Conditional Balance:
Improving Multi-Conditioning Trade-Offs in Image Generation

Supplementary Material

8. Appendix A. — Content-Style Tradeoff

In Sec. 3 we study the relationship between content and style and experiment with various conditioning settings. In this section we expand on our evaluation set and on additional experiments regarding B-LoRA [12].

8.1. Evaluation Set

Our initial evaluation set contains 10 artistic styles: *Beatrix Potter, Claude Monet, Egon Schiele, John Singer Sargent, Pablo Picasso, Studio-Ghibli, Utagawa Kuniyoshi, Vincent Van-Gogh, Winslow Homer,* and *Xu Beihong*. For each style, we use five "Easy" and five "Complex" prompts for evaluation: *"A bull moose," "A grizzly bear," "A dragon flying in the sky," "A portrait of a woman," "A tabby cat sitting,"* (Easy) and *"A girl wearing a black and white striped shirt riding a bull moose in the Alaska wilderness," "A family of panda bears wearing kimonos and sharing some tea," "Two dragons, a green one and a red one, flying in a purple sky," "A man wearing sunglasses and a woman watching the sunset from a mountain top," "A ginger tabby cat riding a bicycle in Amsterdam next to a river"* (Complex).

In Sec. 6 we extend our evaluation set with 22 additional styles and 10 additional prompts. The additional styles are: *Pixar, Pixel Art, Edvard Munch, Franz Marc, John James Audubon, Oswaldo Guayasamin, Henri Matisse, Wassily Kandinsky, Ilya Repin, Gustav Klimt, Voxel Art, Vector Art, Anime, Henri De Toulouse-Lautrec, Yoshitaka Amano, Cyberpunk, Concept Art, Low Poly, Gustav Courbet, Paul Cézanne, Jean Metzinger* and *Georges Seurat*. The additional prompt are: *"An old TV set," "A colorful fishbowl," "A house in a village," "A bartender leaning on his bar," "A brown horse galloping"* (Easy) and *"A robot wearing a fedora holding a flower," "A humpback whale floating in the sky carried by large colorful balloons," "A fantasy castle with blue pointy rooftops located on a hill in a green valley," "An orc and a blond wood-elf sitting in a tavern drinking beer as friends," "Gandalf the Gray riding a horse while casting a spell with his wooden staff"* (Complex).

To generate the evaluation set we use 4 randomly chosen seeds: 10, 20, 9787, and 140592. For text-only generation we use all four seeds, for Canny conditioning we use 10 and 9787, and for Depth conditioning we use 20 and 140592.

8.2. B-LoRA Experiments

In Sec. 3 we investigate the content-style tradeoff by using StyleAligned [17]. We expand this study for B-LoRA using the same evaluation set. Unlike StyleAligned, B-LoRA requires training residual weights prior to inference, which prevents applying stylization over a random set of layers for each evaluated image. Instead, we show the tradeoff between content and style using our balancing strategy and compare it to B-LoRA. Following Sec. 3 we use Dino [5] and Clip [27] embeddings to evaluate style and content, respectively, over various layer combination choices for both text conditioning and structure conditioning experiments.

We report both Qualitative and Quantitative evaluations in Fig. 10. As illustrated, our strategy balances content and style for mutual conditioning. In the case of 'Text Conditioning' (left) we can see that choosing style sensitive layers by our layer ranking yields a dramatic improvement in style over B-LoRA without sacrificing content, even when basing the stylization on only five self-attention layers. In this case we observe that choosing 20 layers yields a good balance between content and style. In the case of 'Structure Conditioning' (right) using a structure control map yields more stability in content even for a high number of stylization layers. For this reason, we find that choosing 40 layers yields the optimal balance between content and style.

In both cases we observe that using excessive style may lead to issues caused by content drift from the style image. When using a structure map (image D. in Fig. 10) the impact can be marginal but when using a text condition alone (image A. in Fig. 10) we can sometime lose the content of the image overall.

9. Appendix B. — Analysis

9.1. Painting Collections

Our style and content analysis is conducted over five collections each. For our style analysis we use five different objects and constrain their structure with a Canny map: *Car, House, Rabbit, Bottle,* and *Chair.* We generate 10 image clusters by various artistic styles: *Vincent Van-Gogh, Claude Monet, Georges Seurat, Paul Signac, Edvard Munch, Winslow Homer, John Singer Sargent, Edward Hopper, Paul Cézanne,* and *Berthe Morisot*. We choose these styles as they show variance in color and texture patterns but all have relatively realistic geometric style. The entire collections for *Car* and *Rabbit* are presented in Fig. 33.

The geometric sensitivity analysis was focused on limiting content conditionals from layers sensitive to geometric style. We choose five different objects: *Cat, Wolf, Cow, Shark,* and *Horse.* We choose to concentrate on animals as

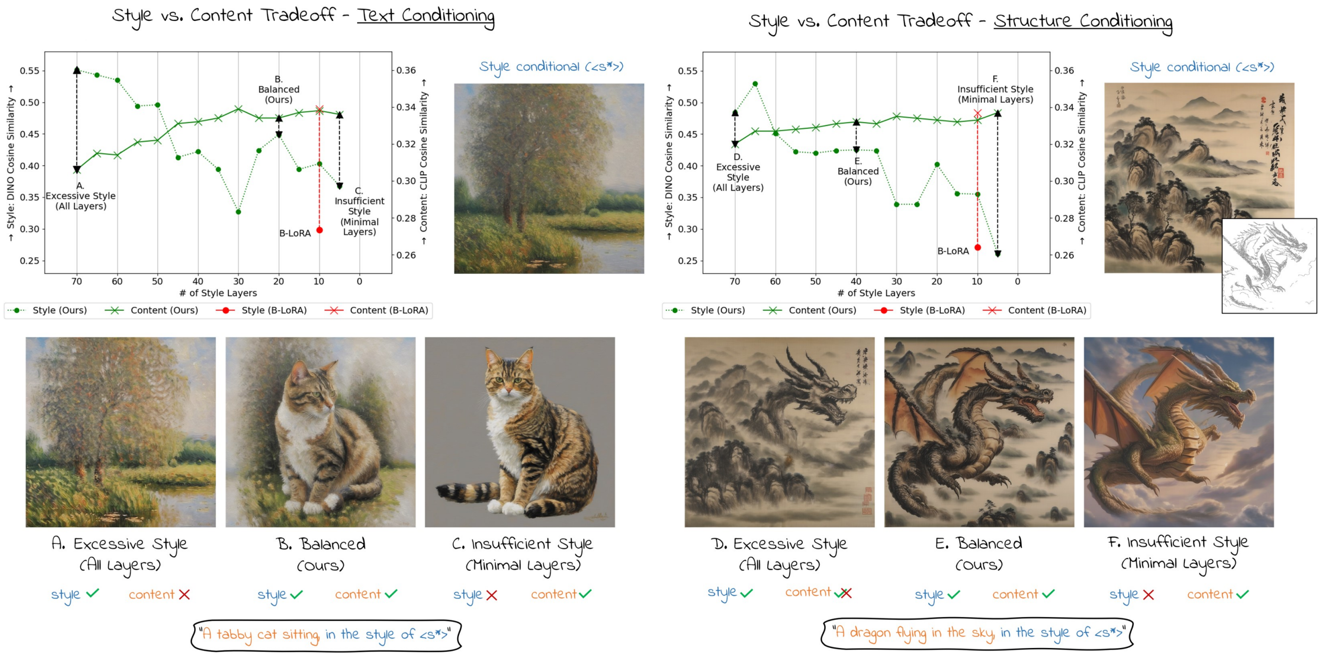


Figure 10. **Content-Style Tradeoff - B-LoRA.** *Investigating the content-style tradeoff for B-LoRA vs. balancing using various number of stylization layers using our balancing strategy. As can be seen, using our strategy leads to balanced results for both 'Text' and 'Structure' conditioning and improves the overall generated image quality over imbalanced B-LoRA.*

they tend to have more fluid interpretations in art paintings which is key for varying geometric style through the collections. We do not use a conditioning map to constrain structure as geometric style freedom is dependent on structure freedom. We generate 10 image clusters using the following styles: *Jean-Michel Basquiat, Egon Schiele, Franz Marc, Vincent Van-Gogh, Ernst Ludwig Kirchner, Henri Matisse, Jean Metzinger, Edvard Munch, Pablo Picasso,* and *Utagawa Kuniyoshi.* The entire collections of *Cat* and *Wolf* are presented in Fig. 34 and Fig. 35, respectively, in both their color version and black and white version which was used in the analysis.

9.2. Layer Rankings

Using our analysis method results with a ranking for each layer at each timestep. To better understand the ranking choices we show the mean and standard deviation of the layer rank over timesteps (Fig. 11). As can be seen, both style and geometry show a high correlation with the Up layers of the denoising UNet, while style seems to show a significant correlation also with Down layers.

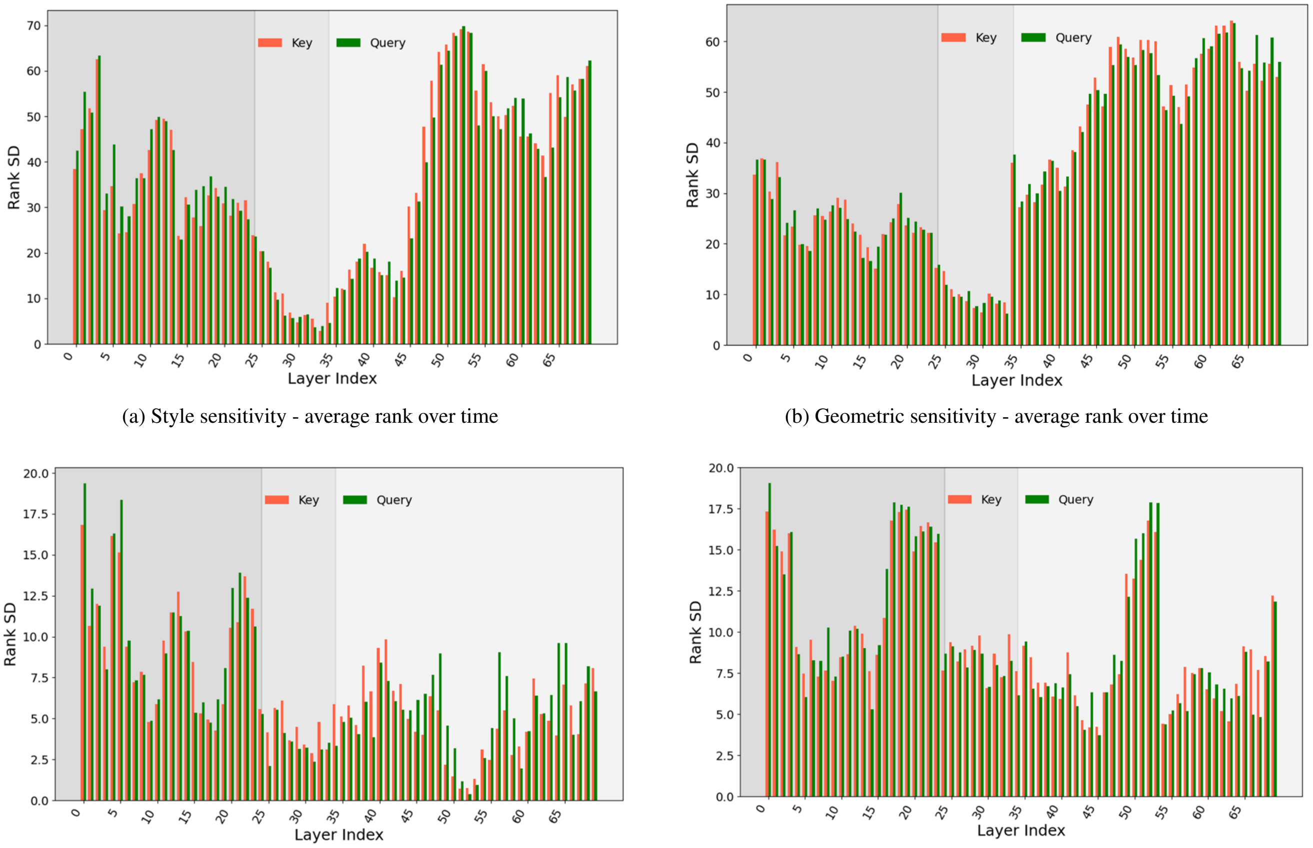
We present an example of a choice of 30 layers of Key layers for both style and geometry in Fig. 12. As can be seen the majority of layers show consistency over time while some layers change on various timesteps.

9.3. Style Layers Ablation Study - λ_S

In Sec. 3 of the main manuscript, we examine the trade-off between style and content in conditional image generation and provide an initial evaluation of the effectiveness of applying stylization to a subset of self-attention layers, as determined by our analysis method. In this subsection, we further explore our method's ability to identify style sensitivities by assessing the performance of layers marked as **not** style-sensitive. We compare their impact to both random selection and our previous results from Sec. 3. To achieve this, we generate the evaluation set described in Sec. 3 using different subset sizes of style layers. However, instead of applying stylization to the k-most style-sensitive layers, we now apply it to the k-most **insensitive** layers.

Our experimental findings are illustrated in Fig. 13. Similar to the approach in Sec. 3 of the main manuscript, we provide quantitative evaluations for both style similarity (top left) and content similarity (top right), alongside a qualitative example for visual demonstration.

