagle’s feather with detailed texture.

9.2. Procedurally Generated Prompts

1. anxious man big ears
2. black escape artist man
3. bored physician girl
4. elderly personal care aide boy
5. excited old psychic person
6. firefighter woman closed eyes
7. gay stablehand woman
8. hispanic barista man with black flowing hair
9. lawyer woman diamond earrings
10. man wearing purple skirt
11. necromancer man brown eyes
12. nurse person green eyes
13. person wearing gown
14. sad fairy girl hazel eyes
15. seamstress girl standing
16. software engineer boy big ears
17. teenage gay nurse man
18. waiter man beard waving
19. white boy with curly hair
20. woman with red spiky hair

9.3. Image Captions

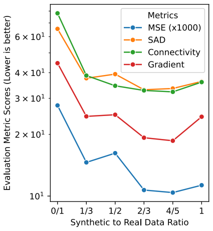
1. Close-up of a new basketball ball
2. Dairy products on a wooden table
3. Deliciously refined tangerines
4. Dog in a hat laborer looking at the camera
5. Dried betel nuts or areca nuts
6. Flying bird from black smooth lines
7. Fresh artichokes close-up on dark background
8. Fresh lemon with lemon essential oil
9. Glasses of tasty Negroni cocktail
10. Green bush or wall of shrubs
11. Heart sticker with the flag of Tajikistan
12. Intertwined white textile fibers
13. Number 14 made of wooden blocks
14. Piggy bank with a vernier caliper
15. Shiba Inu dog in a birthday cap
16. Skyscraper building in 3D render
17. Varnished beige elegant shoes
18. White and brown chicken wings
19. White bread toast with honey
20. Young smiling woman posing in a studio

10. Matting Experiments

We trained a matting model from Dai et al, CVPR 2023 [7] under default settings on various training sets - comprised of images from both the MAGICK dataset and the Deep Image Matting (DIM) dataset. The ratio of datasets used varied; for example, a 1/3 ratio means 1/3 of the training images were from MAGICK, and 2/3 from DIM. Our results are in Fig. 17.

We evaluated the models on the DIM test set using the standard metrics. Our findings showed that a combined dataset approach yielded better results than using either the MAGICK or DIM datasets alone. The optimal performance was achieved with a mixture where 1/5 of the data was from MAGICK and 4/5 from DIM.

We conclude that the MAGICK dataset is indeed useful for image matting, even though it was primarily designed for image generation - resulting in a considerable domain difference between the two datasets.



Ratio	**MSE**	**SAD**	**Conn**	**Grad**
0	.0277	65.49	78.03	44.54
1/3	.0146	37.59	38.76	24.49
1/2	.0162	39.38	34.54	24.95
2/3	.0107	**33.03**	32.78	19.27
4/5	**.0104**	33.51	**32.26**	**18.62**
1	.0113	36.27	36.00	24.39

Figure 17. Matting results using MAGICK.

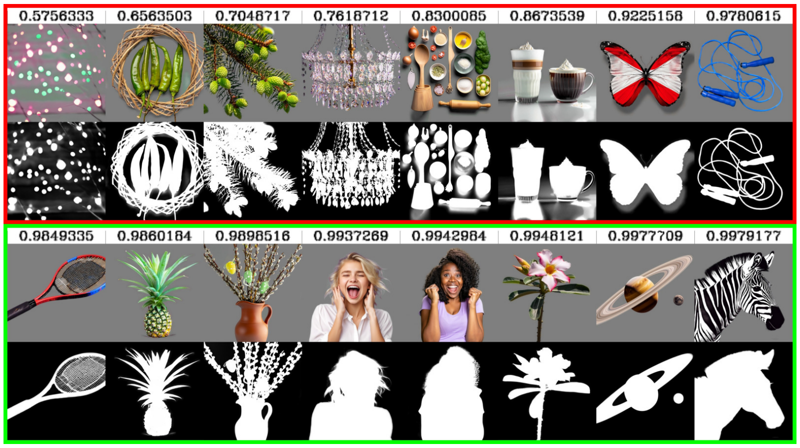


Figure 18. **Automatic selection:** Randomly generated images with similarity scores increasing from left to right. The top 50%, highlighted in green, are kept while the rest are discarded. Samples with high similarity almost always have accurate alpha mattes.

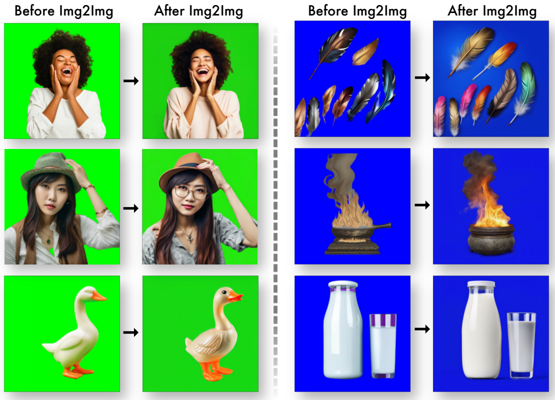


Figure 19. **SDEdit's Effect:** 6 more examples continuing Fig. 7 in the main paper. Note the extra detail given by SDEdit.

11. Qualitative Alpha-to-RGB Results

In this section, we showcase many examples of Alpha-to-RGB generation, as described in Sec. 4.

There are many artistic applications our dataset here, such as applying styles to text.

Here are the figures we've included in this section:

1. Text Stylization, along with comparisons to baselines: See Fig. 23
2. Optical Illusions: See Fig. 22.
3. More variants of the letter S, continuing Fig. 13: See Fig. 21.
4. Other Results: See Fig. 24.

12. Creation Effort

Our dataset is comprised of both automatically selected images and manually selected images. The automated part, forming 110k out of MAGICK's 150k images, involves negligible human effort. This process, as described in the paper, uses 32 A100 GPUs for three weeks, incurring only computational costs. For manual section, comprising 40,000 images, we hired 5 workers who each worked 112 hours at a rate of $0.3 USD per sample. Four of these workers were annotators, and the fifth was in charge of quality control. Human anotation requires only one mouse click per sample. Future work could include training a classifier model to replace the annotators.

