Stroke2Sketch: Harnessing Stroke Attributes for Training-Free Sketch Generation

Supplementary Material

A. Analysis and ablation

A.1. Stroke stylization

One of the main challenges in sketch extraction is how to transfer stroke attributes from a reference sketch to reconstruct the content image's sketch. As discussed in the main paper's related work section, previous approaches often rely on algorithmic simulations to emulate specific stroke styles. However, the vast diversity of sketch styles in real-world references makes it impractical to enumerate and simulate all possible styles algorithmically.

Our proposed approach introduces a novel solution by leveraging key-value (K-V) exchanges in attention mechanisms to transfer stroke attributes. This method allows dynamic adaptation of reference stroke properties to the content sketch during the generation process. However, as shown in the third column of Fig. 13 (a), direct K-V exchanges can sometimes distort structural elements, such as curves, leading to incomplete or misaligned strokes.

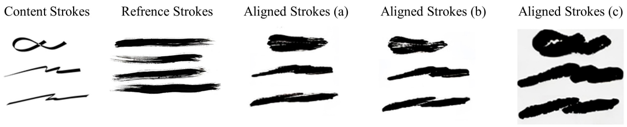


Figure 13. Stroke alignment results. The first two columns show the content strokes and reference strokes, respectively. Column (a) displays results with direct K-V exchanges, showing partial curve distortion. Columns (b) and (c) show improvements using contour guidance and stroke details propagation enhancement, respectively, highlighting the balance between stroke consistency and content preservation.

To address these limitations, we integrate contour guidance and the SDPE module into the generation process. These enhancements enable the system to retain structural integrity while achieving stroke style consistency. As demonstrated in Fig. 13, column (b) shows results with contour guidance applied, which helps preserve critical outlines while aligning strokes. Column (c) illustrates the output with both contour guidance and SDPE, achieving a balance between stroke stylization and content preservation.

While these methods improve stroke consistency, they can occasionally compromise the semantic expression of the content. To mitigate this, we introduce user-adjustable parameters, allowing users to fine-tune the balance between style fidelity and content preservation based on specific application requirements. In the following section, we detail the default parameters used in our experiments and provide the rationale for their selection.

A.2. Experimental configuration

We operate using the Stable Diffusion v2.1-base model[*] [32], leveraging DDPM inversion [16] for input image inversion and the DDIM scheduler for denoising over 50 steps. Following [1], cross-image attention layers are employed at specific resolutions (32×32 and 64×64) during denoising, enhancing stroke injection. The injection timesteps and additional settings are summarized in Tab. 3. Further, object prompts are extracted using BLIP-2[†] [20], and contour detection is performed using TEED [37] and U2-Net[‡]. To ensure semantic segmentation, the unsupervised self-segmentation technique from [30] is applied.

Hyperparameter	Value/Methodology
Model	Stable Diffusion v2.1-base[*]
Inversion	DDPM inversion [16]
Denoising Scheduler	DDIM, 100 steps (30 steps skip)
Resolution for SFI	32×32 (steps 10–70) 64×64 (steps 10–90)
Contrast Strength	$\zeta = 1.67$
Contour Mask	U2-Net[‡]
Contour Detection	TEED [37]
Guidance Scales	$\beta_{sg} = 5, \beta_{text} = 0.1$ (steps 20–100)
Self-Segmentation	Patashnik et al. [30]
Contour Guidance	$\gamma = 0.25$
Prompt Extraction	BLIP-2[†] [20]
Device	CUDA NVIDIA RTX 3090
Seed	42

Table 3. Hyperparameter settings for Stroke2Sketch experiments.

A.3. Ablation study analysis

As discussed in Sec. 4.4 of the main paper, we performed ablation studies to validate the contributions of the DAM, SPM, and SDPE. Quantitative results in Tab. 2 and qualitative comparisons in Fig. 12 demonstrate the critical roles of these components in achieving high-quality sketch generation.

Removing any component results in significant performance degradation, as reflected in both metrics and visual outputs:

Configuration B: Without DAM. Removing DAM results in ArtFID increasing from 32.45 to 38.67 and FID

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

[†] https://huggingface.co/docs/transformers/main/model_doc/blip-2

[‡] https://github.com/xuebinqin/U-2-Net

increasing from 22.43 to 26.53, indicating weaker style-content alignment and semantic consistency. LPIPS worsens to 0.672, highlighting the loss of content fidelity. Visually, as shown in Fig. 12, the absence of DAM causes noticeable content leakage, leading to inconsistent stroke thickness and blurred object boundaries. For example, the foreground details, such as facial contours and clothing edges, become misaligned, disrupting the overall semantic clarity.

Configuration C: Without SPM. Without SPM, ArtFID increases to 36.89, FID worsens to 30.47, and LPIPS rises to 0.637, reflecting reduced semantic alignment. Fig. 12 shows that this configuration struggles to preserve high-level abstractions, with many fine details either omitted or misplaced. For instance, the strokes in object outlines lose coherence, and elements such as eyes or limbs become poorly defined. This highlights the importance of SPM in maintaining semantic coherence and ensuring structural integrity.

Configuration D: Without SDPE. The removal of SDPE leads to the most significant degradation, with ArtFID increasing to 40.53 and FID and LPIPS scores worsening to 32.44 and 0.598, respectively. Visually, Fig. 12 reveals that sketches become overly coarse and noisy, with significant background interference and a lack of refinement in stroke details. For example, small textures and edges appear cluttered, reducing the clarity and aesthetic quality of the sketch. SDPE is essential for refining fine-grained details and suppressing noise propagation.