As shown in the top left plot, applying stylization to layers identified as not style-sensitive (blue graph) results in a significant reduction in the style similarity of the generated images compared to stylization using the style-sensitive layers (green graph). Furthermore, applying stylization to these layers leads to lower style similarity than random layer selection, reinforcing the effectiveness of our method in correctly identifying both style-sensitive and



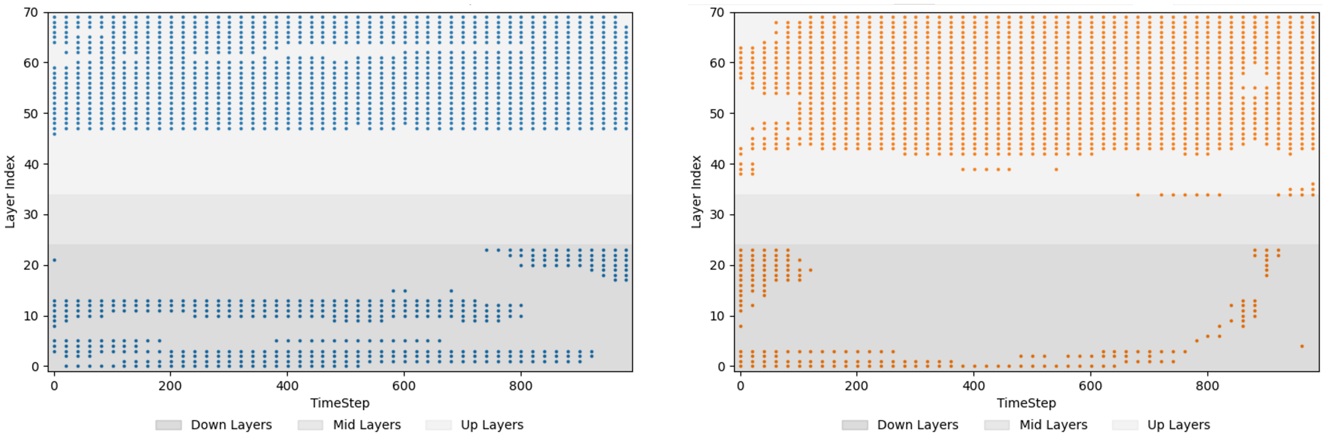
(a) Style sensitivity - average rank over time

(b) Geometric sensitivity - average rank over time

(c) Style sensitivity - standard deviation over time

(d) Geometric sensitivity - standard deviation over time

Figure 11. **Average Layer Rank.** We show the average layer grade over all time steps for style and content sensitivity analysis. The top row depicts the first instance of each plot, and the bottom row duplicates them. As can be seen, various Up layers are important for both general style and geometric style. While geometric style seems to be more reliant on Up layers, some general style aspects seem to rely on Down layers. (Down, Middle and Up layers are divided by gray colored areas in the plot from left to right, respectively.)



(a) Style sensitive layer choice

(b) Geometry sensitive layer choice

Figure 12. **Layer Decision Example.** *We show an example of the layer choice for* $\lambda_S = 0.43$ *using 30 Key layers for style (left) and geometry (right). As can be observed the majority of layers show consistency over time while a some layers change for different timesteps.*

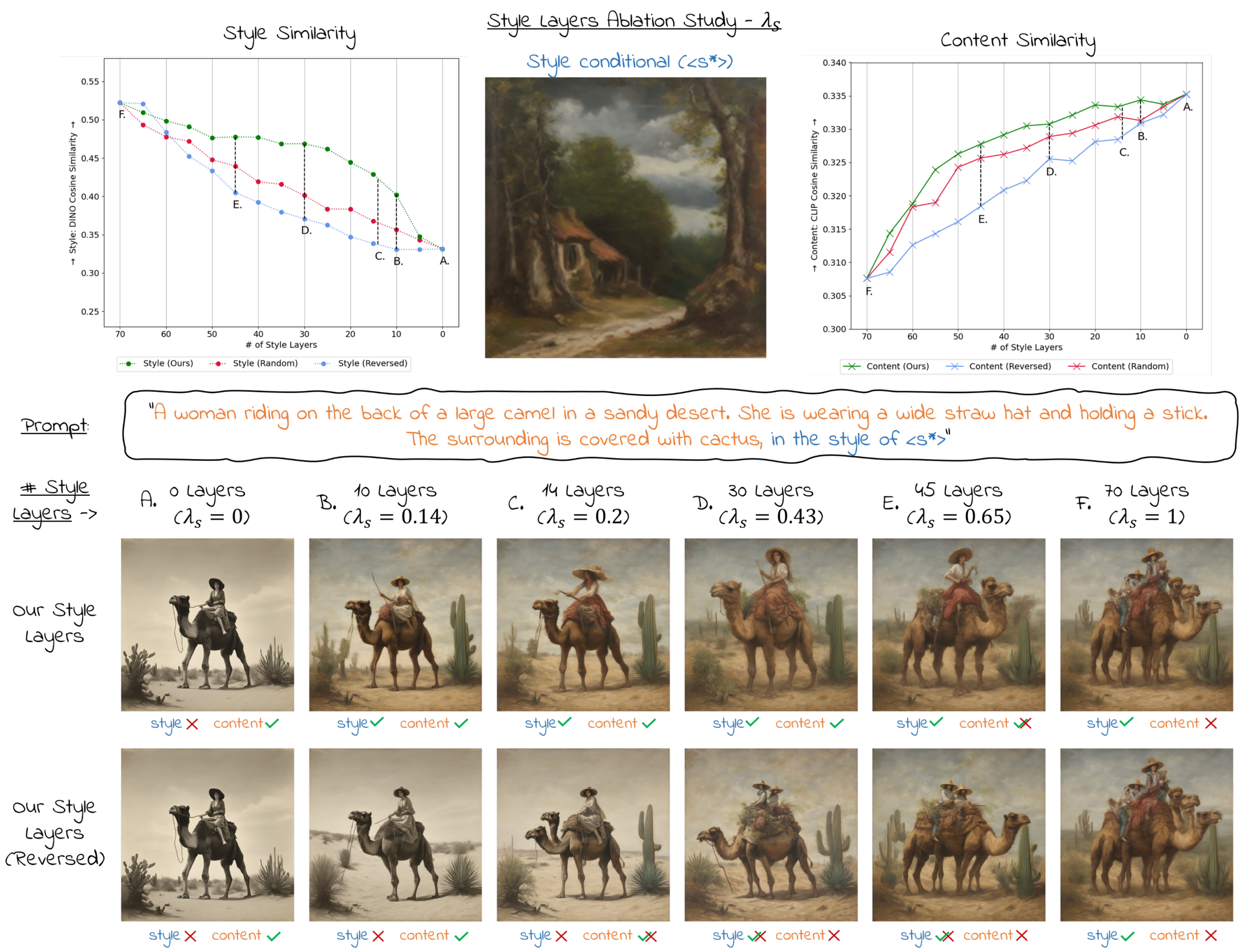


Figure 13. **Style Layers Ablation.** *Ablation study for the ability of our sensitivity analysis method. We investigate and compare the method's ability to identify both style-sensitive and style-insensitive layers by using both layer rankings for stylization, where using insensitive layers is marked in blue and using sensitive layers is marked in green. We show quantitative results (top) and compare both scenarios to a random layers selection (red) and show a visual example for demonstrative purposes (bottom).*

style-insensitive layers. Notably, these insensitive layers appear to be entirely unrelated to style, as stylizing them negatively impacts the overall stylization quality. Additionally, using the insensitive layers causes a slower increase in style similarity, which only starts to improve around (E.), when the style-sensitive layers take effect. Observing the content similarity plot (top-right), we can confirm our hypothesis that injecting style information into style-insensitive layers (blue) is not only ineffective for style similarity but also degrades content similarity. This degradation results in lower content similarity compared to both style-sensitive layers (green) and randomly selected layers (red.)

The qualitative impact of these findings is visually demonstrated in the image sequence at the bottom of Fig. 13. This sequence compares the interpolation effect of applying λ_S on style-sensitive layers (top row) versus style-insensitive layers (bottom row). Each column ($A.-F.$), corresponding to points in the quantitative graphs, represents an increasing λ_S value. Columns (A) and (F) show generated images for (A) no image style conditioning and (F) full layer style conditioning, both presenting a suboptimal result. The following observations can be made: (B) When using only 10 style-sensitive layers, the generated image already exhibits strong style representation while maintaining content integrity, whereas using style-insensitive layers results in no noticeable stylization effect. (C) Utilizing 14 sensitive layers maintains the previous quality, whereas using insensitive layers introduces content artifacts without achieving style alignment, likely due to injecting style information into content-related layers. (D) Applying 30 sen-

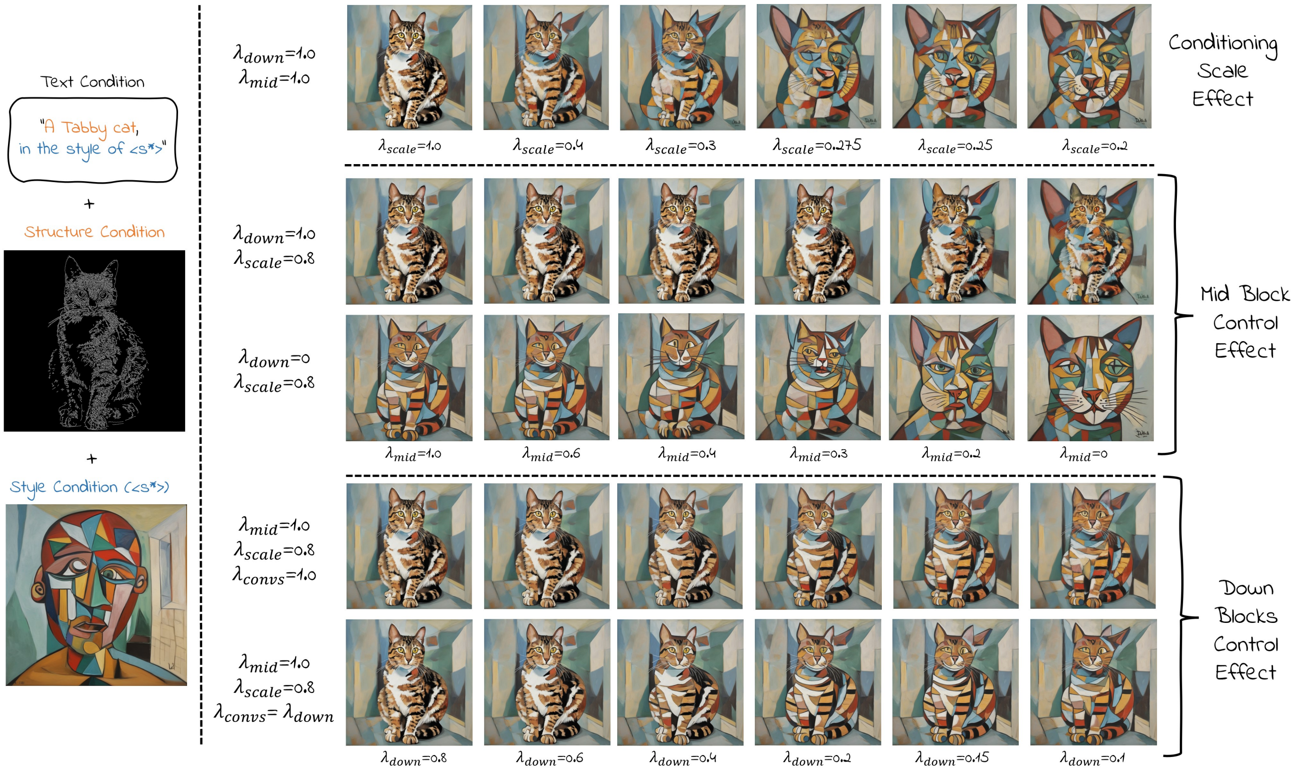


Figure 14. **Geometric Style Ablation.** *Ablation study for our geometric scaling factor. We demonstrate the effect of* $\lambda_{scale}, \lambda_{mid}, \lambda_{down}$ *on the geometric style of the generated image. As can be seen, reducing results in a gradual decrease of fine details control and enables the model to gradually increase its geometric freedom which peaks around* λ_{down} *value of 0.15. However, reducing control by lowering* λ_{scale} *(top row) or* λ_{mid} *(middle rows) results in abrupt loss of of general mask details around the value of 0.3, which leads to neglecting the content condition overall.*

sitive layers further improves both style and content alignment, while using insensitive layers results in some style improvements but also introduces content artifacts. (E) With 45 layers, stylization using our identified sensitive layers begins to introduce content artifacts, while using insensitive layers finally achieves reasonable style alignment as style and content-related layers start to blend in both cases.

9.4. Geometric Style Ablation Study - λ_T

Following our analysis in Sec. 5, we conduct an ablation study to investigate the impact of the residual outputs of ControlNet on the generated images. ControlNet fine-tunes a copy of the denoising UNet encoder and extracts its outputs from the Down and Middle layers. These residuals are subsequently injected into the main UNet during generation at the Up and Middle layers, respectively. In our ablation study, as shown in Fig. 14, we examine the effect of each of these layers and compare their influence to that of the default ControlNet conditioning scale, which reduces the conditioning effect on the output image. We define the parameters λ_{scale}, λ_{down}, and λ_{mid} to control the default conditioning scale, the Down layer residuals, and the Mid layer residuals, respectively. The default parameter λ_{scale} limits the conditioning effect by scaling the residuals, while λ_{down} and λ_{mid} restrict conditioning by applying it over fewer timesteps. Additionally, since some residuals are injected through convolution layers that are not analyzed by our method, we introduce λ_{convs} to similarly limit the influence of convolutional-based layers.

As observed in Fig. 14, each layer group exerts a distinct effect on the generation process. Adjusting λ_{scale} and λ_{mid} (top three rows) results in an uneven interpolation between full conditioning and no conditioning. In these cases, the generated images exhibit minimal changes across most λ values (1.0 to approximately 0.3) before transitioning sharply (from 0.3 to 0.1) to images without any conditioning constraints. In contrast, interpolating over λ_{down} (bottom two rows) reveals that the generated images progressively relax their adherence to the fine details of the conditioning structure image. This allows geometric style elements to emerge without compromising the broader structure of the image. Moreover, our experiments demonstrate that λ_{convs} plays a significant role in incorporating geometric information in a visually pleasing manner.