13. Limitations

MAGICK's main strength is its size - comprising of 150k images. However, its main limitation is that it is synthetic - inheriting both strengths and weaknesses from current diffusion models. For example, a sample with the caption "stop sign" will have a good alpha matte, but might spell "stop" incorrectly, as SDXL struggles with text.

14. Extended Dataset Preview

In addition to Fig. 2, which showcased 100 samples, this section presents an additional 1300 randomly selected samples from our dataset in figures Fig. 25, Fig. 26, Fig. 27, and Fig. 28.

The 1400 matted images exhibited in this document surpass the size of the previously largest general-purpose matting dataset [35], which contained 726 objects. Furthermore, the 1400 samples illustrated in this document represent less than 1% of the entire MAGICK dataset, encompassing 150,000 samples, each at double the resolution of any figure depicted here.

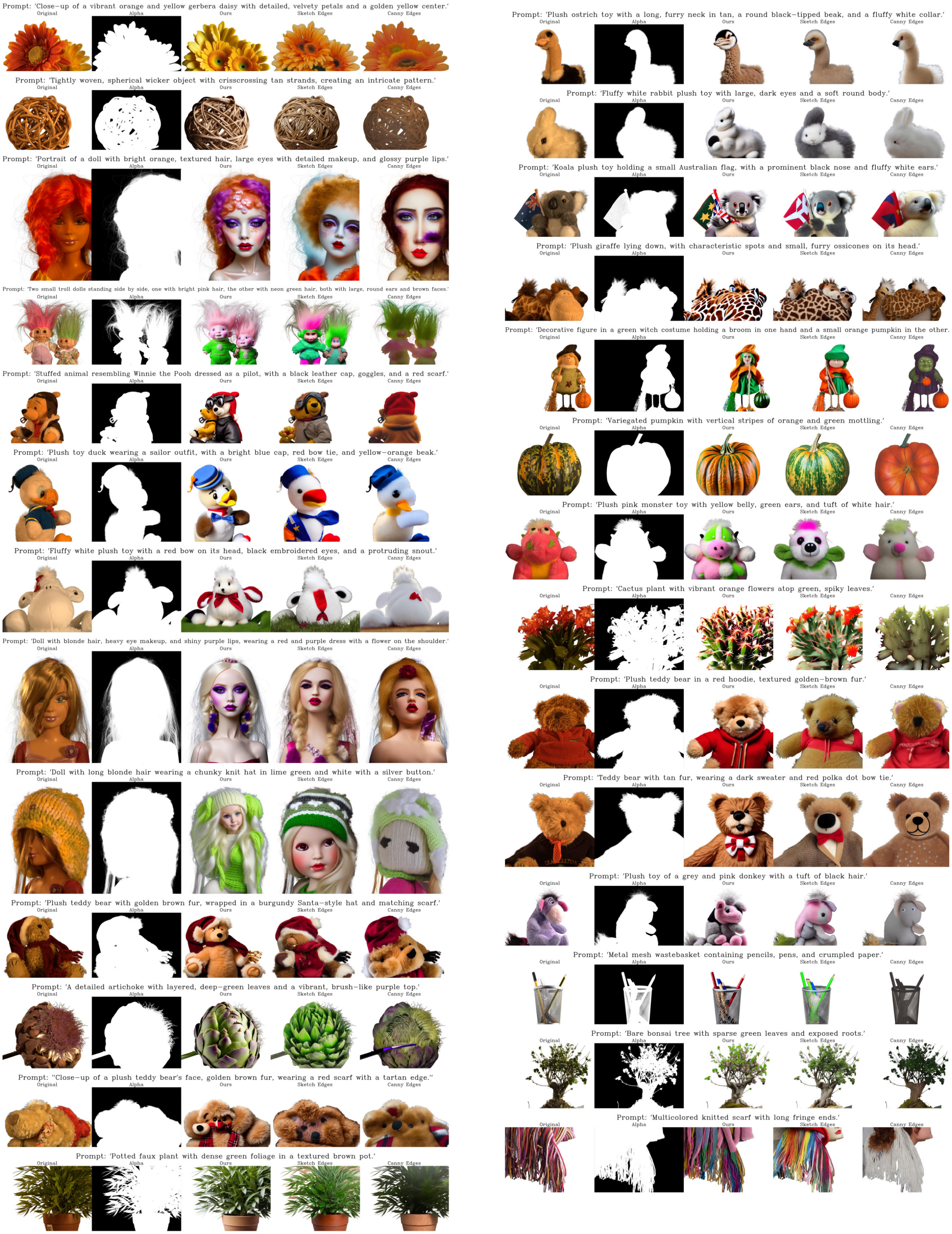

Prompt: 'Close−up of a vibrant orange and yellow gerbera daisy with detailed, velvety petals and a golden yellow center.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Tightly woven, spherical wicker object with crisscrossing tan strands, creating an intricate pattern.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Portrait of a doll with bright orange, textured hair, large eyes with detailed makeup, and glossy purple lips.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Two small troll dolls standing side by side, one with bright pink hair, the other with neon green hair, both with large, round ears and brown faces.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Stuffed animal resembling Winnie the Pooh dressed as a pilot, with a black leather cap, goggles, and a red scarf.'
Original
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush toy duck wearing a sailor outfit, with a bright blue cap, red bow tie, and yellow−orange beak.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Fluffy white plush toy with a red bow on its head, black embroidered eyes, and a protruding snout.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Doll with blonde hair, heavy eye makeup, and shiny purple lips, wearing a red and purple dress with a flower on the shoulder.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Doll with long blonde hair wearing a chunky knit hat in lime green and white with a silver button.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush teddy bear with golden brown fur, wrapped in a burgundy Santa−style hat and matching scarf.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'A detailed artichoke with layered, deep−green leaves and a vibrant, brush−like purple top.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: "Close−up of a plush teddy bear's face, golden brown fur, wearing a red scarf with a tartan edge."
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Potted faux plant with dense green foliage in a textured brown pot.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush ostrich toy with a long, furry neck in tan, a round black−tipped beak, and a fluffy white collar.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Fluffy white rabbit plush toy with large, dark eyes and a soft round body.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Koala plush toy holding a small Australian flag, with a prominent black nose and fluffy white ears.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush giraffe lying down, with characteristic spots and small, furry ossicones on its head.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Decorative figure in a green witch costume holding a broom in one hand and a small orange pumpkin in the other.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Variegated pumpkin with vertical stripes of orange and green mottling.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush pink monster toy with yellow belly, green ears, and tuft of white hair.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Cactus plant with vibrant orange flowers atop green, spiky leaves.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush teddy bear in a red hoodie, textured golden−brown fur.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Teddy bear with tan fur, wearing a dark sweater and red polka dot bow tie.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Plush toy of a grey and pink donkey with a tuft of black hair.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Metal mesh wastebasket containing pencils, pens, and crumpled paper.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Bare bonsai tree with sparse green leaves and exposed roots.'
Original
Alpha
Ours
Sketch Edges
Canny Edges
Prompt: 'Multicolored knitted scarf with long fringe ends.'
Original
Alpha
Ours
Sketch Edges
Canny Edges

Figure 20. Continuing from Fig. 12, here we present all of the images used in the user study. Our algorithm was compared against both baselines.