Configuration A: Full Method. The full method achieves the best performance, with ArtFID, FID, and LPIPS scores of 32.45, 22.43, and 0.530, respectively. Qualitatively, as seen in Fig. 12, this configuration produces sketches that closely align with the reference stroke style while preserving the semantic structure of the content. Fine details, such as facial features and object edges, are rendered with high precision, demonstrating the effectiveness of integrating DAM, SPM, and SDPE.

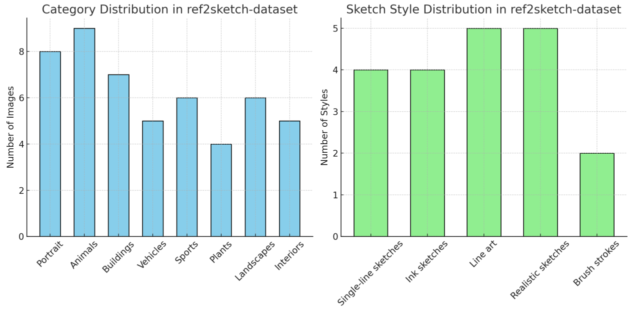


Figure 14. Overview of the Stroke2Sketch-dataset: Left - category distribution; Right - sketch style distribution. Zoom in to view details.

A.4. Hyperparameter effects

We demonstrate in Fig. 15, Fig. 16, and Fig. 17 how varying the hyperparameters γ, β_{sg}, and ζ provides users with greater control over the sketch generation process. These parameters influence the balance between style fidelity, content preservation, and abstraction, enabling customization based on specific user needs. Observing the results across various sketches, we note the interplay of these parameters with the pretrained diffusion model priors and the initial contour extraction quality.

Effect of γ (Contour weight): The parameter γ determines the influence of content image contours on the final sketch. As shown in Fig. 15, increasing γ results in sketches with more pronounced alignment to the original content structure, improving realism. For example, at $\gamma = 0.25$ (our default setting), the contours are well-preserved while maintaining the reference stroke style. However, higher values of γ (e.g., $\gamma = 0.6$) lead to excessive adherence to the content outline, compromising the transfer of stylistic features. Conversely, very low values (e.g., $\gamma = 0.15$) result in sketches with diminished structural coherence, favoring abstraction.

Effect of β_{sg} (Stroke guidance scale): The parameter β_{sg} controls the weight of stroke attributes transferred from the reference image. In Fig. 16, we observe that lower values of β_{sg} (e.g., $\beta_{sg} = 2$) yield sketches with reduced stylization, leaning more toward content fidelity. As β_{sg} increases, the reference stroke features become more prominent, with the optimal balance achieved at $\beta_{sg} = 5$. However, excessively high values (e.g., $\beta_{sg} = 15$) can lead to exaggerated stylization, overshadowing the content image's structural elements.

Effect of ζ (Contrast strength): The parameter ζ enhances contrast in the attention maps, aiding in stroke detail refinement. As shown in Fig. 17, low values of ζ (e.g., $\zeta = 0.8$) result in sketches with softer, less defined strokes. The default setting ($\zeta = 1.67$) provides a balanced output with clear stroke details and stylistic alignment. Increasing ζ beyond 3.5 introduces over-sharpening, leading to unnatural and overly rigid strokes.

Combined effects and user control: By varying these parameters in combination, users can control the degree of abstraction and stylization. For instance, increasing γ while decreasing β_{sg} emphasizes content realism, which is suitable for architectural sketches. In contrast, lowering γ and increasing β_{sg} enhances artistic abstraction, ideal for expressive line art. Default settings of $\zeta = 1.67$, $\gamma = 0.25$, and $\beta_{sg} = 5$ provide a general-purpose configuration that balances stroke style consistency with content preservation. Users can further refine these parameters based on their specific objectives.

B. Evaluation details

B.1. Stroke2Sketch-dataset

As described in the main paper, the Stroke2Sketch-dataset was created to assess the human perception of different sketch extraction methods. Fig. 14 provides a detailed visualization of the category distribution and sketch style diversity in the ref2sketch-dataset. This comprehensive dataset serves as a benchmark for evaluating both stylistic fidelity and semantic alignment in sketch generation tasks.

B.2. Baseline implementations

When comparing to alternative methods, we used the following implementations or demo websites:

- Ref2sketch: https://github.com/ref2sketch/ref2sketch
- Semi-ref2sketch: https://github.com/Chanuku/semi_ref2sketch_code
- Informative-drawings: https://github.com/carolineec/informative-drawings
- IP-Adapter: https://github.com/tencent-ailab/IP-Adapter
- InstantStyle: Huggingface demo https://huggingface.co/spaces/InstantX/InstantStyle
- InstantStyle-plus: https://github.com/instantX-research/InstantStyle-Plus
- CSGO: Huggingface demo https://huggingface.co/spaces/xingpng/CSGO
- RB-Modulation: Huggingface demo https://huggingface.co/spaces/fffiloni/RB-Modulation

B.3. Quantitative results on Stroke2Sketch-dataset

As shown in Tab. 1 in the main paper, our method achieves the lowest ArtFID and FID values among both training-based and training-free baselines, demonstrating its superiority in style fidelity and content preservation. Although our LPIPS value is slightly higher than Semi-ref2sketch [34], this discrepancy is expected due to the unique emphasis on stroke consistency in our approach. Notably, LPIPS, as a pixel-level similarity metric, does not fully capture the complexity of reference-based sketch extraction, where abstract artistic effects and semantic alignment are crucial. This limitation is evident in user evaluations, where our method consistently outperforms baselines, as detailed in Sec. 4.2 of the main paper.