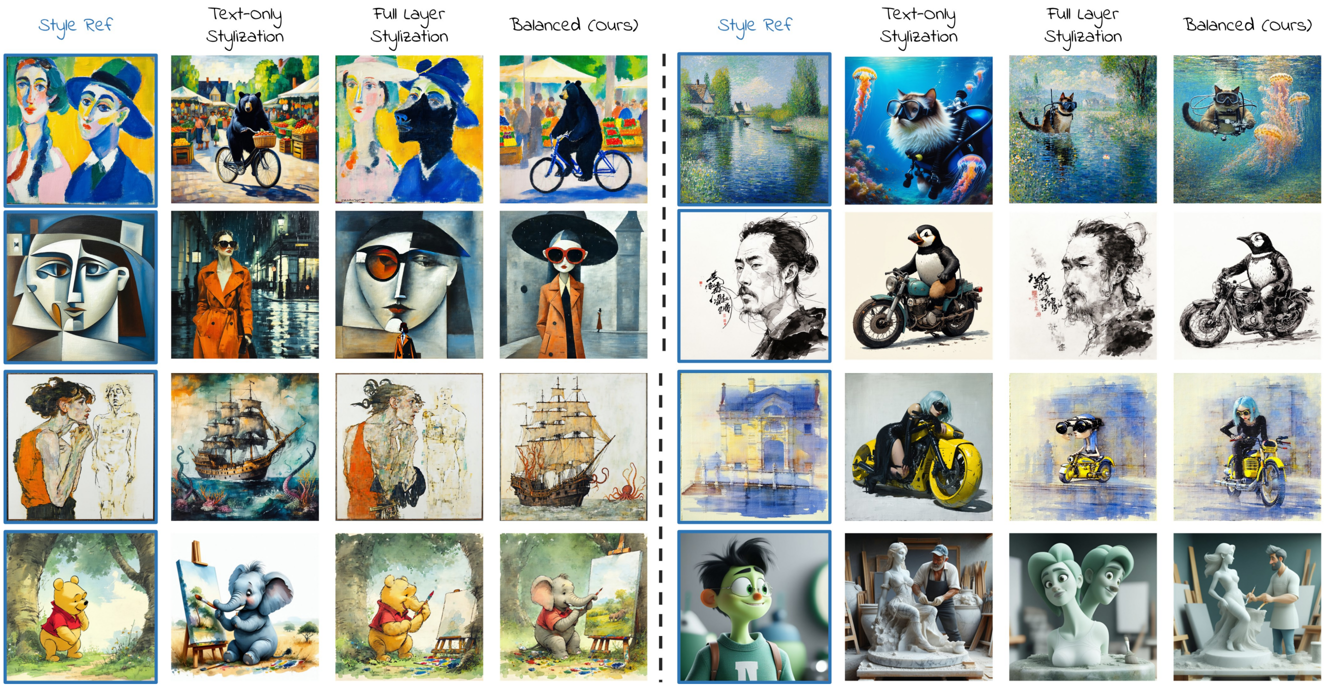


Figure 15. **SD3.5-Large Results.** *Results for text-only stylization, full layer stylization and balanced conditioning with SD3.5-Large. Prompts: "A black bear riding a bicycle in bustling market. The market is full of stands selling fruits and vegetables," "a Siamese cat wearing scuba diving gear, horizontally scuba diving in a deep blue sea, watching a school of colorful jellyfish," "A woman wearing a dark orange trench coat and large sunglasses walking in the cold streets of London," "A penguin riding a motorcycle," "A pirate ship sailing in the ocean and being attacked by the Kraken," "A light-blue haired woman wearing a black attire and black steampunk goggles, leaning on a futuristic yellow motorcycle," "An elephant painting the Savannah. He is sitting in front of a canvas holding a paint brush with his trunk," and "A sculptor working in his studio. He is in the middle of sculpting a marble statue which starts to resemble a female figure."*

These findings are consistent with our analysis in Sec. 5, which highlight the high sensitivity of the Up layers in the denoising UNet to geometric style. From this, we conclude that λ_T, which controls the conditioning injections in the Up layers over the timesteps of the generation process, enables interpolation over the amount of geometric style present in the output image.

10. Appendix C. — Stable Diffusion 3

10.1. Style Conditioning

Recently, the Stable Diffusion 3 (SD3) model family [11] was released, offering new image generation diffusion models. Unlike SDXL [26] which uses a UNet based on Self-Attention and Cross-Attention layers, the SD3 models are based on a Joint-Attention layers which processes both image information and text information. For this reason to apply balanced conditioning on these models we adapted the stylization ideas suggested by Hertz et al. [17], which are Self-Attention based, to the Join-Attention architecture.

Like Hertz et al. [17], we apply stylization by applying AdaIN [19] between the attention features of a generated style image and the target image, and sharing the features of the Keys and Values on their projections. Since the Joint-Attention layer concatenates the Query, Key, and Value projections to the text encodings, we apply these operations before the concatenation to prevent changes in the text fea-

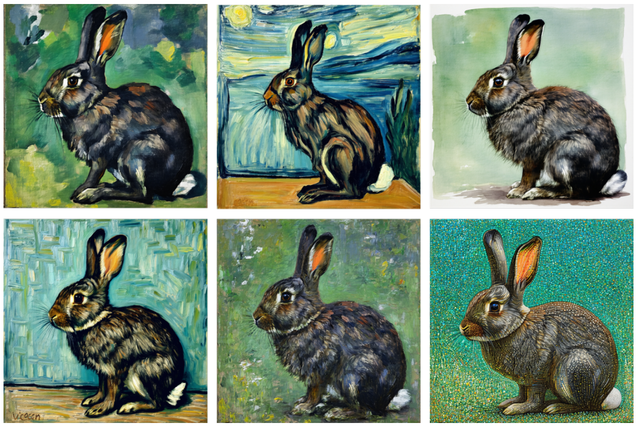


Figure 16. **Collection Examples.** *Representatives from the style clusters generated during the SD3.5-Large analysis. Each sample represents a different style.*

	Easy						Complex						Easy + Complex	
	Text		Depth		Canny		Text		Depth		Canny		Averaged	
Methods	Content	Style	Content	Style	Content	Style	Content	Style	Content	Style	Content	Style	Content	Style
Jeong et al.	0.277	0.625	0.278	0.563	0.293	0.485	0.292	0.630	0.289	0.579	0.316	0.487	0.289	0.578
InstantStyle	0.299	0.431	0.303	0.365	0.308	0.311	0.340	0.439	0.345	0.411	0.352	0.346	0.323	0.396
B-LoRA	0.296	0.493	0.304	0.376	0.308	0.327	0.352	0.443	0.363	0.361	0.364	0.302	0.329	0.404
StyleAligned	0.273	0.592	0.295	0.459	0.300	0.408	0.319	0.537	0.348	0.429	0.355	0.383	0.310	0.492
B-LoRA (Balanced - 10 Layers)	0.291	0.548	0.291	0.533	0.293	0.528	0.346	0.495	0.349	0.479	0.352	0.499	0.319	0.515
B-LoRA (Balanced - 20 Layers)	0.286	0.575	0.288	0.547	0.292	0.540	0.339	0.522	0.344	0.509	0.350	0.511	0.315	0.537
StyleAligned (Balanced)	0.297	0.504	0.296	0.501	0.297	0.497	0.351	0.482	0.349	0.480	0.353	0.468	0.323	0.489

Table 2. Comparison of methods across Easy and Complex prompts conditioned with and without Depth and Canny Conditioning.

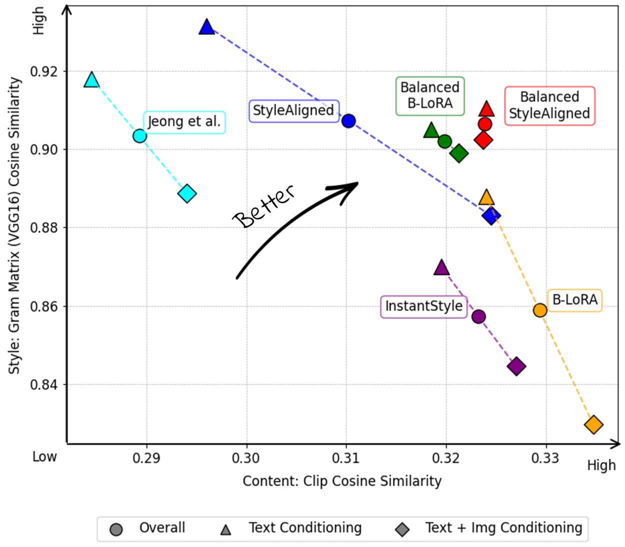


Figure 17. **Gram Based Evaluation.** *Style and content evaluation using Gram-Matrix representation and Clip embeddings.*

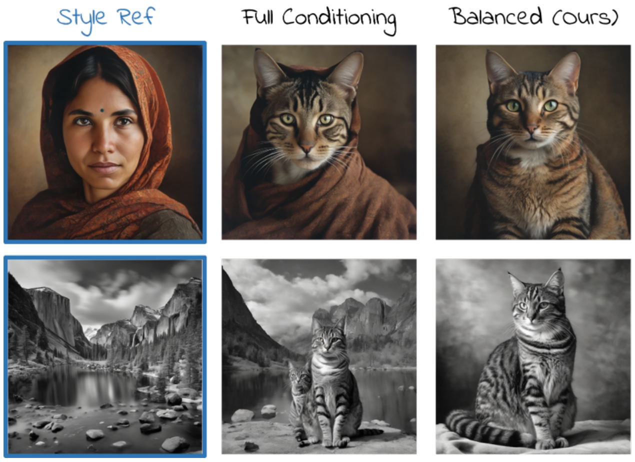


Figure 18. **Photographic-Editing Styles.** *"A tabby cat". Style sensitive layers analyzed by our method are not limited to painting styles, but show sensitivities to other styles like photographic editing styles. Notice that using full conditioning may result with content drift and artifacts.*

tures. During our experiments we noticed that unlike Hertz et al. which applies AdaIN only on the Key and Query projections of the attention layers, applying AdaIN on the Value projections significantly contributes to the stylization of the target image and is key for achieving a satisfying result. For this reason we incorporate this change to our style conditioning algorithm for SD3.5-Large.

10.2. Analysis

We use our method presented in Sec. 4 to find SD3.5-Large style sensitivities as presented in Sec. 5.1. Fig. 16 shows examples from the collections generated during the analysis process where each example is taken from a different style cluster.

As shown in and Fig. 15, SD3.5-Large, like SDXL, struggles with complex conditioning combinations. This is particularly evident in the "Text-Only Stylization" columns, where including only the artist's name in the prompt often results in style mismatches and, in extreme cases, the complete omission of the target style. Conditioning on all layers for style leads to significant content drift, whereas our balancing method effectively aligns the generated results with both the content and style of the reference image. The balanced results we present were achieved by applying style conditioning to 28 out of 38 joint-attention layers ($\lambda_S = 0.73$), which we identified as the optimal setting.

11. Appendix D. — Results

11.1. Qualitative Results

To demonstrate the ability of our conditioning strategy we show additional results produced by the Balanced versions of StyleAligned and B-LoRA for "Text Only" (Fig. 21), "Text + Canny" (Fig. 22), and "Text + Depth"' (Fig. 23). We share additional qualitative comparisons between the balanced versions of StyleAligned and B-LoRA with the benchmark methods from Sec. 6 in Figs. 24 to 27. In addition, we compare ourself to two additional recent methods: RB-Modulation [31] and InstaStyle [8]. Since both methods provide access to their model only through an interactive web interface we could not evaluate their results using Canny or Depth conditioning. For this reason we refrained

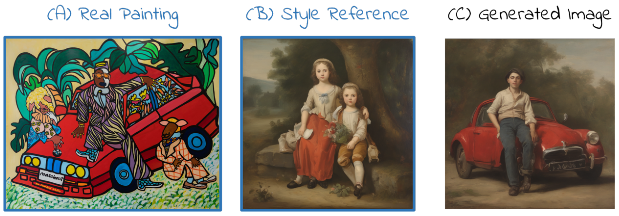


Figure 19. **Unfamiliar Styles.** *A limitation example, where the style of an artist (A) is unknown to the base model, causing a mismatch in the reference image (B) which then results with a style mismatch in the target image (C), even when the content matches the artist.*

from using their methods in our main comparison in Sec. 6. For fairness, we present a qualitative comparison in Fig. 28 using a style reference and a text prompt without any structure conditioning.

11.2. Style Variation

To assess the generality of our method in identifying style-sensitive layers, we extend our experiments to styles beyond artistic paintings. Fig. 18 illustrates the applicability of our style layers to photographic-editing styles. Notably, the same style layers identified for artistic paintings enable our method to generate photographs that adopt a reference style while preserving content integrity and avoiding unwanted artifacts and content drift from the style image.

11.3. Quantitative Results

We present a breakdown of our quantitative results in Tab. 2. We show the results over Easy and Complex prompts, for all conditioning types: "Text Only," "Text + Depth," and "Text + Canny." We add an evaluation result for balancing B-LoRA based on 20 layers, which we found optimal for text conditioned generation. In addition we show the balanced version used in Sec. 6 which is based on 10 layers.

Since style representation is still an active area of research, we provide an additional evaluation of our results using Gram matrices [15] as style descriptors instead of Dino [5] features. As shown in Fig. 17, the Gram matrix based results are consistent with those presented in Sec. 6, further highlighting the effectiveness of our method in achieving a balanced representation of both content and style.

11.4. User Study Details

Study Design and Participant Demographics The user study aimed to quantitatively evaluate the impact of balancing methods on the perceived quality of images conditioned on content and style prompts. A total of 42 anonymous participants took part, representing diverse backgrounds. The cohort included 62% male participants, distributed across the following age groups: 16 participants aged 25–32, 10 aged 33–38, 10 aged 39–45, and 6 participants aged over 45. Professional affiliations spanned research and development (31%), computer science graduate studies (24%), professional artistry (19%), and UX design (9%), ensuring a broad spectrum of expertise relevant to the evaluation task.

Experimental Setup The study consisted of three tasks: a multi-choice comparison and two A/B tests, detailed in the main manuscript. Each task was designed to assess how well balanced and imbalanced methods align with both content and style, as perceived by users. The stimuli were generated by sampling from our dataset of text prompts and style reference images. Stratified sampling ensured a balanced representation of prompt complexity ("easy" vs. "complex") and conditioning techniques (e.g., Canny, Depth, and Text-only).

To eliminate biases, no style image was repeated across tasks, and the presentation order of images was randomized. Importantly, participants were not informed of the underlying generation method. The study was conducted online, with participants completing the evaluation independently, ensuring no researcher supervision or bias influenced the results.

Tasks and Protocols

1. **Multi-Choice Test**: Participants selected the best image from six options based on alignment with both the text prompt and style reference. Two of the six options in each instance were generated using balanced methods. The test encompassed 15 unique content-style pairings to ensure variety and robust statistical analysis.
2. **A/B Tests**: Each test involved binary comparisons between a method and its balanced counterpart. One test focused on B-LoRA, while the other evaluated StyleAligned. Both tests followed the same content-style alignment criterion and included six unique pairings for each method.

Figures illustrating the test interfaces and sample questions can be found in Fig. 36 and Fig. 37.

Statistical Analysis. Results were analyzed using a Chi-Squared Test for Independence to assess the preference for balanced versus imbalanced methods. The null hypothesis assumed no difference in user preference. For the multi-choice test, the expected probability of selecting balanced methods was set at $\frac{1}{3}$, based on their representation among the six options. The observed preferences significantly diverged from the null hypothesis, as shown in Tab. 1 of the main manuscript.