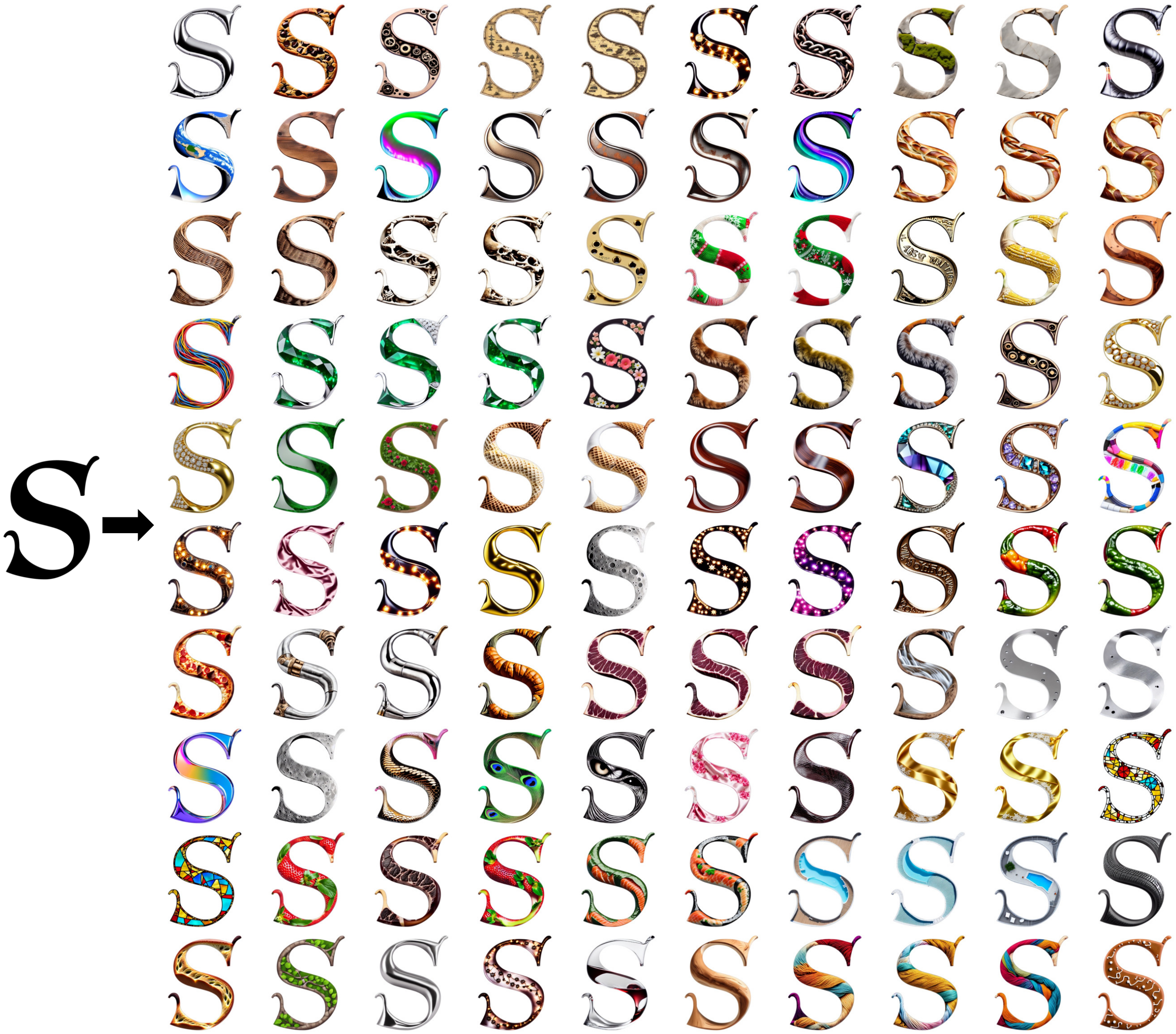


Figure 21. **Text Stylization**: *This image is very high resolution - please zoom in!* Continuing Fig. 13, we take the alpha mask of the letter S (inverted here for visibility), and apply our Alpha-to-RGB algorithm from Sec. 4 to it using 100 different prompts.