Informative-drawings [5], designed to work with predefined styles, performs well on similar styles but lacks the flexibility to generalize to arbitrary reference sketches.

B.4. Quantitative results on FS2K dataset

In addition to the Stroke2Sketch-dataset, we evaluated our method on the FS2K dataset. Tab. 4 highlights our method's superior performance compared to specialized sketch extraction methods (Ref2sketch [2], Semi-ref2sketch [34]) and recent style transfer methods (StyleID [7]). Our method achieves the lowest FID (128.84) and LPIPS (0.4057) values, showcasing its robustness in producing high-quality sketches with strong semantic and stylistic fidelity.

While Ref2sketch and Semi-ref2sketch demonstrate reasonable performance due to their focus on training with paired data, they lack the flexibility to adapt to varied and abstract reference sketches. StyleID, although effective in style transfer tasks, struggles with precise alignment when handling content-specific sketches. In contrast, our approach leverages contour guidance and cross-image attention to preserve both structural details and stylistic nuances, ensuring high-quality results even in complex scenarios.

Methods	LPIPS	FID
Ref2sketch	0.5309	228.15
Semi-ref2sketch	0.4540	185.26
StyleID	0.5494	208.64
Ours	**0.4057**	**128.84**

Table 4. Quantitative results of comparison with baselines on FS2K dataset

B.5. Perceptual Study

Our user study interface (Fig. 18) displays the source content-reference pair as visual anchors alongside four anonymized stylized results in randomized layouts. Participants independently evaluated 20 unique image pairs, with each session limited to 5 minutes to ensure focused judgments. The interface incorporated a training phase showing prototypical examples of high/low content extraction and stroke quality before formal evaluation. We implemented quality control by tracking response times (excluding votes $< 3s$ as rushed) and adding attention-check questions. Detailed voting distributions per image pair and participant demographic profiles (85% with art-related backgrounds) are archived in the supplemental material.

C. Additional Results

As discussed in Sec. 4.1 of the main paper, we compare Stroke2Sketch with eight state-of-the-art methods that support both reference-based and text-based inputs, ensuring a fair evaluation of our approach. This design choice allows for a more equitable comparison, as models requiring only textual prompts or those designed for unrelated tasks (e.g., vector sketch generation or appearance transfer methods such as [1]) are fundamentally different in their objectives and are excluded from the subsequent visualizations.

Fig. 19 and Fig. 20 present additional comparison results across diverse styles and content images, demonstrating the robustness of our method. Meanwhile, Fig. 21 showcase sketches generated by Stroke2Sketch across different datasets, further validating its adaptability to varied styles and semantic requirements.

This focused evaluation highlights the advantages of our approach in achieving consistent stroke fidelity and semantic alignment while excluding comparisons with methods that do not align with the reference-based sketch extraction task.

D. Failure Cases

While our method demonstrates strong performance across a variety of reference styles, certain limitations remain when handling reference sketches with extreme characteristics. Specifically, sketches with overly simplistic or highly complex strokes pose challenges. As illustrated in Fig. 21, cases involving highly abstract continuous single-line references or densely detailed brushstroke references often result in suboptimal outcomes.

For instance, overly thick or abstract strokes can lead to detail loss or distortions in features like facial expressions, particularly in areas such as the eyes or intricate textures. Similarly, when the reference sketch exhibits densely packed details, the model may struggle to balance semantic consistency and stroke fidelity, resulting in either excessive abstraction or loss of critical content elements.

This behavior mimics how human artists adapt their interpretations based on the nature of the reference strokes. However, the challenge of fully decoupling semantic information from stroke attributes while maintaining both fidelity and style remains an open problem. Future work could explore advanced segmentation or attention mechanisms to address these limitations and enhance robustness in extreme cases.