The study design and statistical robustness demonstrate a clear and significant preference for balanced methods, val-

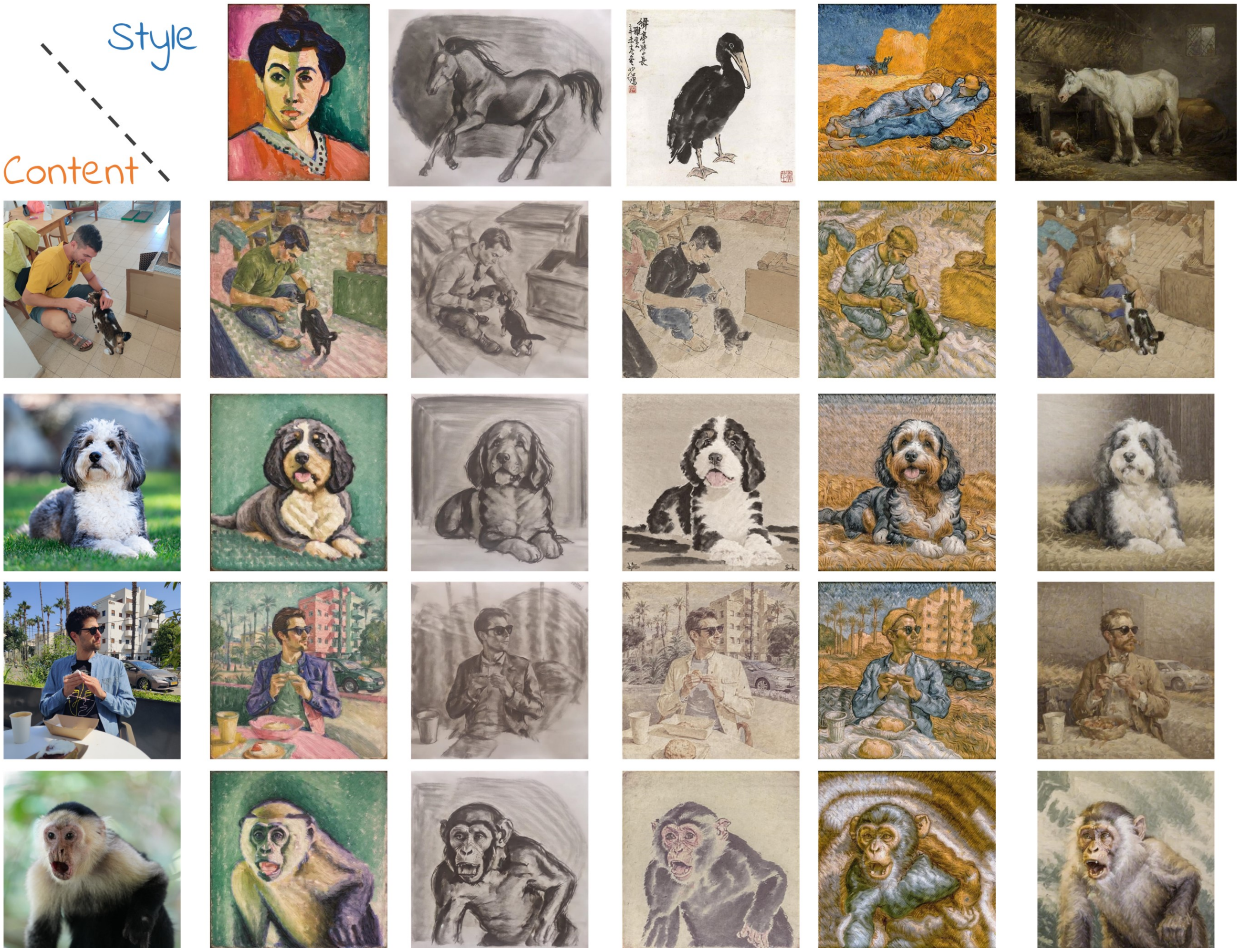


Figure 20. **Style Transfer.** *Results generated using our balanced version of B-LoRA. Please zoom in for a better view.*

idating their efficacy in improving the visual alignment of content and style.

11.5. Method Limitations

As described in the main manuscript, the main limitations of our method arise from its dependence on the capabilities of the base model. For instance, when generating images with a style unfamiliar to the base model, the result may exhibit an unintended style, as the model lacks sufficient knowledge to properly generate the style reference, leading to a style mismatch. We demonstrate these limitations in Fig. 19. As can be seen, the unique style of Enfant Précoce (A) is unknown to SDXL, thus the reference image (B) and output image (C) fail to match his unique style. This issue is not caused by a content misalignment issue, as the style fails to match the artist even when using a prompt that describes the content of an existing painting by the artist. In our demonstration, we use the prompt: *"A man leaning on a red car surrounded by trees."*

12. Appendix E. — Additional Applications

12.1. Style Transfer

To perform style transfer given two content and style images we use our balanced version of B-LoRA. For content alignment we use a Canny edge map as we find it the best option for preserving the structure and alignment of the given content image. For stylization we employ B-LoRA's approach and fine-tune residual LoRA weights on the given style image. Since B-LoRA does not change its stylization layer decision for each timestep, we rank the layers based on their average rank over all timesteps (see Fig. 11). In Sec. 6 we base our balanced version on 10 self-attention layers (20 including cross-attention layers) for fairness reasons, as it closely approximates the number of stylization layers used by B-LoRA. In practice we find that using a larger number of layers improves style fidelity. We experiment by using B-LoRA with various layer decisions (Fig. 10), guided by our layer ranking and we find that basing our choice on the 20 best self-attention layers (40 with cross-attentions) strikes a fine balance between content and style. We show a quantitative ablation between B-LoRA and the two balanced variants in Tab. 2. In addition, we show qualitative examples produced the balanced version of B-LoRA in Fig. 20.

12.2. Material Generation

As shown in Sec. 6, using our balancing strategy yields geometric style freedom when generating artistic images. Another result of this is better generation of material style. We show results in Fig. 29 and Fig. 30. As can be seen, by applying our balancing strategy StyleAligned gains the ability to generate physical aspects of different materials even when conditioned on a content image. The regular version of StyleAligned forces unnecessary conditional information on the output on the content image, which results in patterns that do not match the material.

12.3. ReStyle/ReContent

Copying the works of old masters is a time-honored tradition in the art world, dating back to the origins of painting itself. This practice serves as a tool for artists to refine their techniques and develop their unique personal styles. Throughout history, many renowned painters have engaged in this approach - examples include Vincent van Gogh, who copied works by Jean-François Millet, and Pablo Picasso, who reinterpreted works by Diego Velázquez such as Las Meninas. This tradition has even given rise to several iconic artworks, such as Edgar Degas' studies of Old Masters like Nicolas Poussin and Rembrandt, or Francis Bacon's reimaginings of Diego Velázquez's Portrait of Pope Innocent X. Inspired by this classical method of artistic learning, we utilize our stylization approach, which enables the application of distinctive styles to the works of old masters - a process we call ***ReStyle***. Additionally, our method extends the model's geometric flexibility, allowing for the re-imagining of an artwork's content in innovative ways - a feature we refer to as ***ReContent***.

To achieve this flexibly-conditioned image editing capability, we first use the original artwork as a structural condition, employing either a Canny or Depth map. We then generate the edited image using a relatively high style weight, $\lambda_S \approx 0.55$, a low content weight, $\lambda_T < 0.2$, and a descriptive text prompt. Setting these values for λ_S and λ_T allows for a strong resemblance to the artistic style of the style condition while providing geometric flexibility. This not only ensures a fine resemblance to the style condition but also enables content modifications to the original image through the text prompt. Examples are shown in Figs. 31 and 32. As demonstrated, our approach effectively edits both the style and content of the original image while preserving its underlying structure and general characteristics, even when the generated content shift is significant.

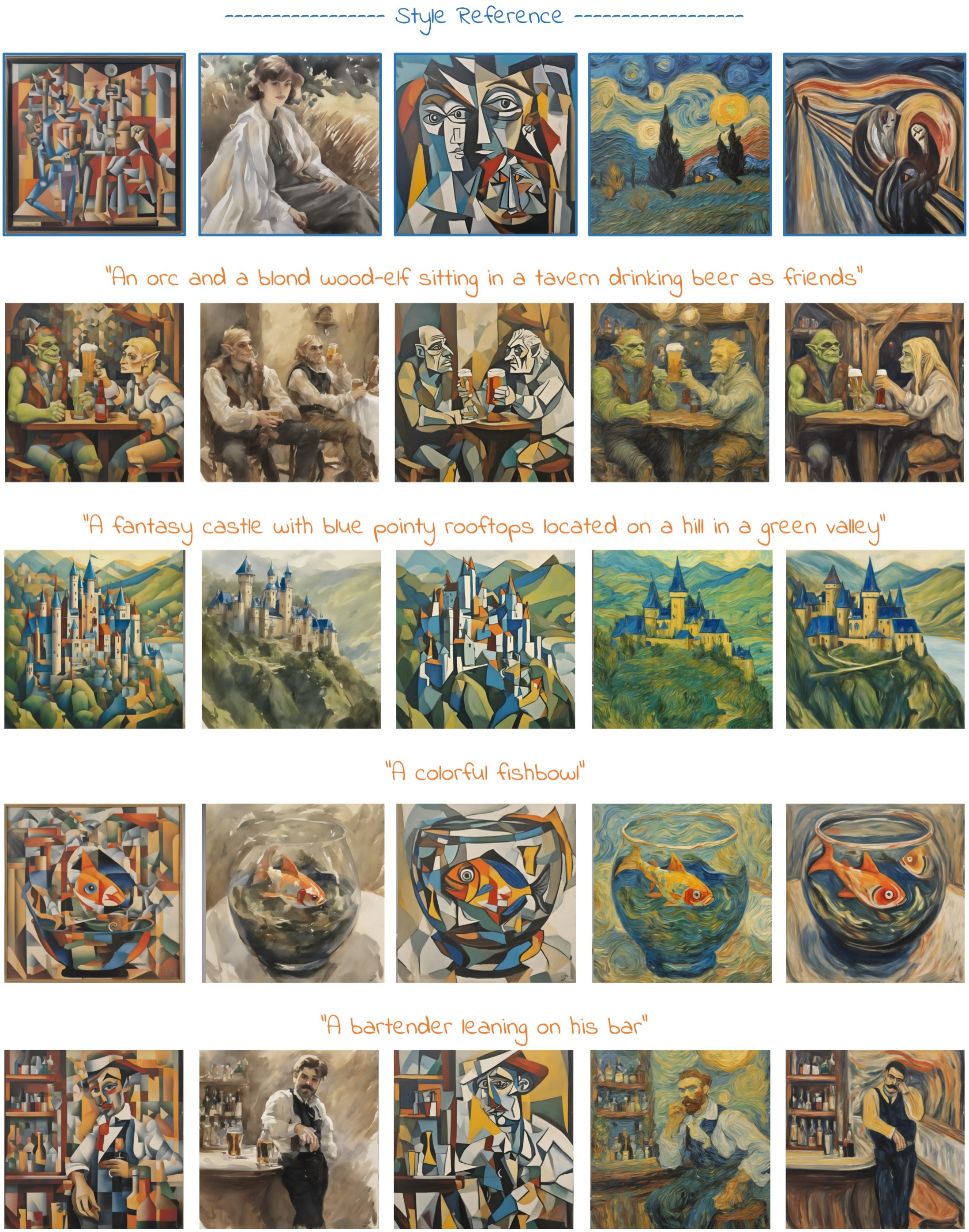

Style Reference
"An orc and a blond wood-elf sitting in a tavern drinking beer as friends"
"A fantasy castle with blue pointy rooftops located on a hill in a green valley"
"A colorful fishbowl"
"A bartender leaning on his bar"

Figure 21. **Text Conditioned Results.** *Zoom in for a better view.*

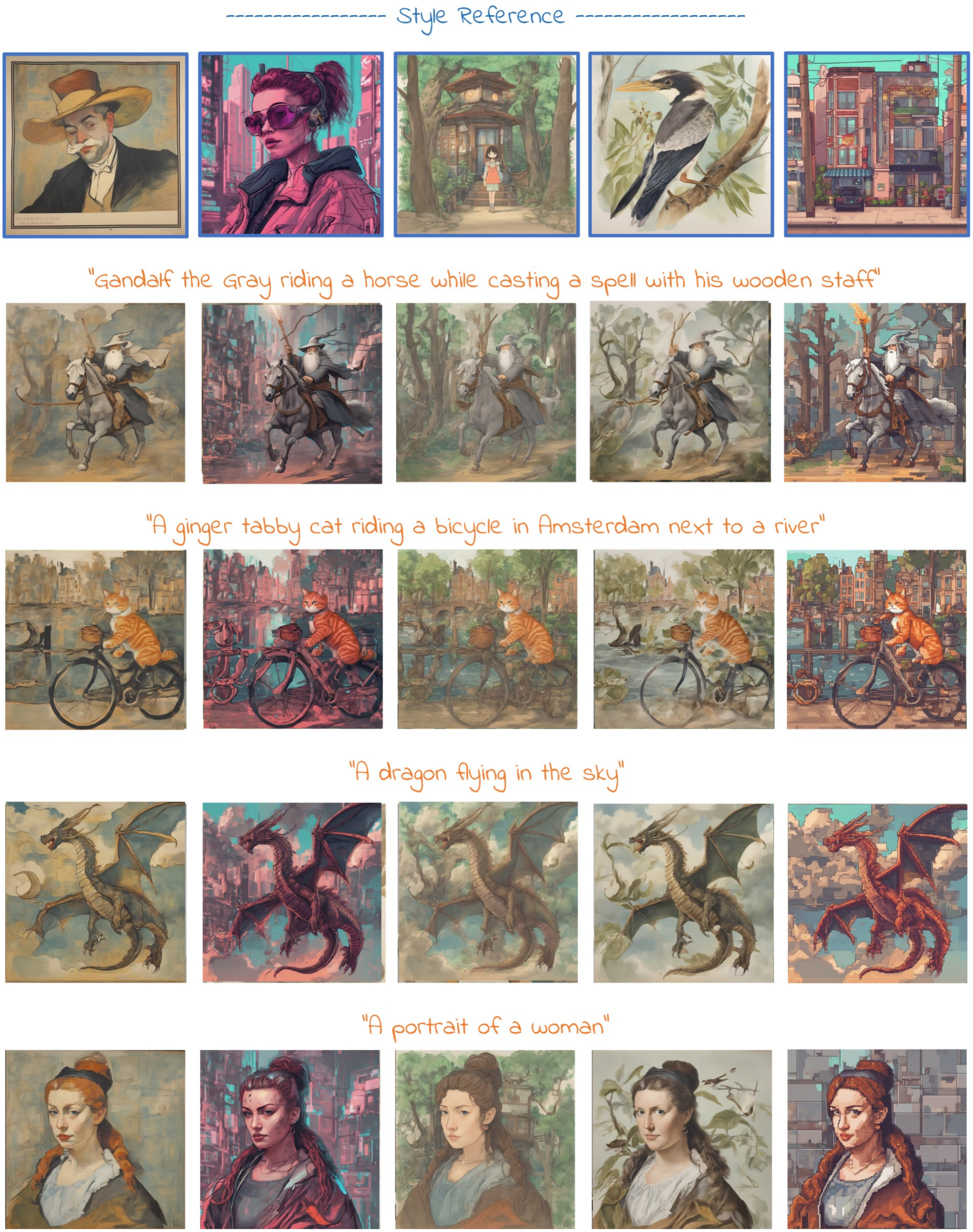

Style Reference
"Gandalf the Gray riding a horse while casting a spell with his wooden staff"
"A ginger tabby cat riding a bicycle in Amsterdam next to a river"
"A dragon flying in the sky"
"A portrait of a woman"