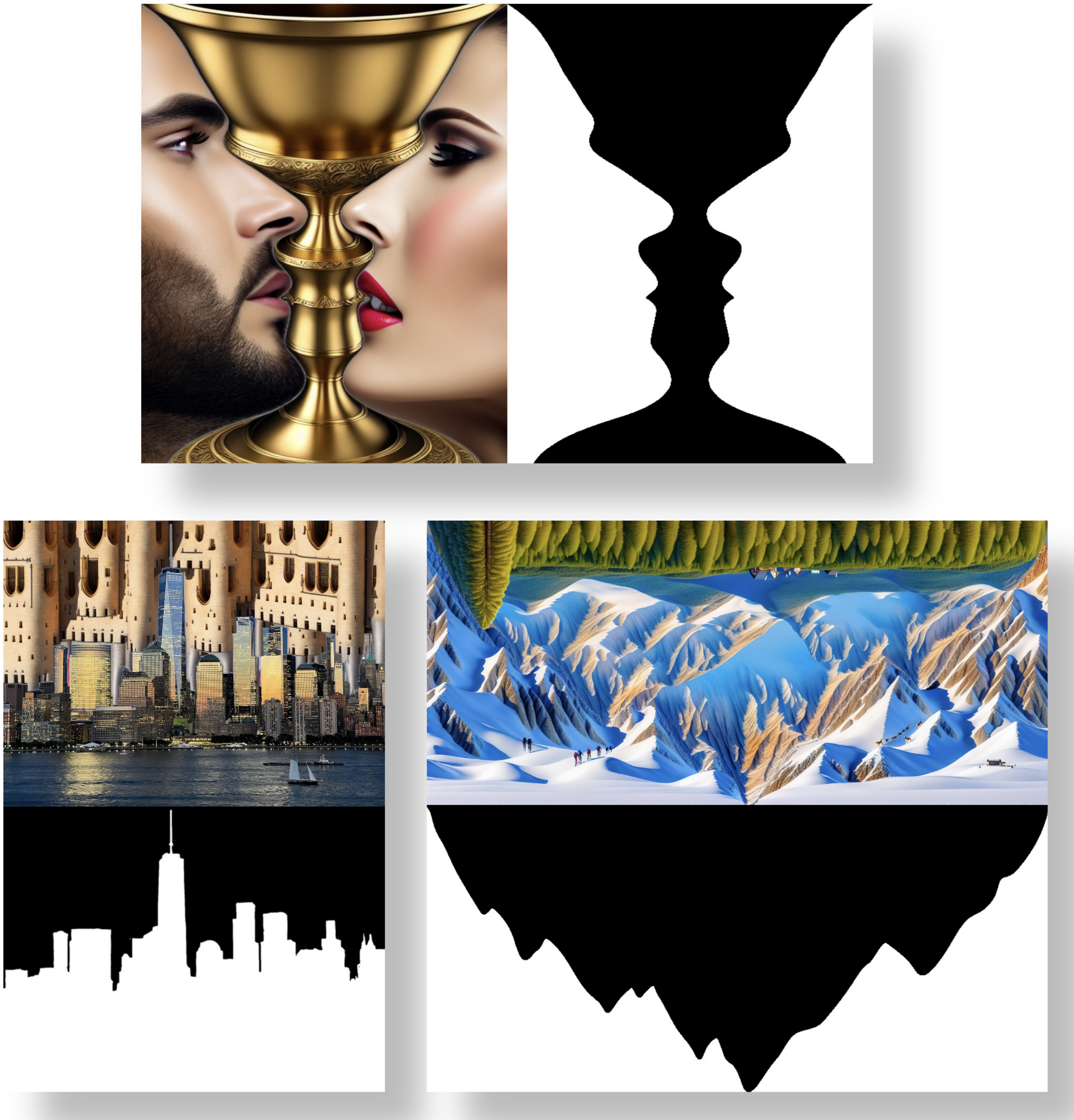
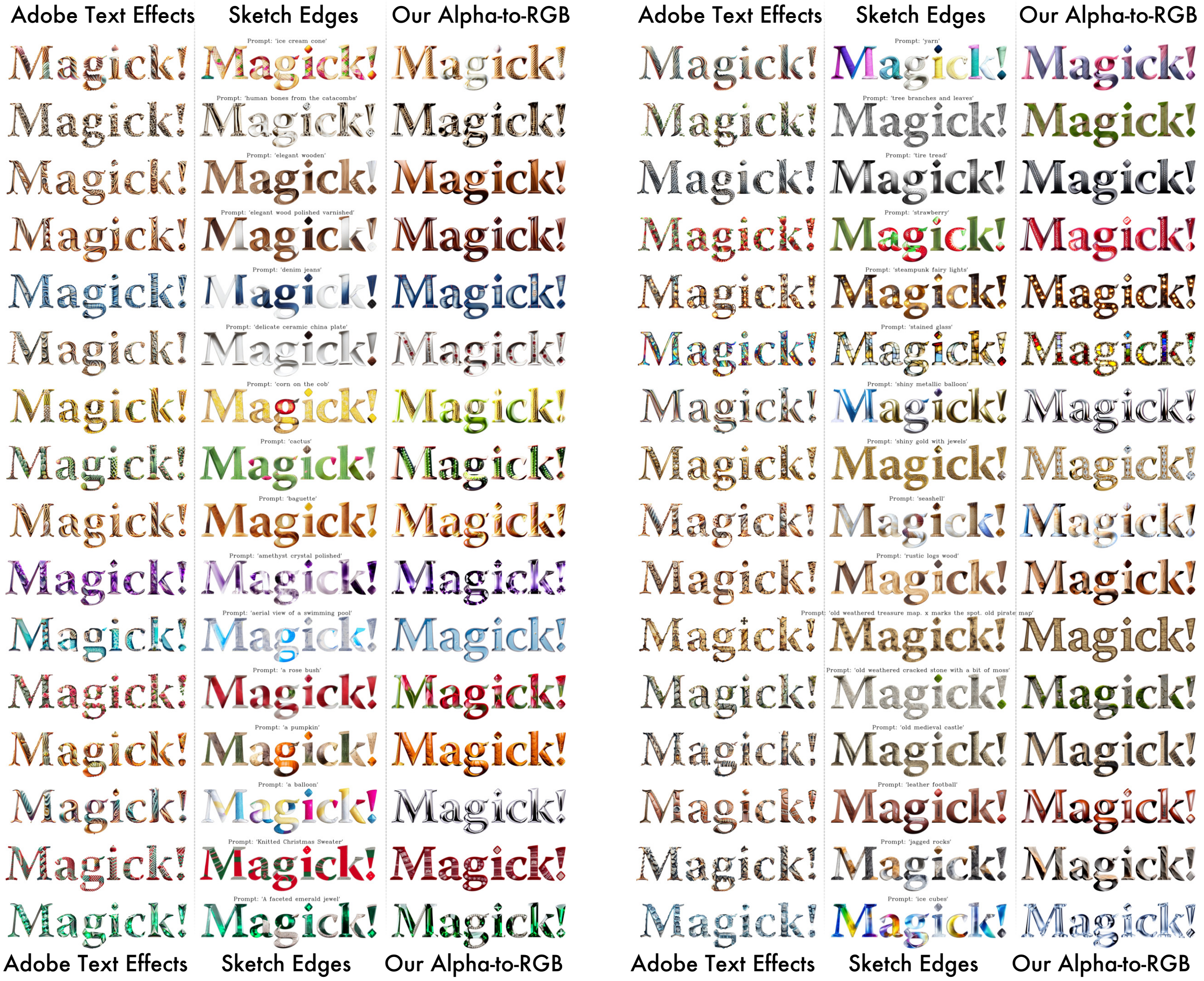