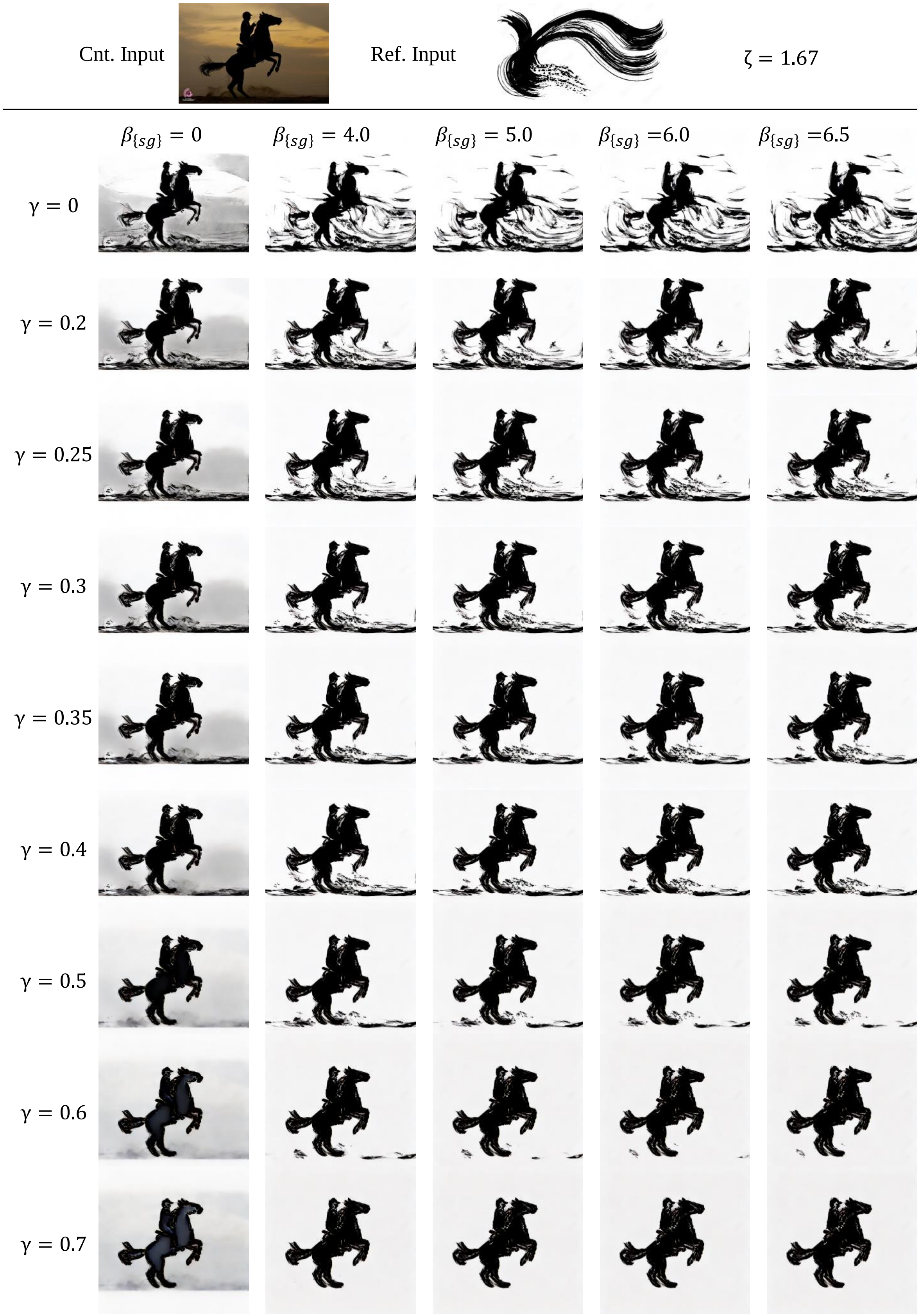


Figure 15. Visualization of γ variations. Increasing γ improves contour alignment but reduces stylistic abstraction. Default setting: $\gamma = 0.25$. Zoom in to view details.