Figure 22. **Canny Conditioned Results.** *Zoom in for a better view.*

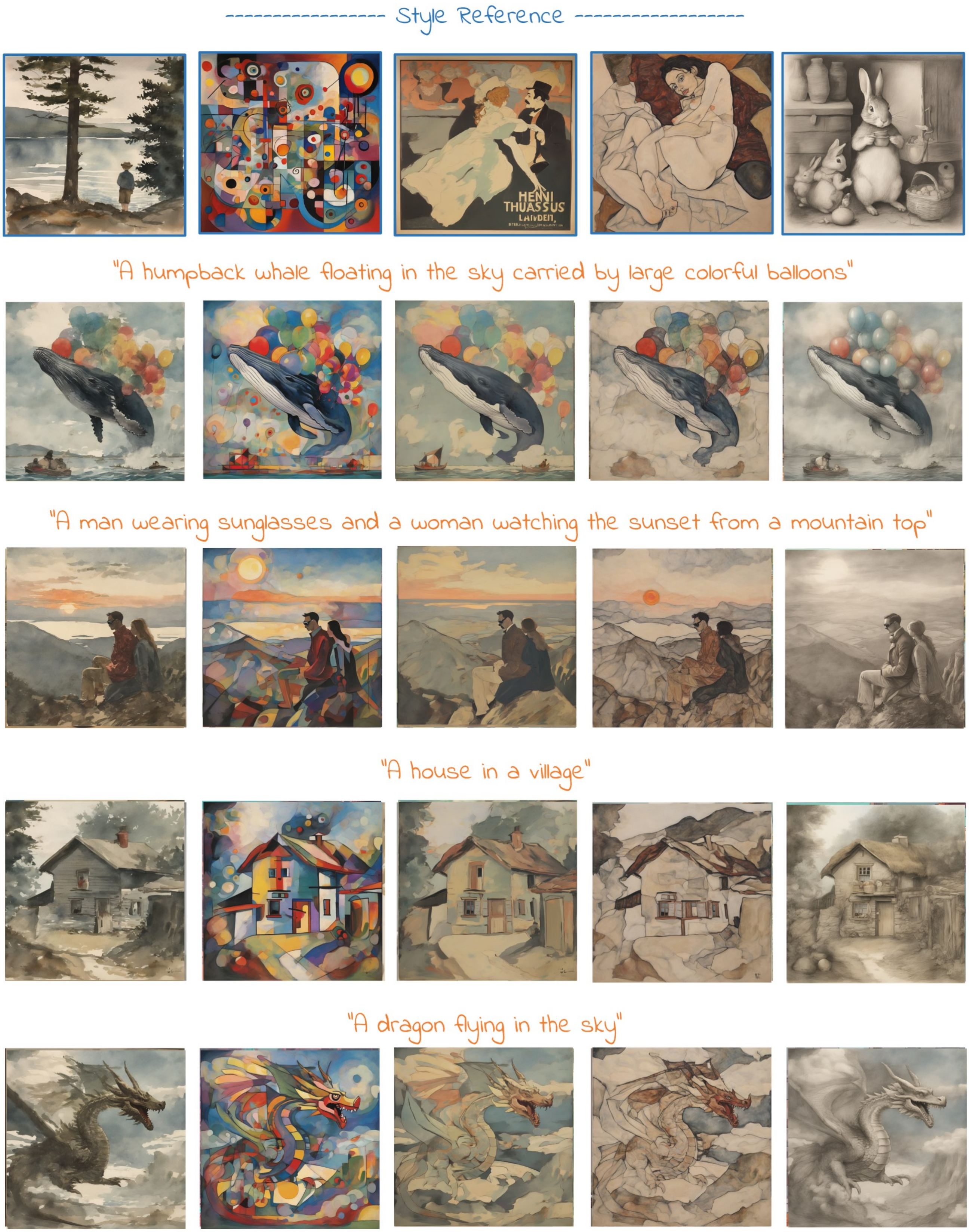

Style Reference
HENNI THUASSUS LAINDEIT
"A humpback whale floating in the sky carried by large colorful balloons"
"A man wearing sunglasses and a woman watching the sunset from a mountain top"
"A house in a village"
"A dragon flying in the sky"

Figure 23. **Depth Conditioned Results.** *Zoom in for a better view.*

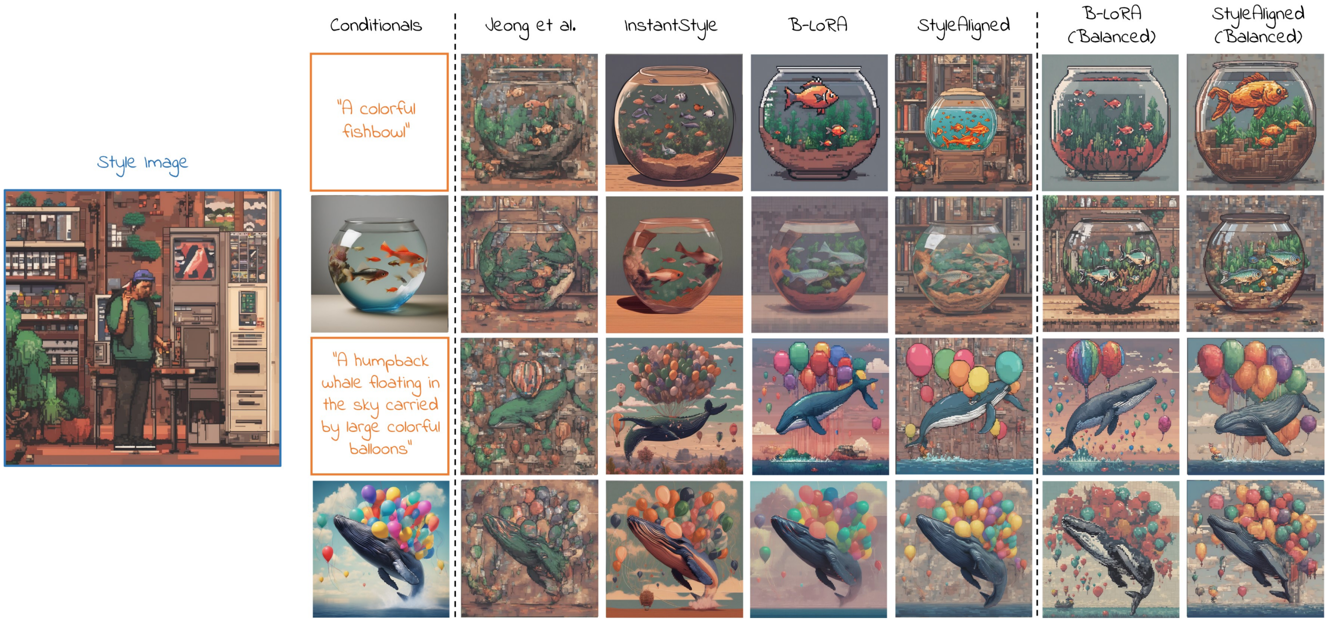


Figure 24. **Qualitative Comparison**. *A comparison of different conditional combinations: Easy vs Complex prompt (two first rows vs. two last rows), Text only vs. Text and content image conditioning (1,3 vs 2,4 rows). As can be seen, both balanced methods achieves consistency over all conditioning combinations while the imbalanced methods show an inconsistent generation quality and in some examples content and style issues.*

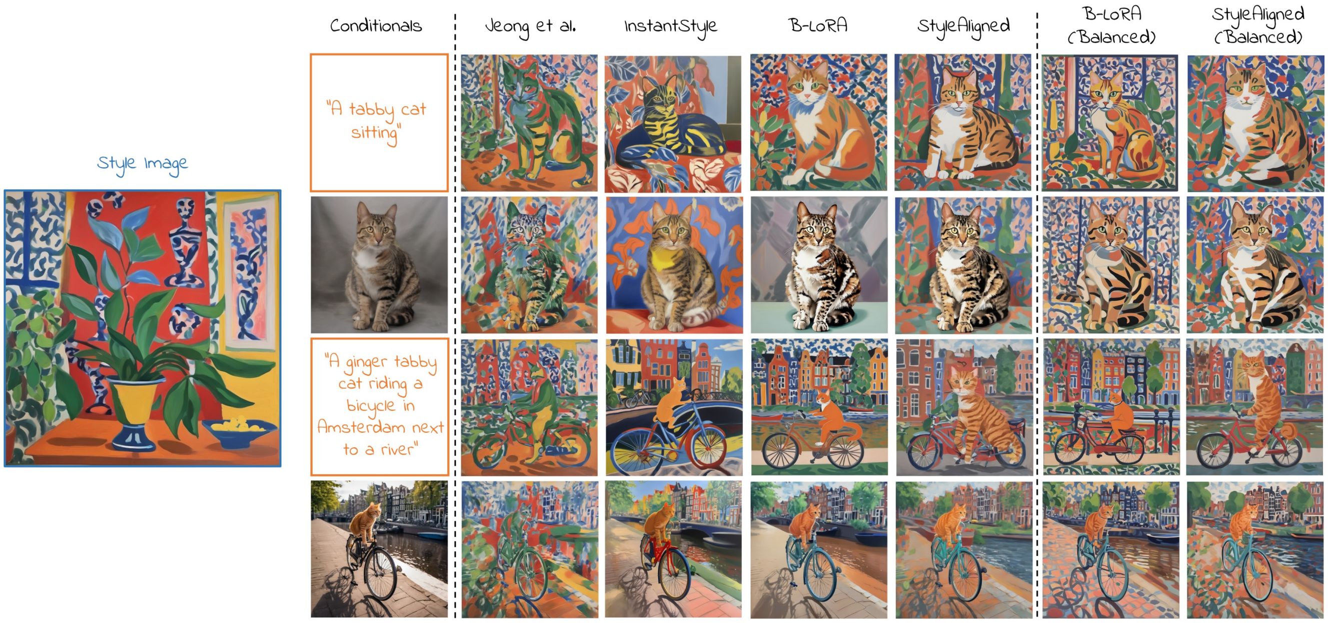


Figure 25. **Qualitative Comparison**. *A comparison of different conditional combinations: Easy vs Complex prompt (two first rows vs. two last rows), Text only vs. Text and content image conditioning (1,3 vs 2,4 rows). As can be seen, both balanced methods achieves consistency over all conditioning combinations while the imbalanced methods show an inconsistent generation quality and in some examples content and style issues.*

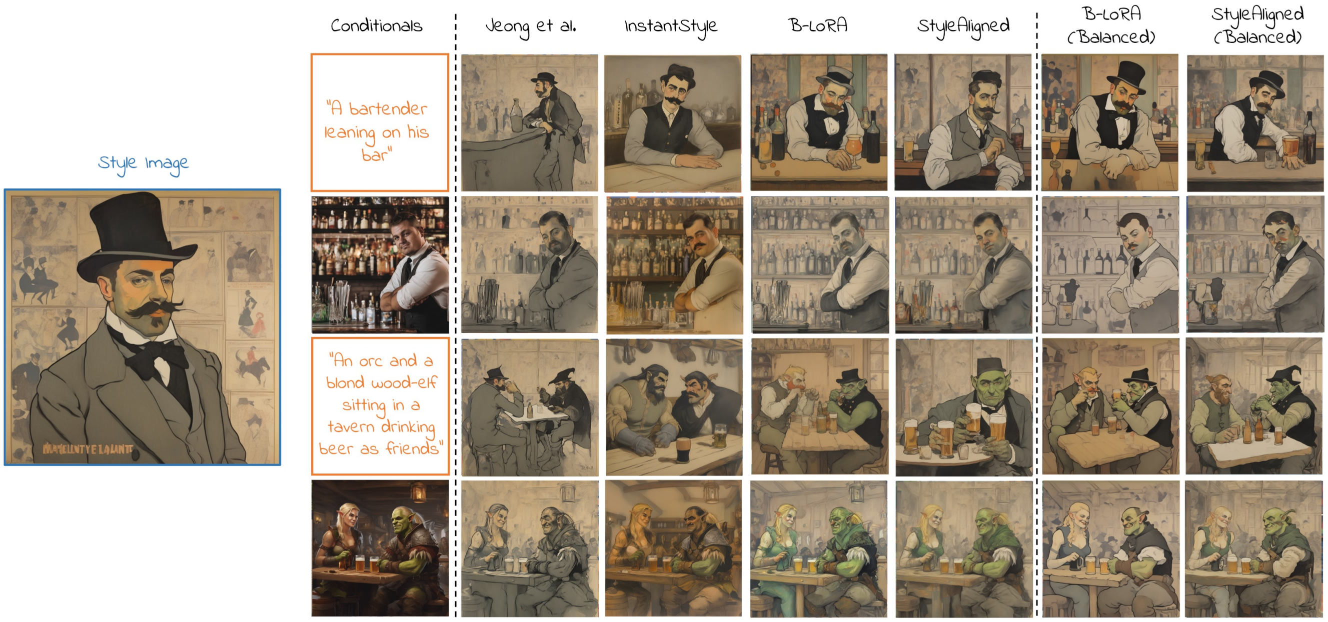

Style Image
Conditionals
Jeong et al.
InstantStyle
B-LoRA
StyleAligned
B-LoRA (Balanced)
StyleAligned (Balanced)
"A bartender leaning on his bar"
"An orc and a blond wood-elf sitting in a tavern drinking beer as friends"

Figure 26. **Qualitative Comparison**. *A comparison of different conditional combinations: Easy vs Complex prompt (two first rows vs. two last rows), Text only vs. Text and content image conditioning (1,3 vs 2,4 rows). As can be seen, both balanced methods achieves consistency over all conditioning combinations while the imbalanced methods show an inconsistent generation quality and in some examples content and style issues.*