Figure 22. **Optical Illusions**: Our Alpha-to-RGB algorithm from Sec. 4 can be used to generate striking optical illusions. In each image, we use two prompts: one for each region of the alpha mask. On the top image, we fill in the classic goblet illusion: we use the prompts "man and woman staring at each other" along with "a brass goblet". On the bottom right image, we use the prompts "a mountain range with snow-capped mountains" and "a mountain range with snow-capped mountains behind a dense green forest". The log cabins and skiers were added after the fact for decoration. And on the left, a photograph of new york city was cut out, and the remaining mask was given the prompt "a medieval castle" and flipped upside-down and composited back onto the image of the city. Please view them upside-down!


Adobe Text Effects
Sketch Edges
Our Alpha-to-RGB
Adobe Text Effects
Sketch Edges
Our Alpha-to-RGB
Prompt: 'ice cream cone'
Prompt: 'human bones from the catacombs'
Prompt: 'elegant wooden'
Prompt: 'elegant wood polished varnished'
Prompt: 'denim jeans'
Prompt: 'delicate ceramic china plate'
Prompt: 'corn on the cob'
Prompt: 'cactus'
Prompt: 'baguette'
Prompt: 'amethyst crystal polished'
Prompt: 'aerial view of a swimming pool'
Prompt: 'a rose bush'
Prompt: 'a pumpkin'
Prompt: 'a balloon'
Prompt: 'Knitted Christmas Sweater'
Prompt: 'A faceted emerald jewel'
Prompt: 'yarn'
Prompt: 'tree branches and leaves'
Prompt: 'tire tread'
Prompt: 'strawberry'
Prompt: 'steampunk fairy lights'
Prompt: 'stained glass'
Prompt: 'shiny metallic balloon'
Prompt: 'shiny gold with jewels'
Prompt: 'seashell'
Prompt: 'rustic logs wood'
Prompt: 'old weathered treasure map. x marks the spot. old pirate map'
Prompt: 'old weathered cracked stone with a bit of moss'
Prompt: 'old medieval castle'
Prompt: 'leather football'
Prompt: 'jagged rocks'
Prompt: 'ice cubes'
Magick!
Adobe Text Effects
Sketch Edges
Our Alpha-to-RGB
Adobe Text Effects
Sketch Edges
Our Alpha-to-RGB

Adobe Text Effects Settings:

Alpha Mask Used:

Figure 23. **Text Stylization:** *This image is very high resolution - please zoom in!* We apply our Alpha-to-RGB algorithm from Sec. 4 to the text "Magick!" in the font "Warnock", using many different styles. We compare it to two baselines: Adobe Text Effects and our Sketch Edges baseline from Sec. 4. Note how the results from Adobe Text Effects don't always conform to the text boundary properly, despite the settings given to it (also depicted in this figure - its boundary mode is set to "tight"). We chose to include the Sketch Edges baseline instead of the Canny Edges baseline because in our user study Fig. 12 it was the stronger-preferred of the two baselines.

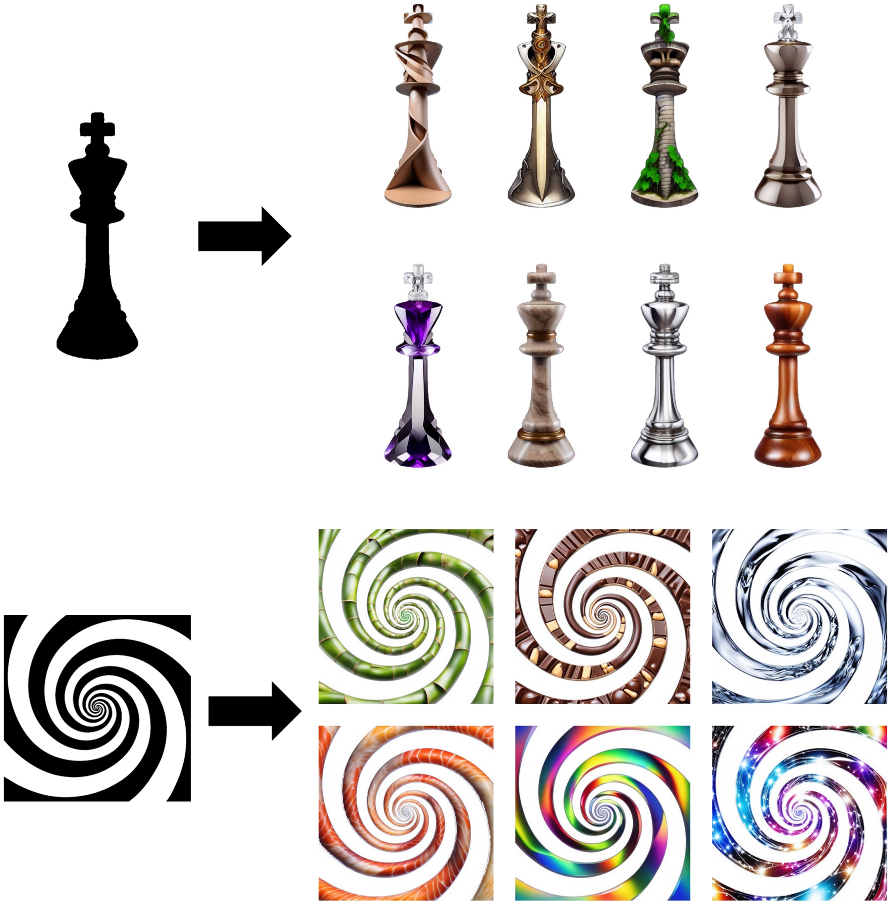


Figure 24. Some more results of our Alpha-to-RGB algorithm 4. The alpha masks are inverted for visibility.