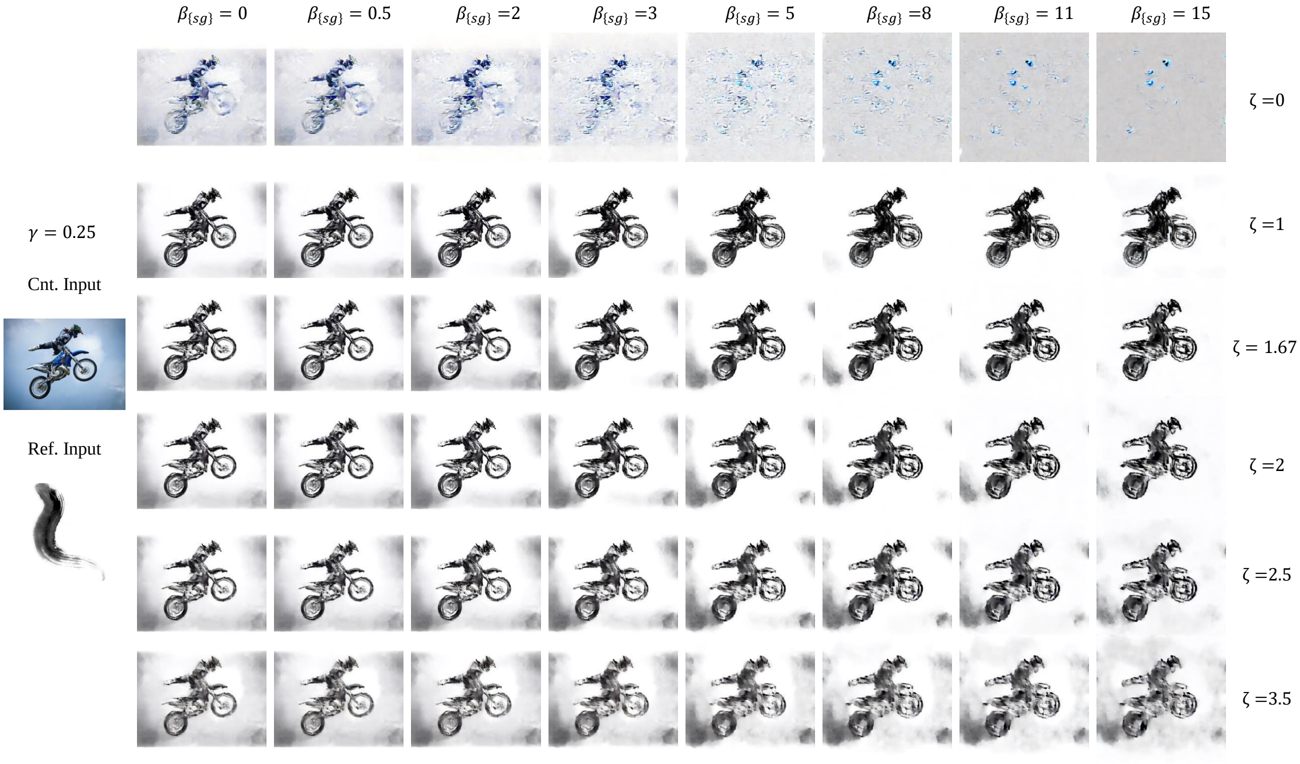


Figure 16. Visualization of β_{sg} variations. Higher β_{sg} emphasizes stroke attributes but may diminish content