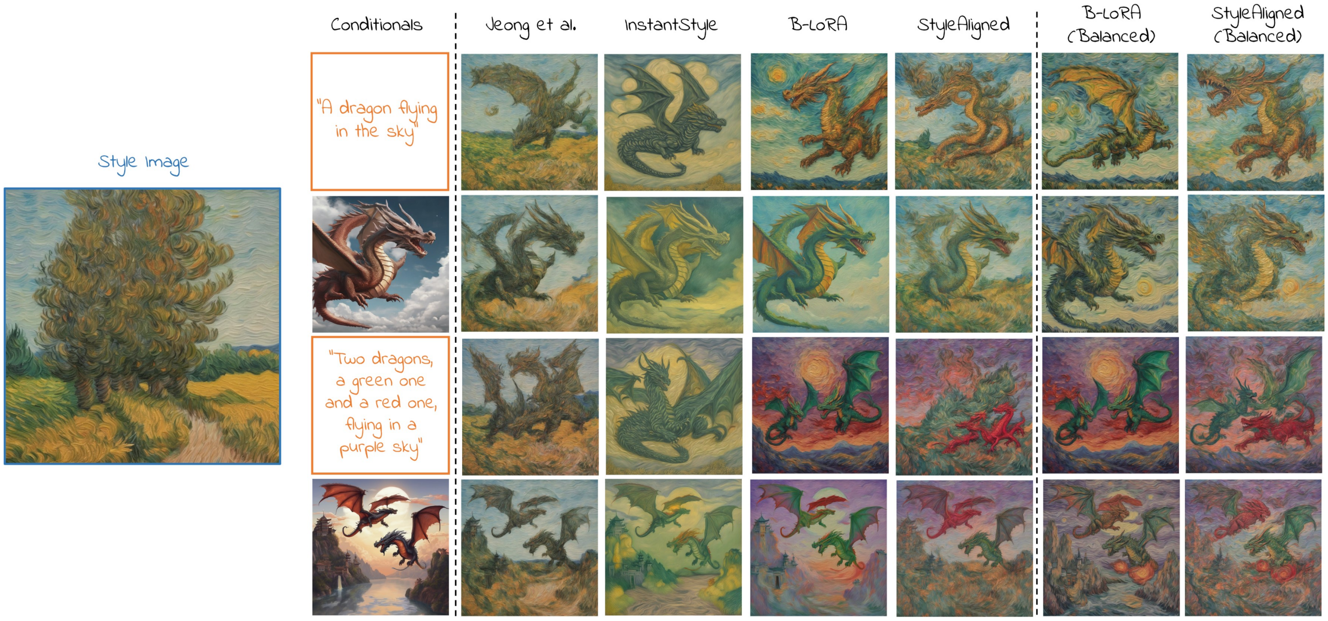


Figure 27. **Qualitative Comparison**. *A comparison of different conditional combinations: Easy vs Complex prompt (two first rows vs. two last rows), Text only vs. Text and content image conditioning (1,3 vs 2,4 rows). As can be seen, both balanced methods achieves consistency over all conditioning combinations while the imbalanced methods show an inconsistent generation quality and in some examples content and style issues.*

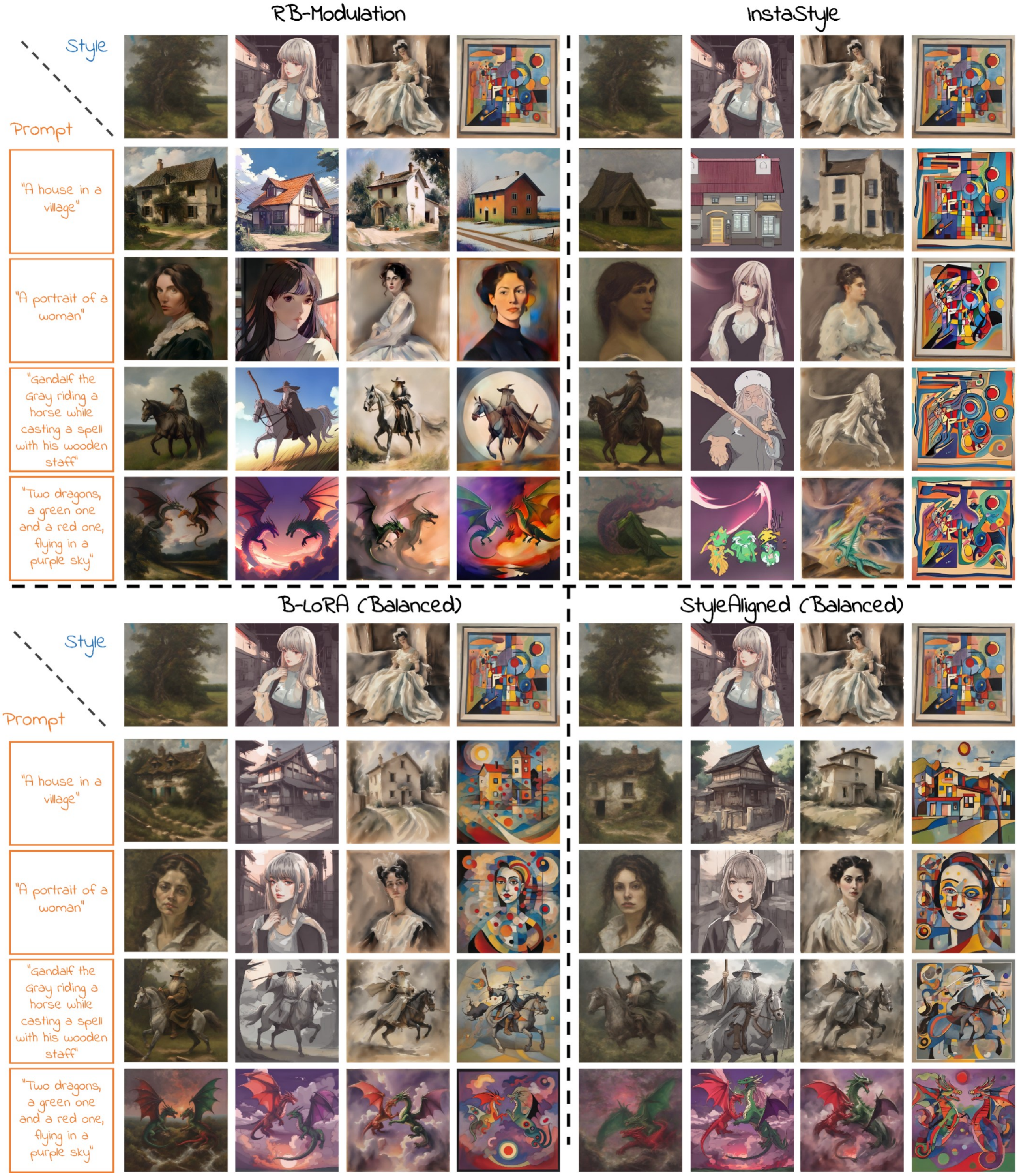

RB-Modulation
InstaStyle
Style
Prompt
"A house in a village"
"A portrait of a woman"
"Gandalf the Gray riding a horse while casting a spell with his wooden staff"
"Two dragons, a green one and a red one, flying in a purple sky"
B-LoRA (Balanced)
StyleAligned (Balanced)
Style
Prompt
"A house in a village"
"A portrait of a woman"
"Gandalf the Gray riding a horse while casting a spell with his wooden staff"
"Two dragons, a green one and a red one, flying in a purple sky"

Figure 28. **Additional Comparisons**. *Prompt + style image conditioned outputs for RB-Modulation (top left), InstaStyle (top right), Balanced B-LoRA (bottom left), and Balanced StyleAligned (bottom right.) Please zoom in for a better view.*

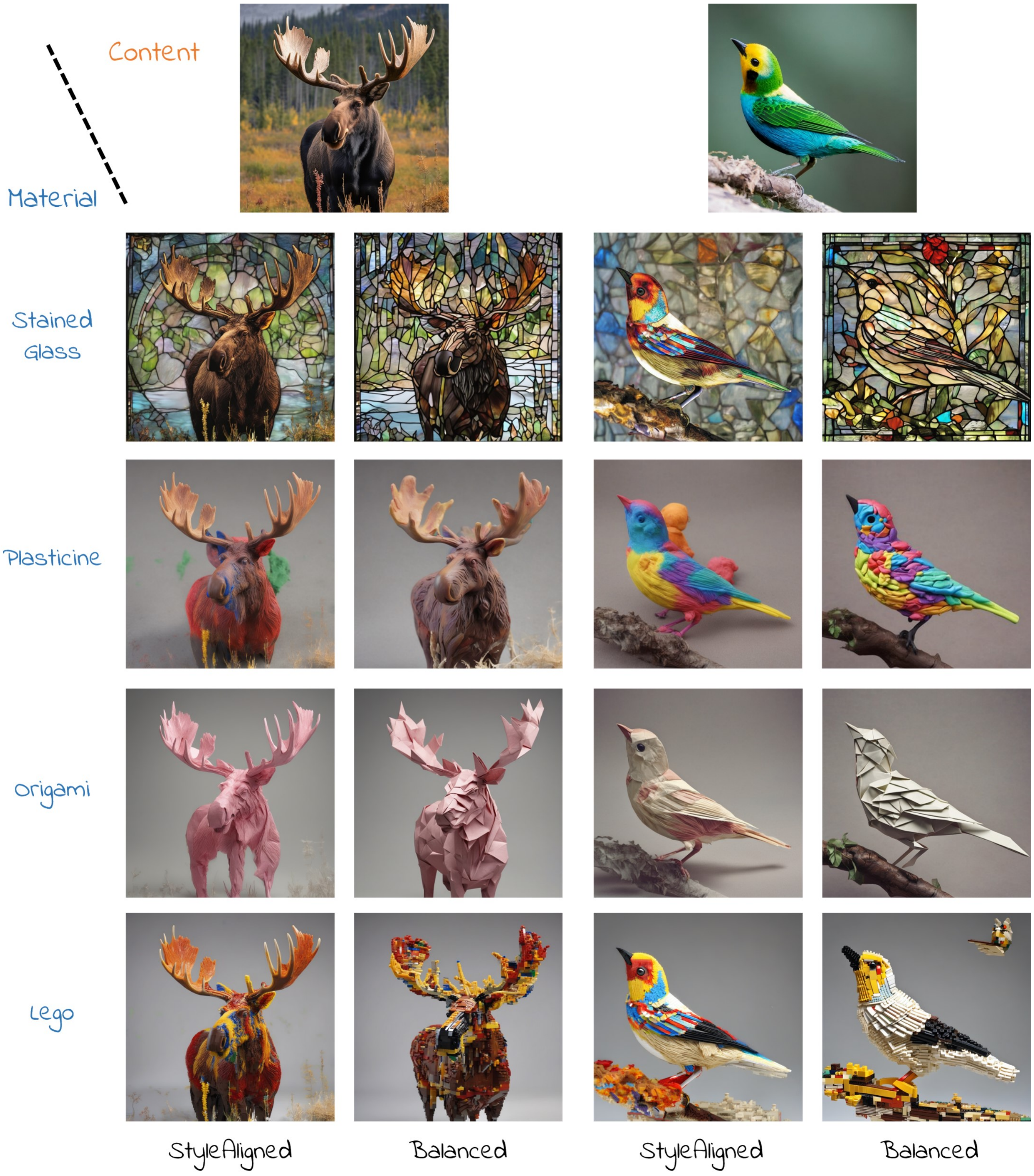


Figure 29. **Material Style Generation**. *A sample of generated images with materialistic style, aligned to content images. Please zoom in for a better view.*

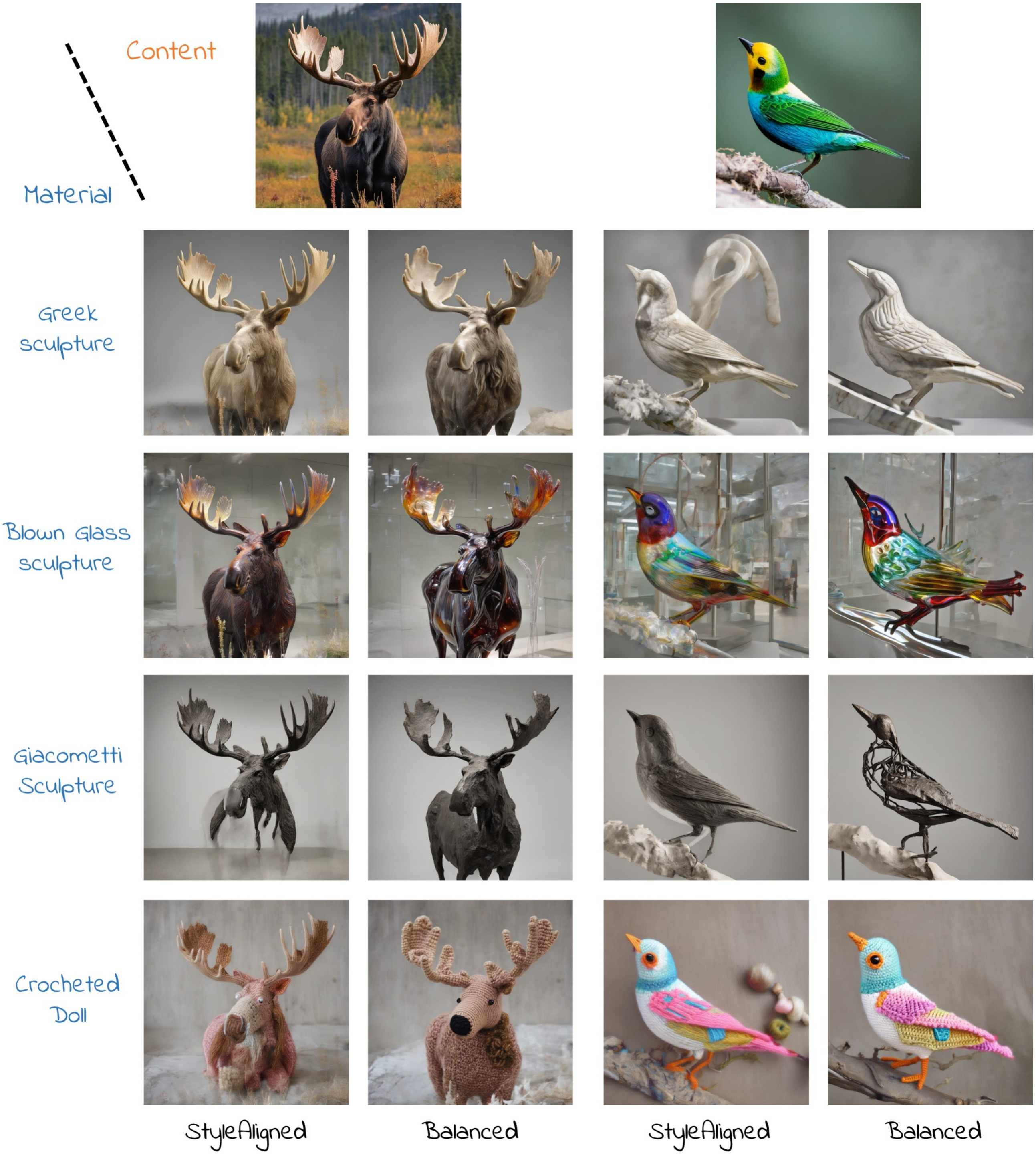


Figure 30. **Material Style Generation**. *A sample of generated images with materialistic style, aligned to content images. Please zoom in for a better view.*