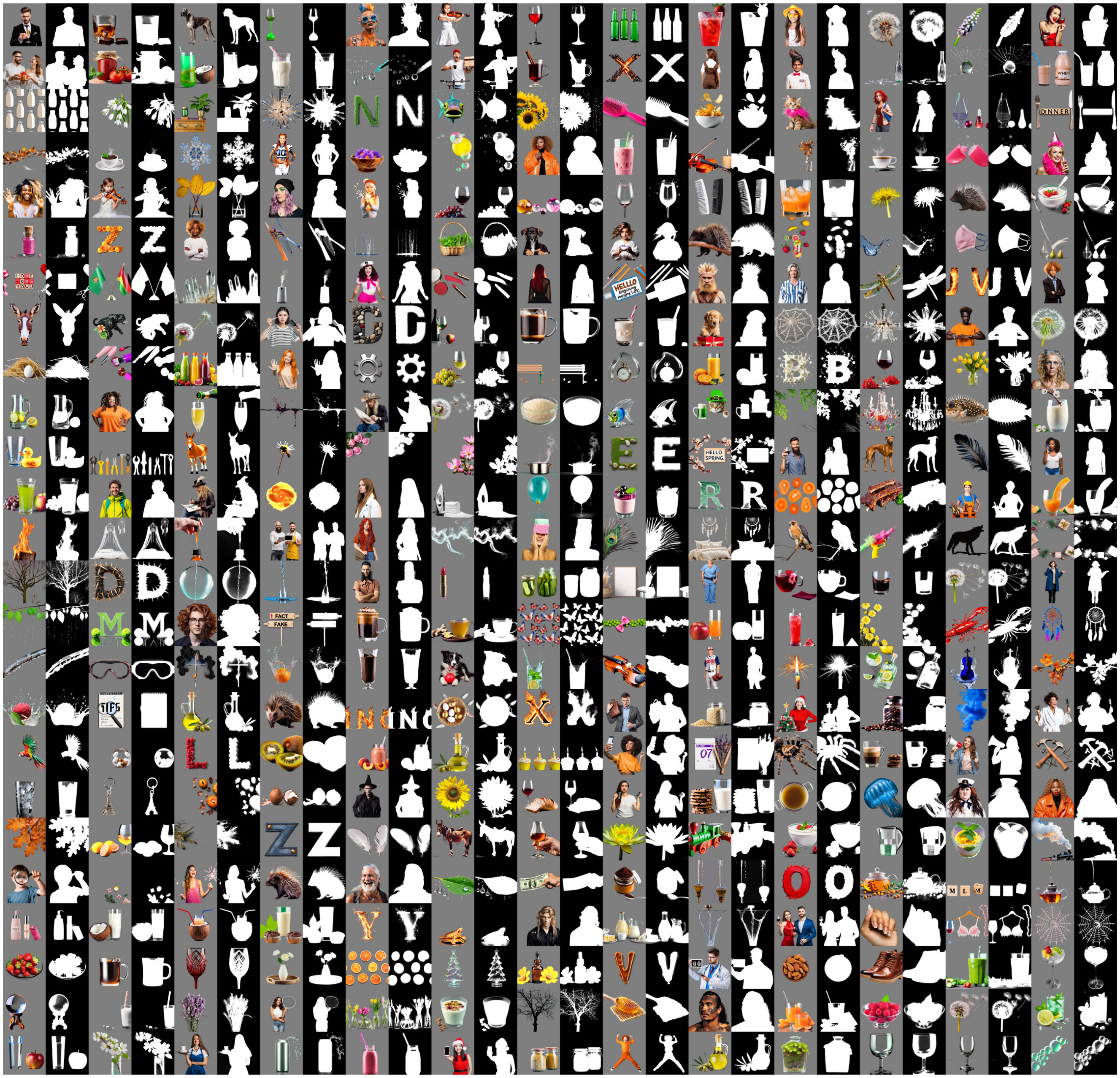


Figure 25. **Dataset Samples Part 1/4**: **Text Stylization:** *This image is very high resolution - please zoom in!* This figure displays 325 random samples from our dataset, along with their alpha masks. Each sample also has a caption, not shown here.