fidelity. Default setting: $\beta_{sg} = 5$. Zoom in to view details.

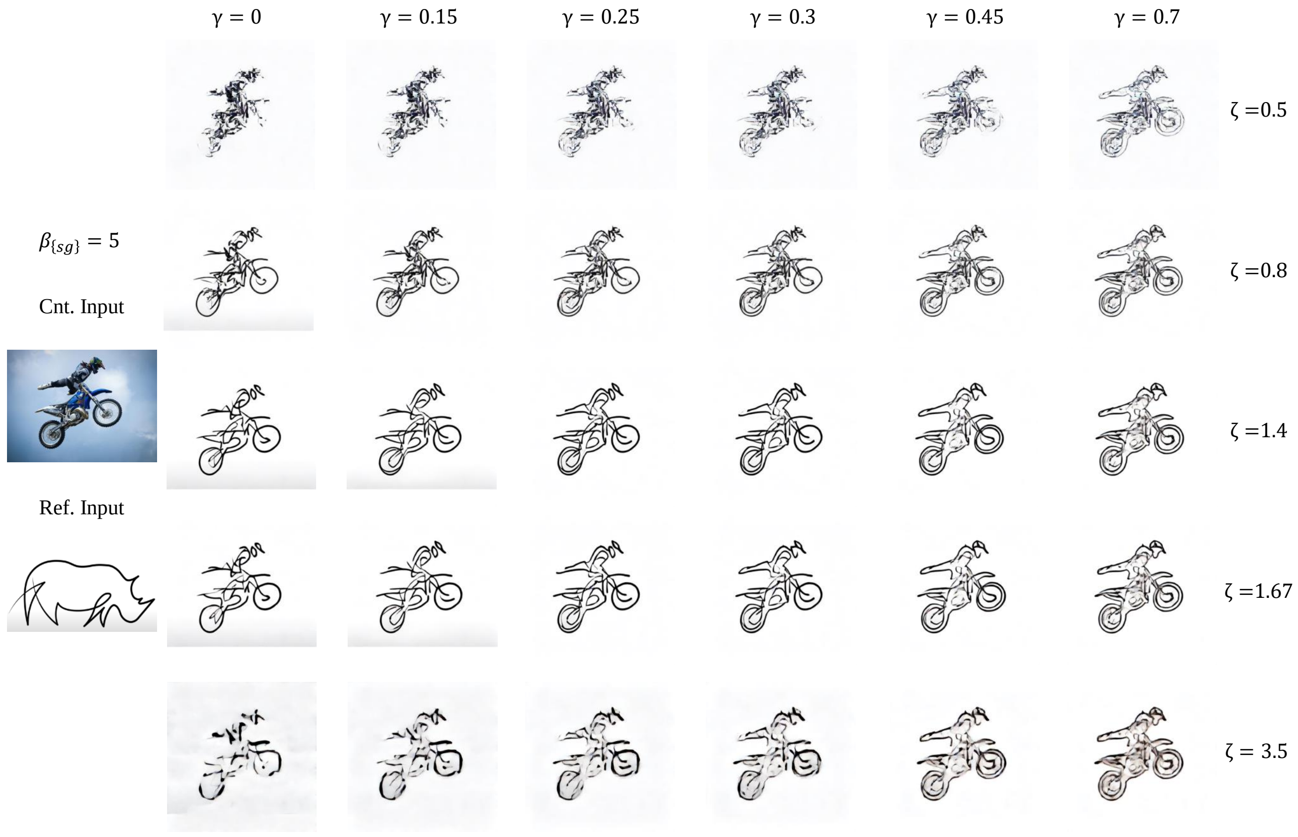
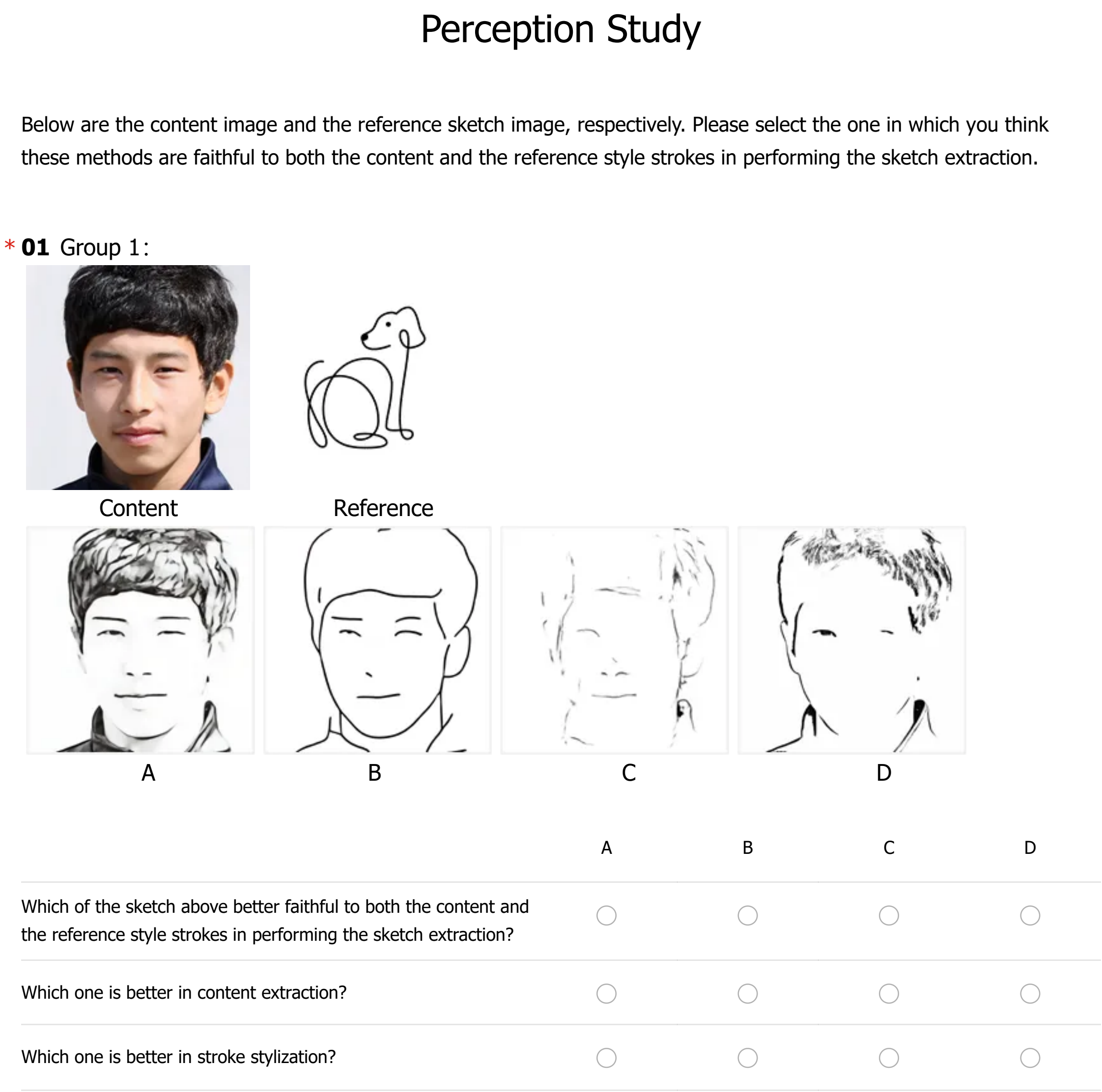