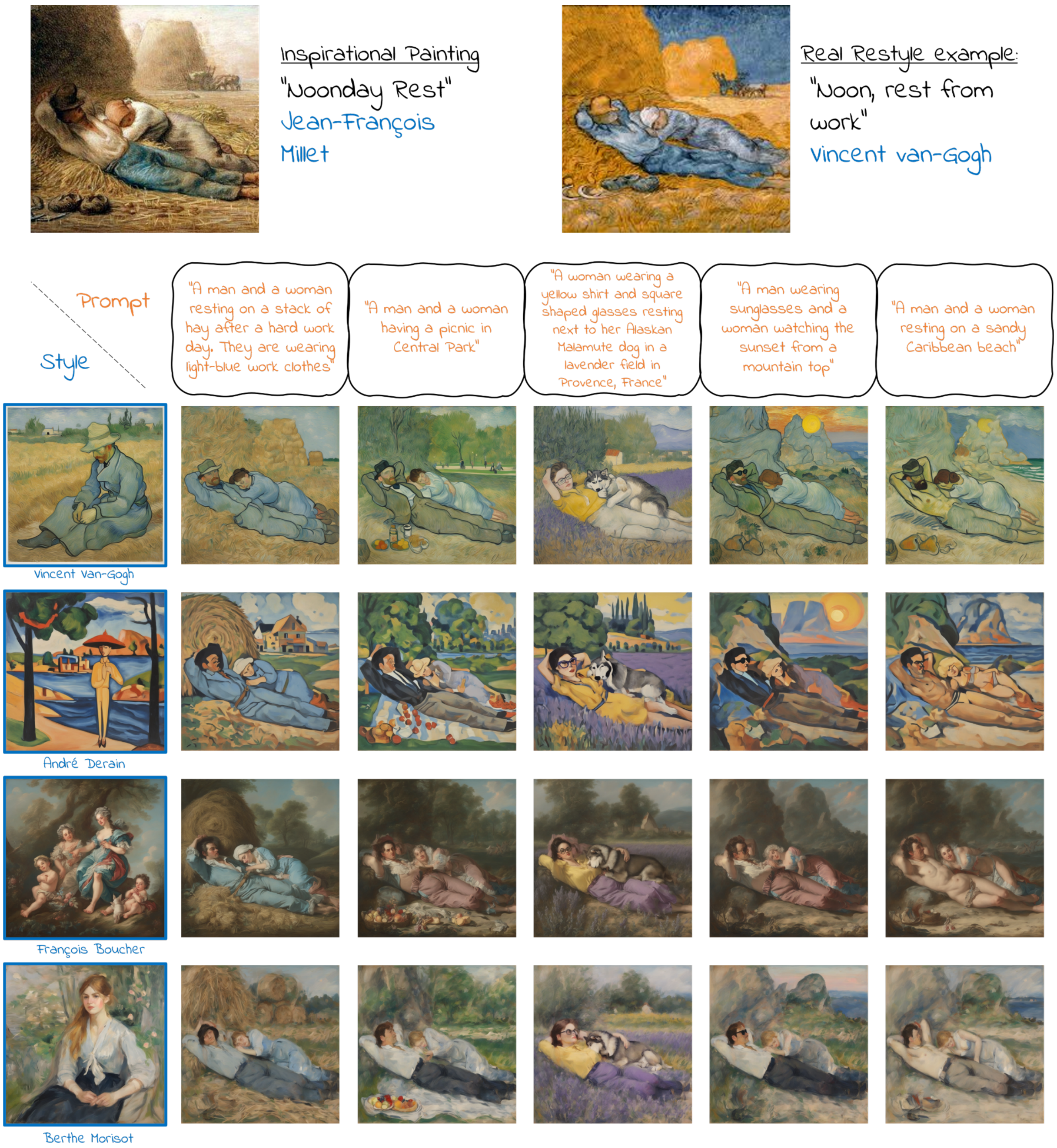

Inspirational Painting
"Noonday Rest"
Jean-François
Millet
Real Restyle example:
"Noon, rest from
work"
Vincent van-Gogh
Prompt
Style
"A man and a woman resting on a stack of hay after a hard work day. They are wearing light-blue work clothes"
"A man and a woman having a picnic in Central Park"
"A woman wearing a yellow shirt and square shaped glasses resting next to her Alaskan Malamute dog in a lavender field in Provence, France"
"A man wearing sunglasses and a woman watching the sunset from a mountain top"
"A man and a woman resting on a sandy Caribbean beach"
Vincent van-Gogh
André Derain
François Boucher
Berthe Morisot

Figure 31. **Restyle/Recontent Example 1**. *An example of restyling a painting inspired by van-Gogh's recreation of "Noonday Rest" by Jean-Francois Millet. The first column shows an example of restyling the content input without changing the original content while the rest of the columns shows an example of ReStyle and ReContent by editing both the image style and content of the output. (Please zoom in for a better view.)*

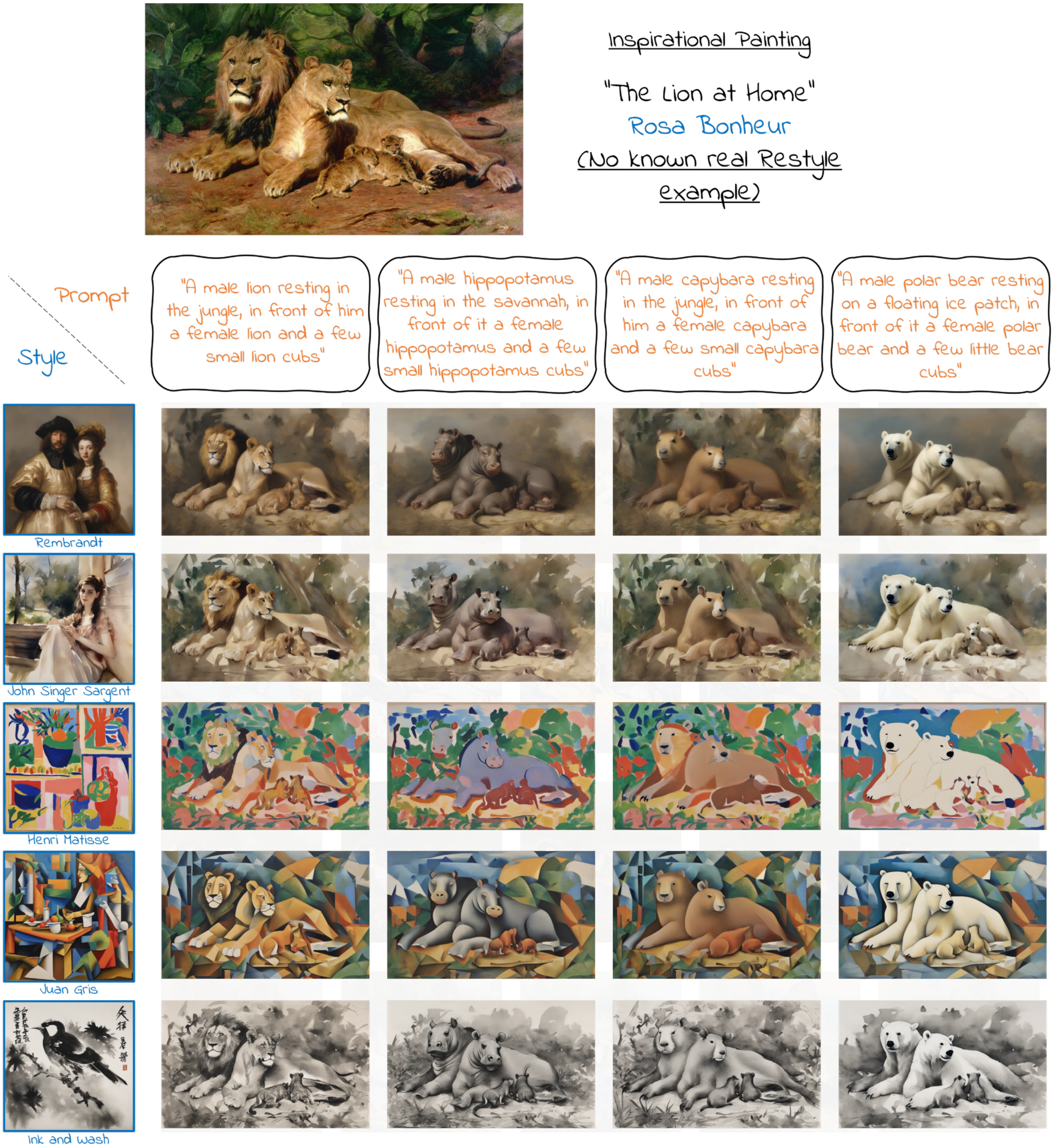

Inspirational Painting
"The Lion at Home"
Rosa Bonheur
(No known real Restyle example)
Prompt
Style
"A male lion resting in the jungle, in front of him a female lion and a few small lion cubs"
"A male hippopotamus resting in the savannah, in front of it a female hippopotamus and a few small hippopotamus cubs"
"A male capybara resting in the jungle, in front of him a female capybara and a few small capybara cubs"
"A male polar bear resting on a floating ice patch, in front of it a female polar bear and a few little bear cubs"
Rembrandt
John Singer Sargent
Henri Matisse
Juan Gris
Ink and wash

Figure 32. **Restyle/Recontent Example 2**. *An example of restyling a painting of Rosa Bonheur: "The Lion at Home". The first column shows an example of restyling the content input without changing the original content while the rest of the columns shows an example of ReStyle and ReContent by editing both the image style and content of the output. (Please zoom in for a better view.)*

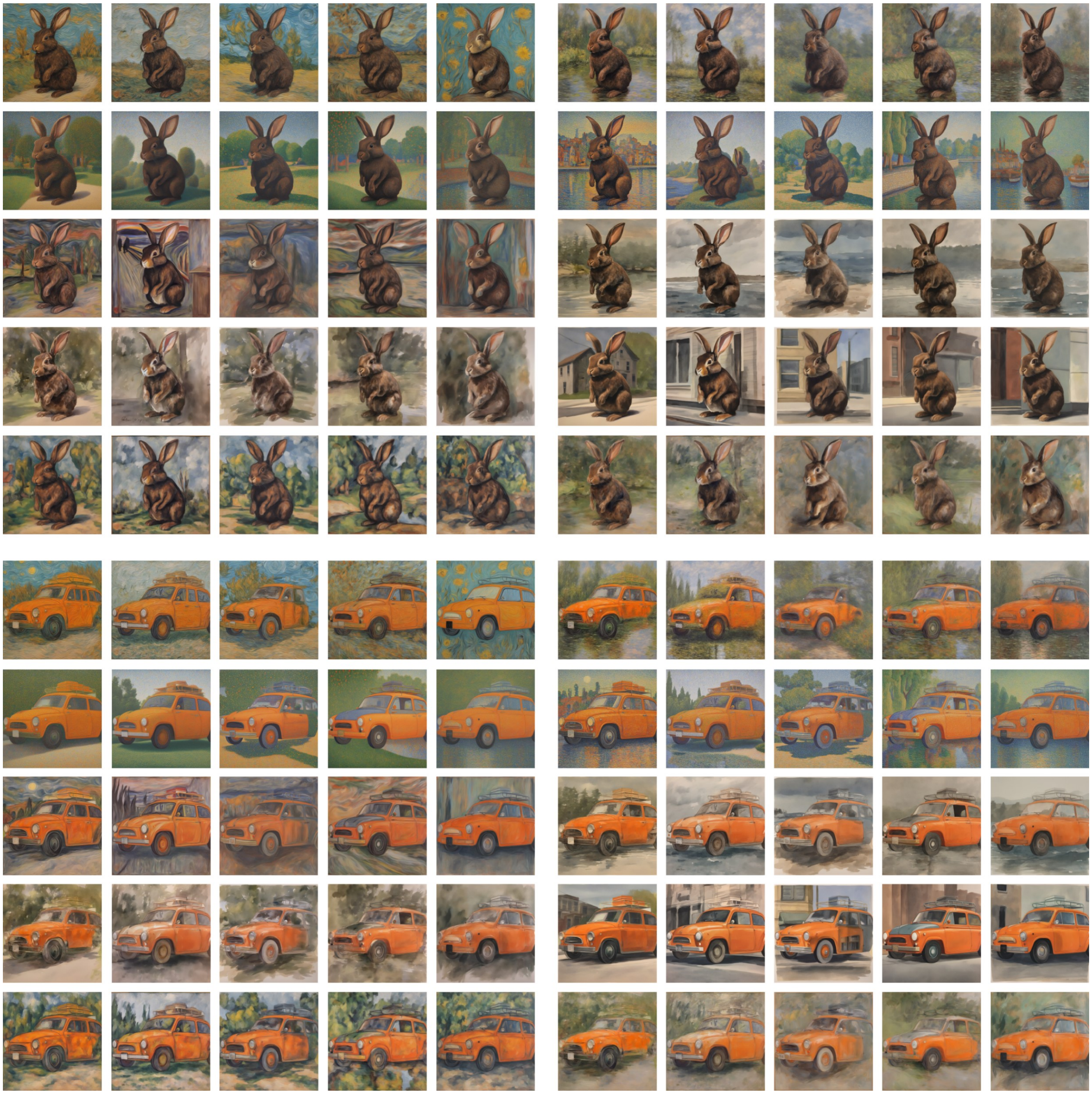


Figure 33. **Style Collections Example**. *An example of a paintings collection used for our style sensitivity analysis.*