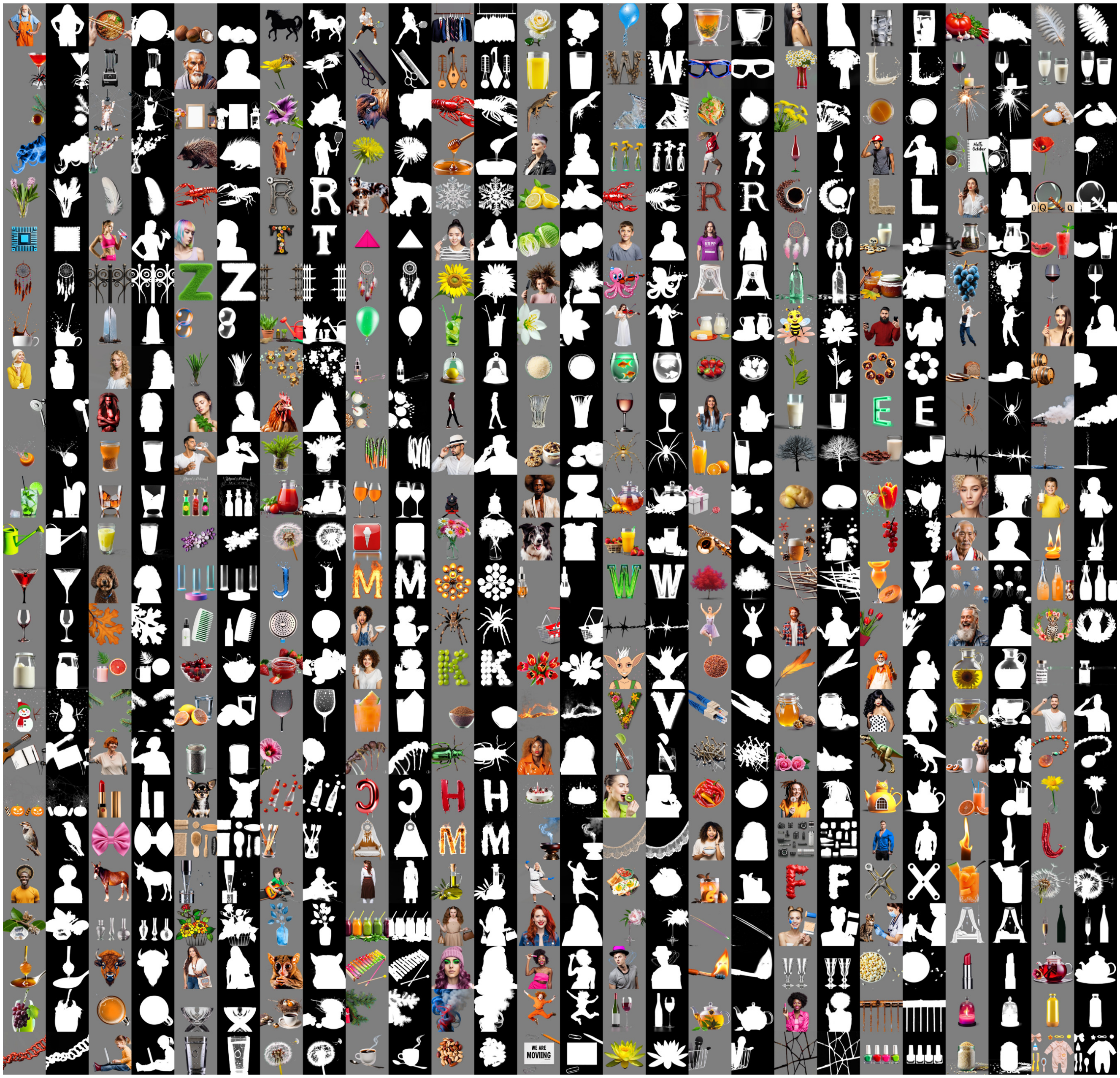


Figure 26. **Dataset Samples Part 2/4**: **Text Stylization:** *This image is very high resolution - please zoom in!* This figure displays 325 random samples from our dataset, along with their alpha masks. Each sample also has a caption, not shown here.

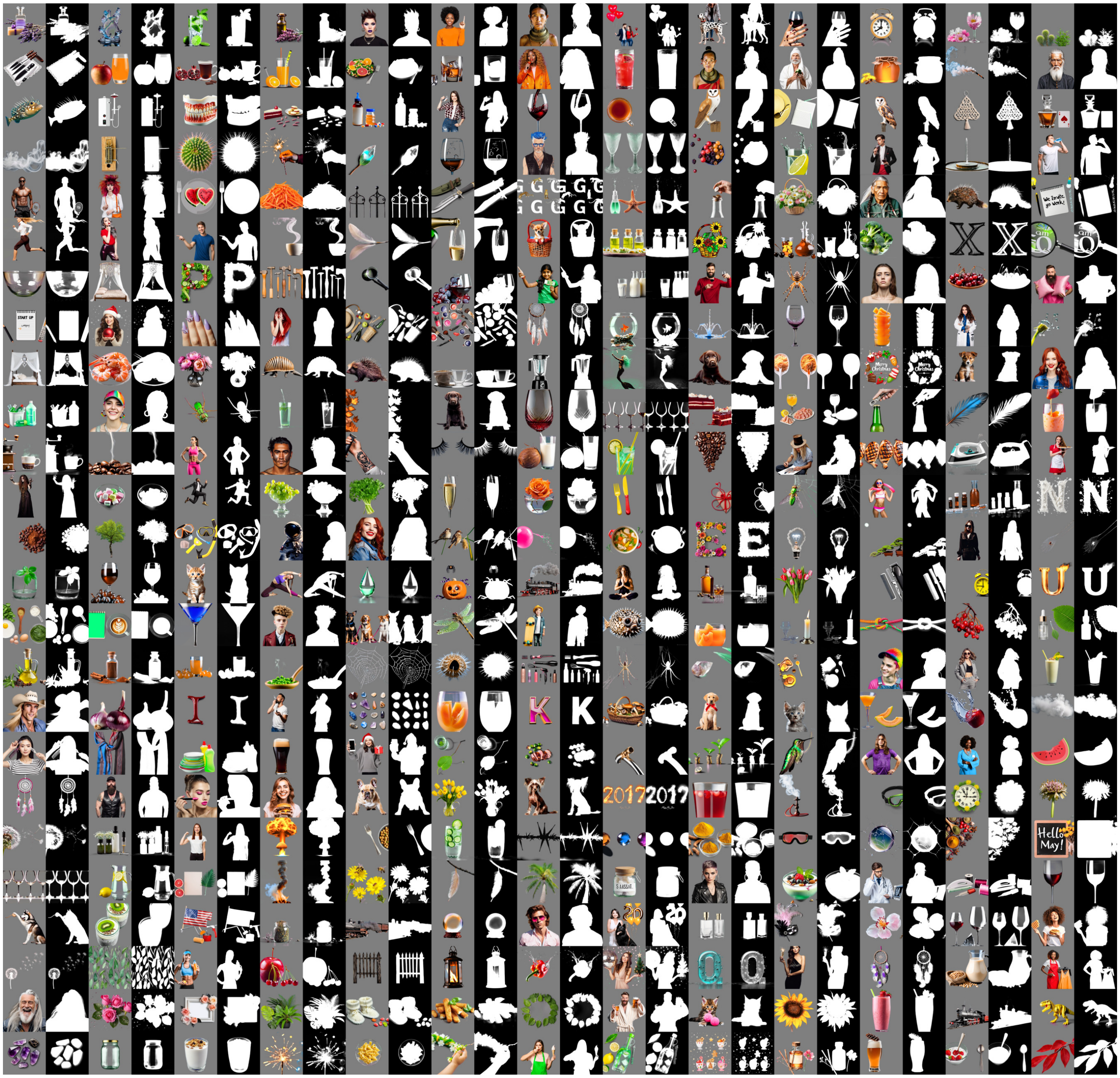


Figure 27. **Dataset Samples Part 3/4**: **Text Stylization:** *This image is very high resolution - please zoom in!* This figure displays 325 random samples from our dataset, along with their alpha masks. Each sample also has a caption, not shown here.