Figure 17. Visualization of ζ variations. Optimal contrast strength is achieved at $\zeta = 1.67$. Excessive ζ introduces over-sharpening effects. Zoom in to view details.



Perception Study

Below are the content image and the reference sketch image, respectively. Please select the one in which you think these methods are faithful to both the content and the reference style strokes in performing the sketch extraction.

* **01** Group 1:

Content

Reference

A

B

C

D

	A	B	C	D
Which of the sketch above better faithful to both the content and the reference style strokes in performing the sketch extraction?	○	○	○	○
Which one is better in content extraction?	○	○	○	○
Which one is better in stroke stylization?	○	○	○	○

Figure 18. Designed user study interface.

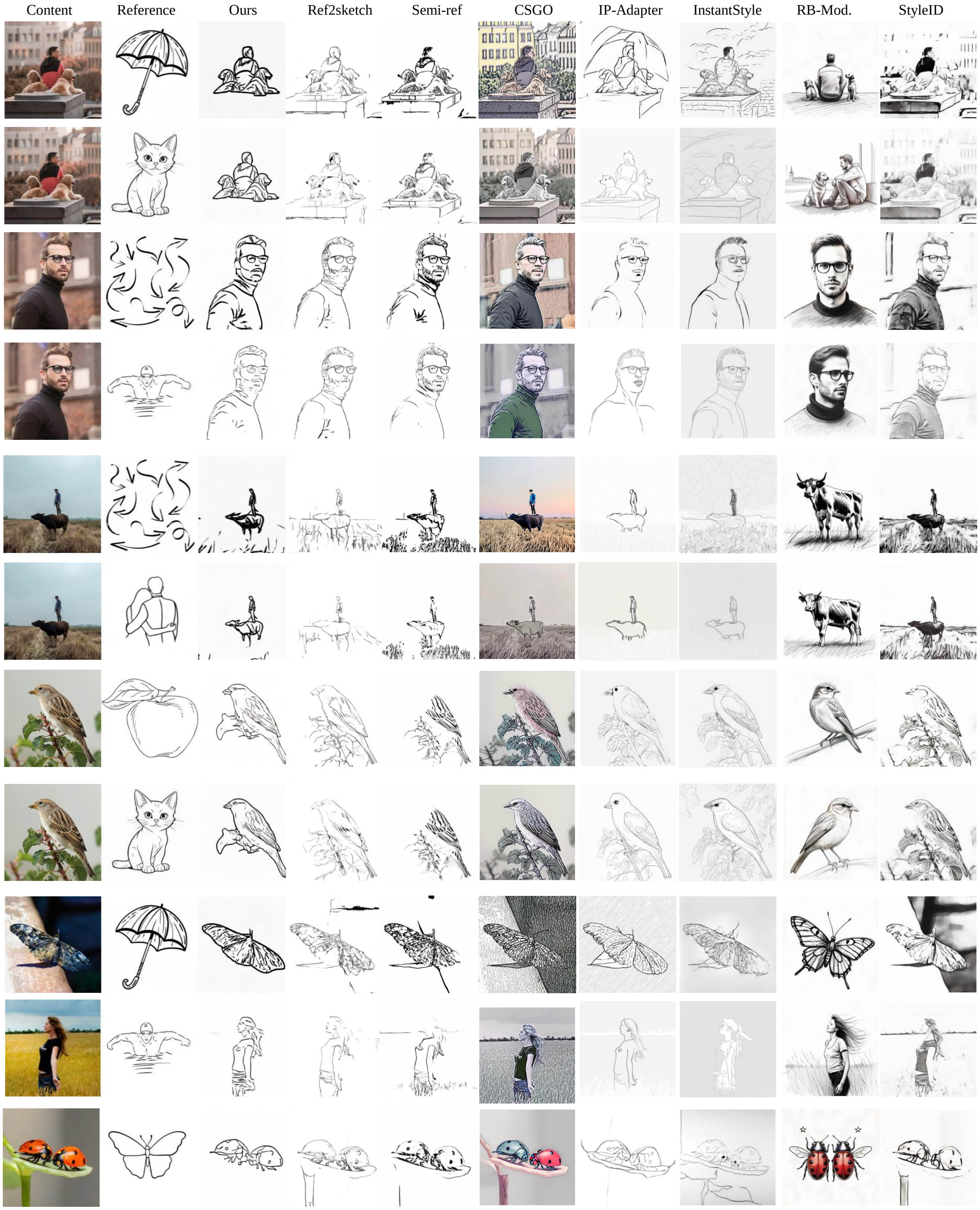


Figure 19. Comparison of sketches generated by Stroke2Sketch and baseline methods, including Ref2Sketch, Semi-ref2Sketch, CSGO, IP-Adapter, InstantStyle, RB-Modulation, and StyleID. Each row presents a content image, reference sketch, and results from different methods. Zoom in to view stroke details, highlighting the accurate alignment of stroke attributes and content semantics achieved by our approach.