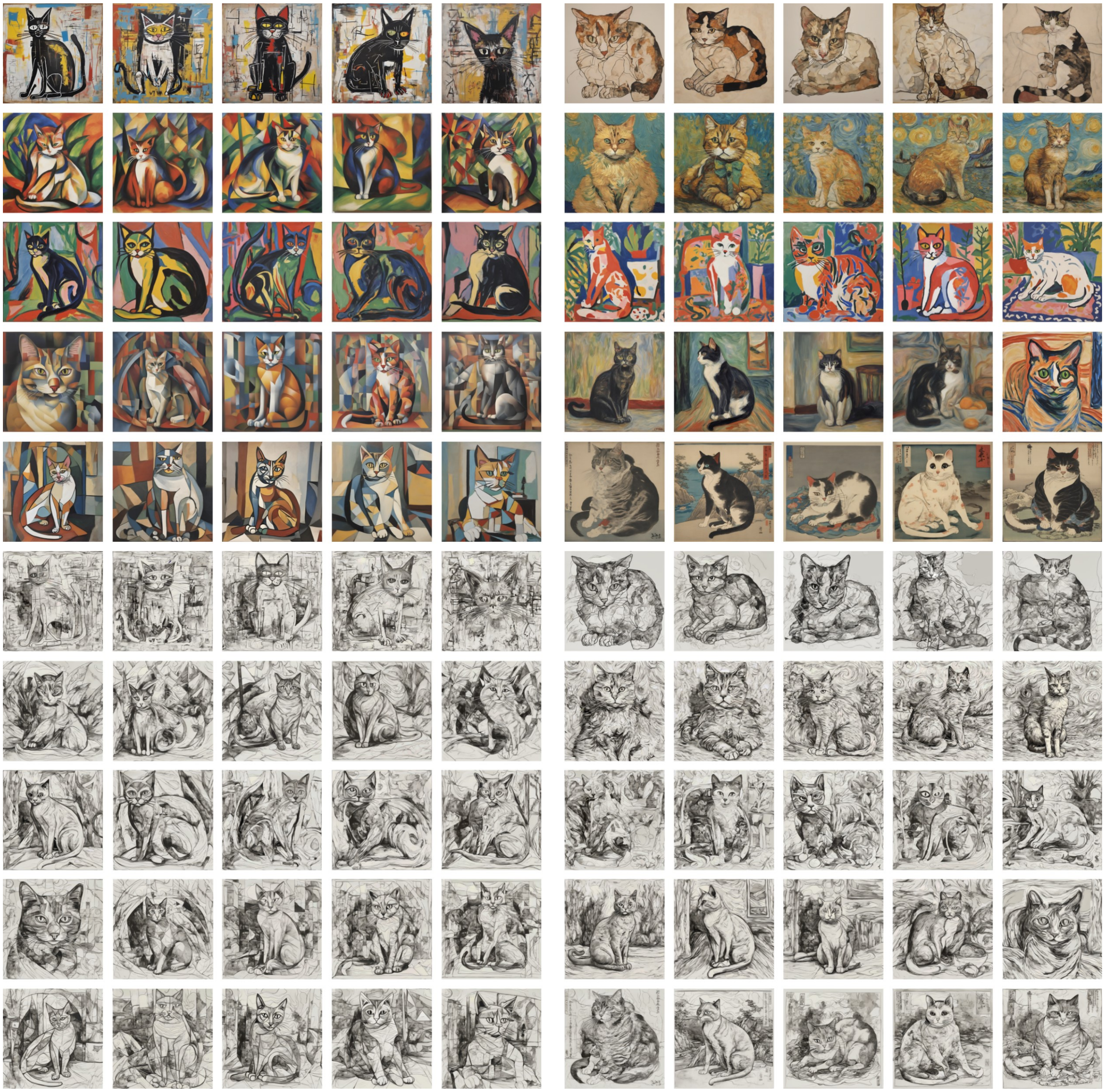


Figure 34. **Geometric Collection Example**. *An example of a paintings collection used for our geometric style sensitivity analysis. Please zoom in for a better view.*

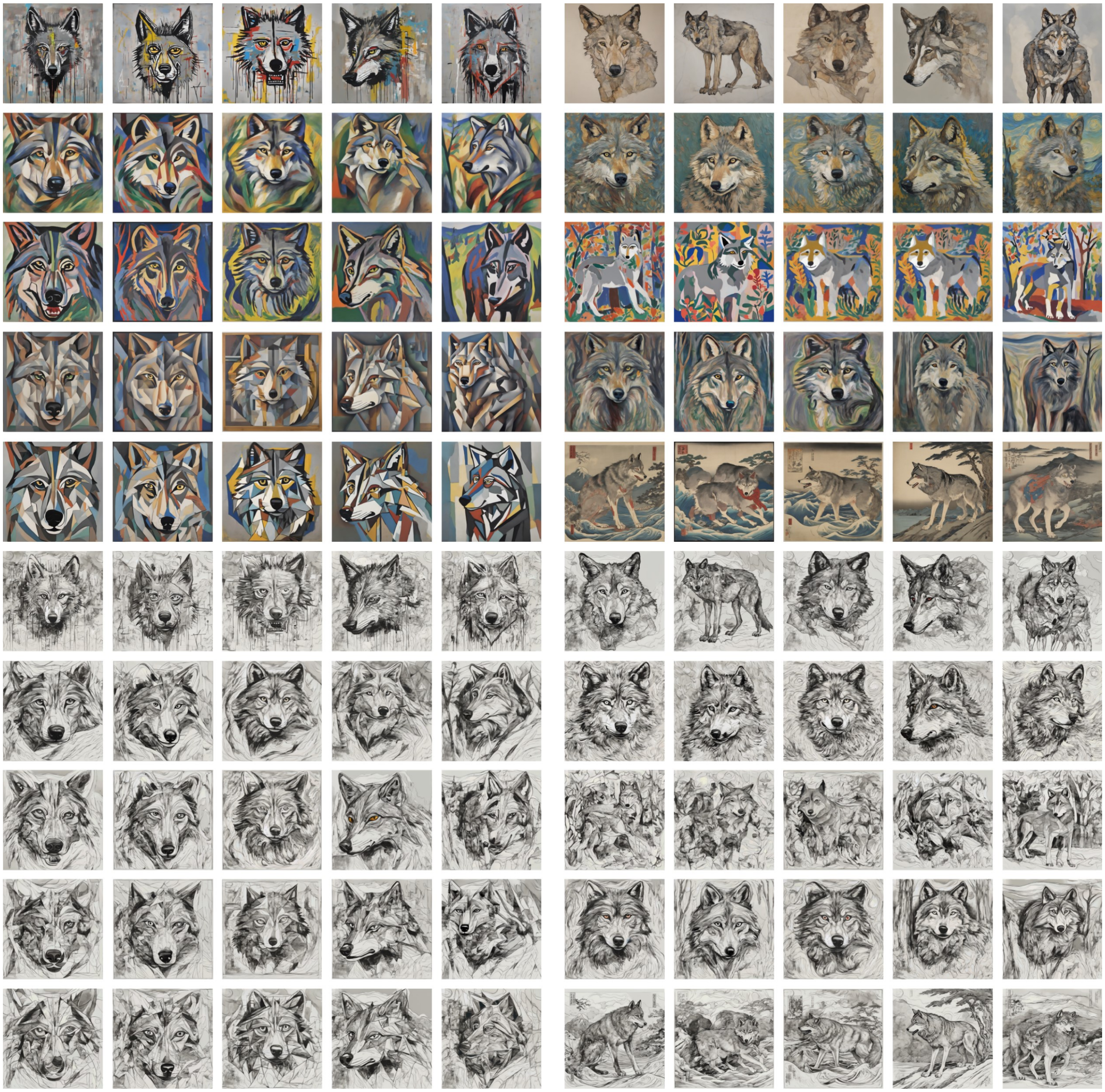
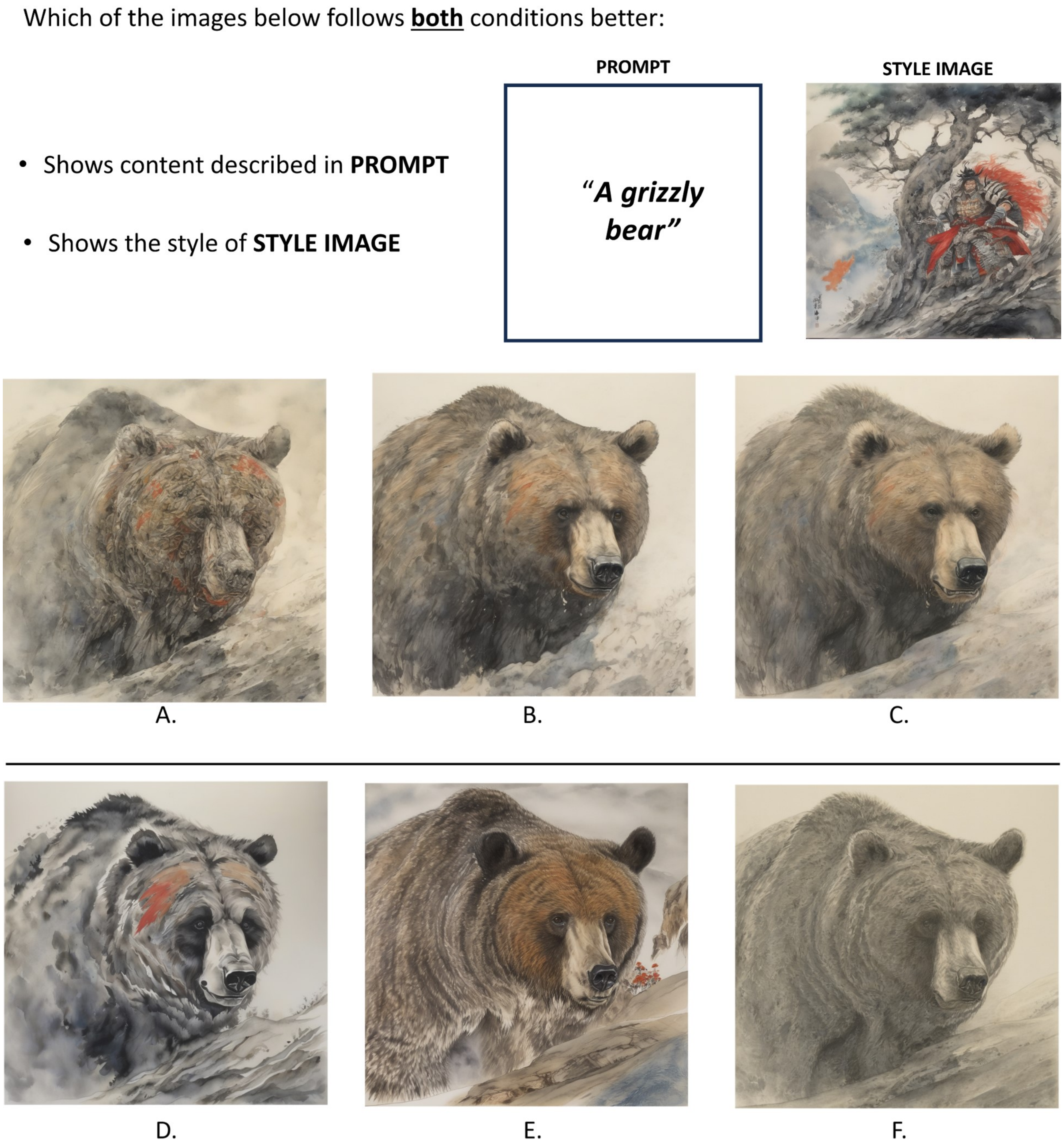


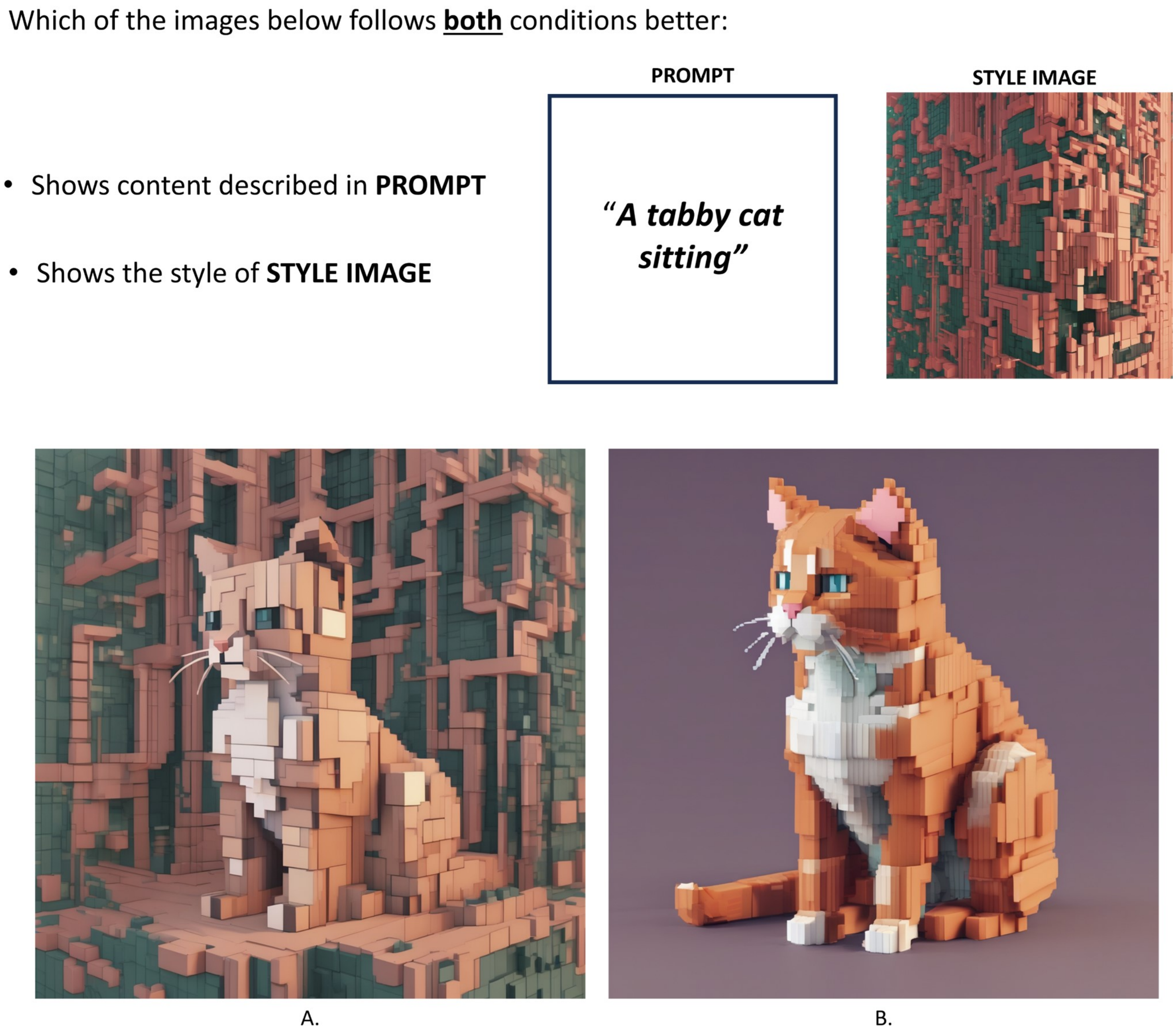
Figure 35. **Geometric Collection Example**. *An example of a paintings collection used for our geometric style sensitivity analysis. Please zoom in for a better view.*



Which of the images below follows **both** conditions better:

- Shows content described in **PROMPT**
- Shows the style of **STYLE IMAGE**

Figure 36. **User Study - Multiple Choice Questions.** *A sample of a multiple choice question from the user study.*



Which of the images below follows **both** conditions better:

- Shows content described in **PROMPT**
- Shows the style of **STYLE IMAGE**

PROMPT

“A tabby cat sitting”

STYLE IMAGE

A.

B.

Figure 37. **User Study - A/B choice Questions.**. *A sample of an A/B choice question from the user study.*