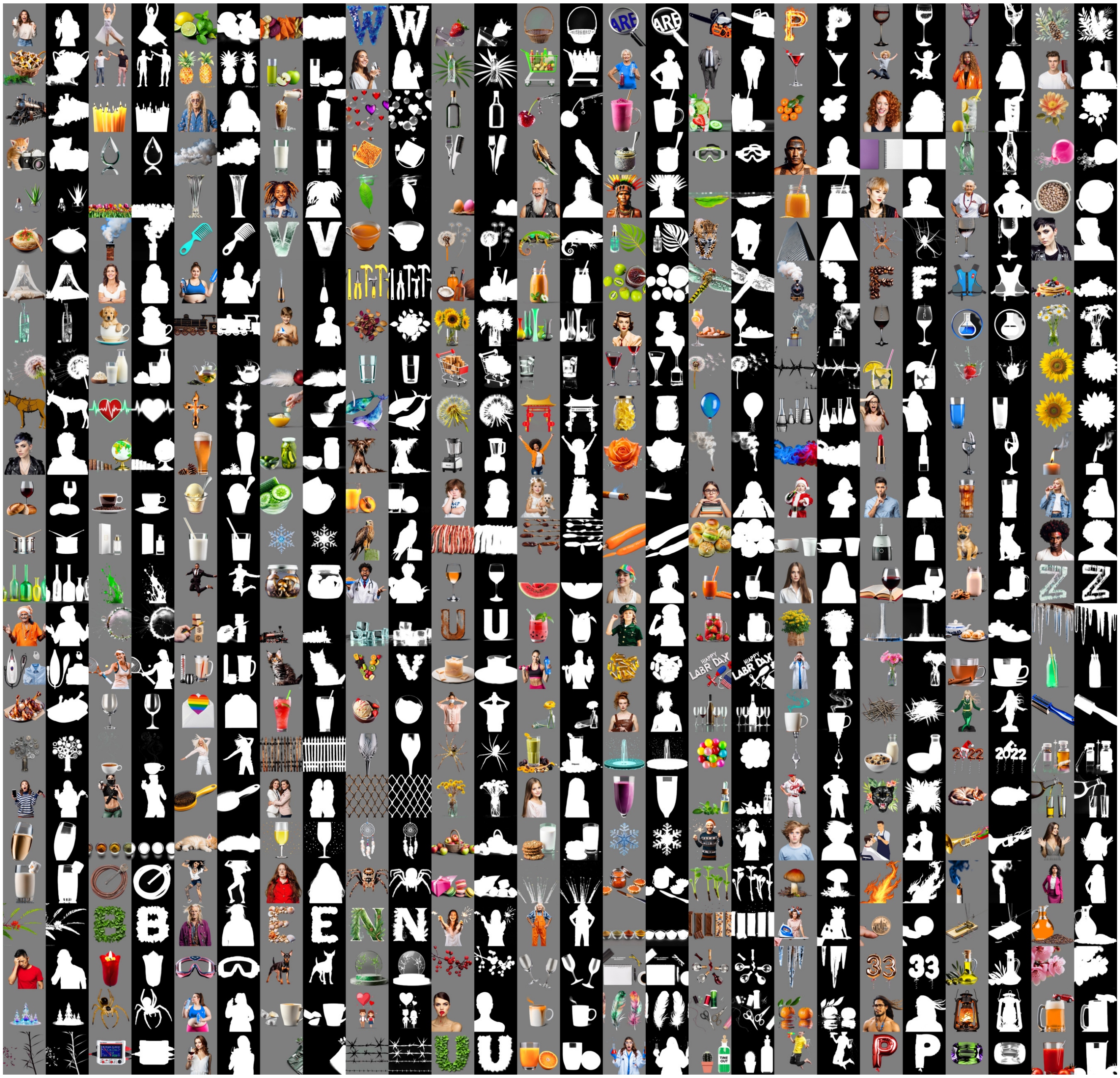


Figure 28. **Dataset Samples Part 4/4**: **Text Stylization:** *This image is very high resolution - please zoom in!* This figure displays 325 random samples from our dataset, along with their alpha masks. Each sample also has a caption, not shown here.

References

[1] Deepfloyd. https://github.com/deep-floyd/if. 2, 4

[2] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go mad, 2023. 2

[3] Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S. Ryoo. Peekaboo: Text to image diffusion models are zero-shot segmentors, 2023. 2

[4] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, 2019. 2

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 4

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2

[7] Yutong Dai, Brian Price, He Zhang, and Chunhua Shen. Boosting robustness of image matting with context assembling and strong data augmentation. In *CVPR*, 2022. 2

[8] Mikhail Erofeev, Yury Gitman, Dmitriy Vatolin, Alexey Fedorov, and Jue Wang. Perceptually motivated benchmark for video matting. In *BMVC*, 2015. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[10] Hugging Face. Image-to-image - stable diffusion. https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/img2img, 2023. Accessed: 2023-11-16. 5

[11] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1, 2

[13] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, , Antonio Torralba Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2

[14] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 2

[15] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 2

[16] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 2

[17] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, 2021. 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Peronaand Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *ECCV*, 2014. 1, 2

[19] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019. 2

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2, 4, 5

[21] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation, 2023. 2

[22] 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*, 2023. 1, 2, 4

[23] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *ICCV*, 2023. 1, 2

[24] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 2

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[26] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *CVPR*, 2009. 1, 2, 8

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[28] 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. *IJCV*, 2015. 2

[29] 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. *NeurIPS*, 2022. 1

[30] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 2

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 4

[32] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *BMVC*, 2016. 2

[33] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget, 2023. 2

[34] Alvy Ray Smith and James F Blinn. Blue screen matting. In *Siggraph*, 1996. 2

[35] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *CVPR*, 2021. 1, 2, 4

[36] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *CVPR*, 2019. 2

[37] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1): 23–34, 2004. 6

[38] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *ICCV*, 2021. 2

[39] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, 2003. 6

[40] Steve Wright. *Digital compositing for film and video*. Taylor Francis, 2010. 6

[41] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In *NeurIPS*, 2023. 2

[42] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 2

[43] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation, 2023. 2

[44] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017. 1, 2

[45] Hanrong Ye, Jason Kuen, Qing Liu, Zhe Lin, Brian Price, and Dan Xu. Seggen: Supercharging segmentation models with text2mask and mask2img synthesis, 2023. 2, 7

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 4

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 7

[48] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. Attention-guided temporally coherent video object matting. In *ACM MM*, 2021. 2

[49] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *CVPR*, 2022. 2

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *CVPR*, 2017. 1, 2, 7