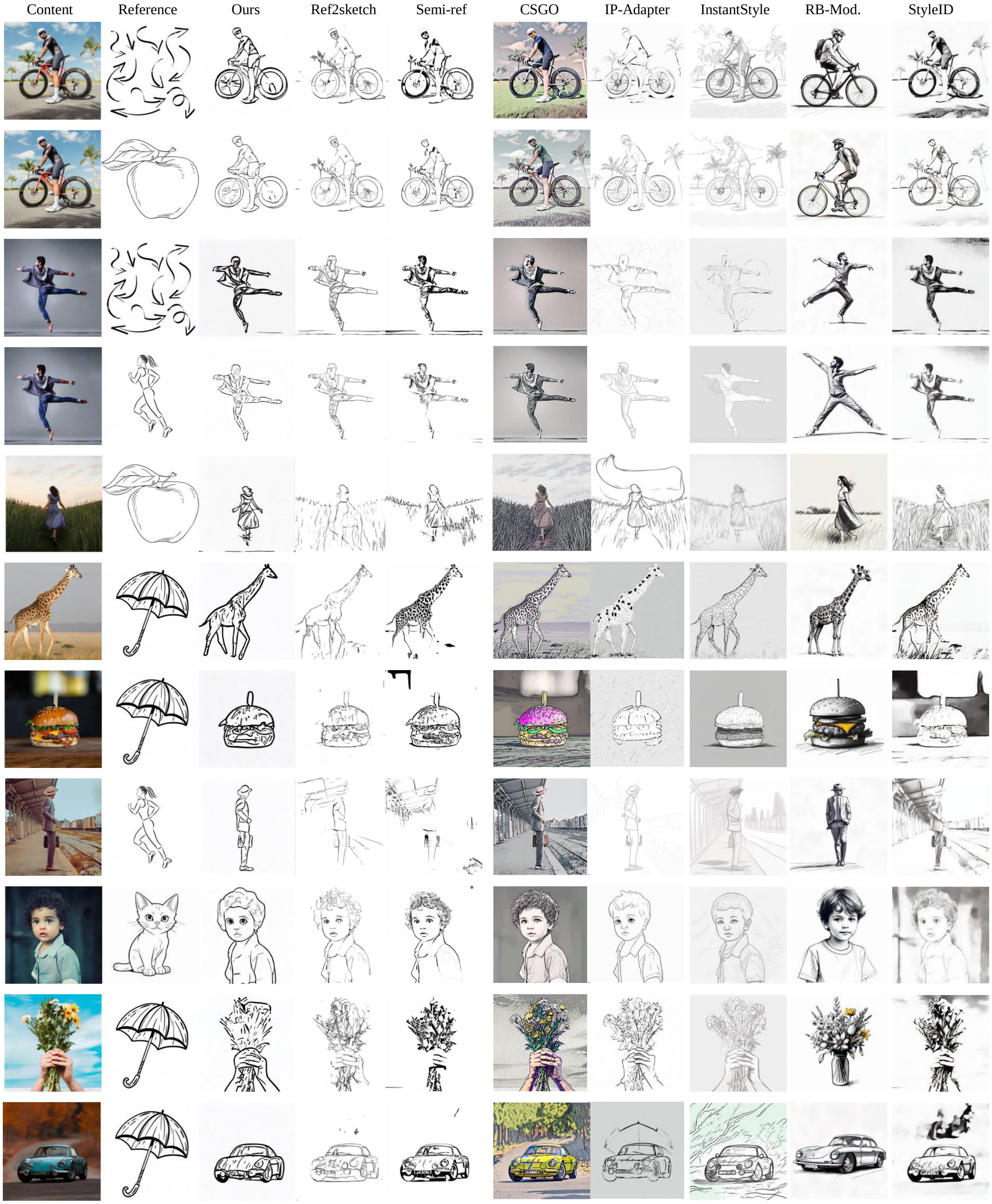


Figure 20. Additional qualitative comparison of Stroke2Sketch against baseline methods. The rows showcase content images, reference sketches, and outputs from various methods. Note the stroke details and style consistency in the results generated by our method. Zoom in to view stroke details for a clearer examination of stylistic fidelity and semantic alignment.

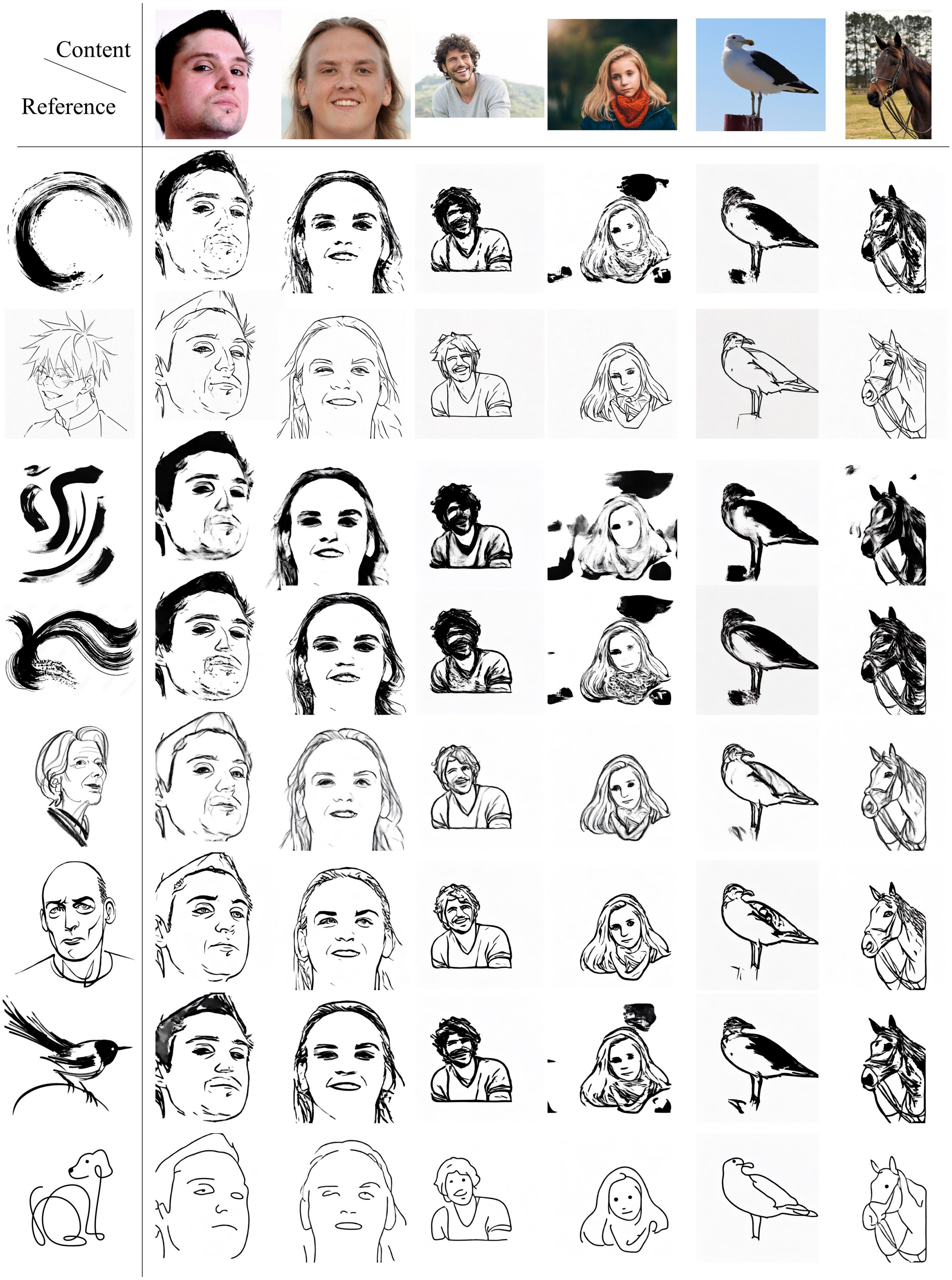


Figure 21. Comparison of sketch generation results using Stroke2Sketch across diverse content and reference styles. The first row shows the content images, and the second row provides the reference sketches representing varied stroke attributes. Rows 3-8 illustrate the generated sketches across a range of abstraction levels and stylistic alignments. Zoom in to view the nuanced details in stroke attributes.