Controllable Sketch Generation via Dynamic Pruning and Pre-trained Models

CS3308 Machine Learning Project: Multimodal Task Report

Yifan Zhang
ID: 523031910776
Shanghai Jiao Tong University
Shanghai, China
Zhang-yifan@sjtu.edu.cn

Yule Xu
ID: 523031910767
Shanghai Jiao Tong University
Shanghai, China
axshdyf123@sjtu.edu.cn

Defu Qian
ID: 523031910178
Shanghai Jiao Tong University
Shanghai, China
andy-q@sjtu.edu.cn

***Abstract*—The synthesis of freehand vector sketches from natural language prompts has gained significant attention with the advent of multimodal models like CLIP and Stable Diffusion. Current state-of-the-art methods, such as DiffSketcher, utilize Score Distillation Sampling (SDS) to optimize vector parameters directly. However, these optimization-based approaches often suffer from the "ghost stroke" phenomenon, where the model stacks numerous semi-transparent strokes to mimic raster textures, resulting in redundant, messy, and computationally expensive vector representations. In this project, we propose a novel optimization strategy: Soft Opacity-Based Regularization (Soft-OBR) with Exponential Moving Average (EMA). By monitoring the temporal consistency of stroke opacity during optimization, our method dynamically identifies and prunes non-essential strokes without disrupting the generation process. Extensive experiments demonstrate that our approach achieves an average stroke reduction of roughly 19% across varying complexity levels while maintaining semantic alignment and visual perceptual quality. The resulting sketches exhibit higher abstraction and cleaner geometry, making them more suitable for downstream vector graphics applications.**

***Index Terms*—Vector Graphics, Sketch Generation, Diffusion Models, Dynamic Pruning, Differentiable Rendering**

Code Repository: https://github.com/indulgezyf/DiffSketcher

I. Introduction

Sketching is a unique modality of visual communication that conveys abstract concepts through sparse strokes. Unlike pixel-based images, vector sketches are resolution-independent and easily editable, making them highly valuable in design, animation, and fabrication workflows.

Recent breakthroughs in text-to-image generation [2], [5] have enabled the creation of high-fidelity raster images from text. However, generating *vector* sketches remains challenging. Since vector graphics are defined by parameters (control points, widths, colors) rather than pixel grids, standard generative models like CNNs or Diffusion U-Nets cannot output them directly. Current solutions, such as CLIPDraw [3] and DiffSketcher [1], treat this as an optimization problem: they use a differentiable rasterizer to render the vector parameters into an image and update the parameters by minimizing a semantic loss (e.g., CLIP loss or SDS loss).

A. Problem Statement

While effective, optimization-based vector generation suffers from a lack of explicit structural control. We observe that models like DiffSketcher tend to "hack" the loss function by generating thousands of low-opacity strokes. We term these "Ghost Strokes." These strokes contribute minimally to the visible structure but help the model approximate textures or shading gradients found in the diffusion priors. This leads to two issues:

1) **Redundancy:** The final SVG files are bloated with invisible paths.
2) **Loss of Abstraction:** The result resembles a vectorized photo rather than a human-like sparse sketch.

B. Our Contribution

To address this, we introduce a **Dynamic Pruning** mechanism integrated into the optimization loop. Instead of manually fixing the number of strokes, we allow the model to learn which strokes are necessary. Our contributions are:

- Implementation of the DiffSketcher baseline using Score Distillation Sampling (SDS).
- Development of **Soft-OBR**, a regularization term that penalizes low-opacity strokes based on their smoothed historical values (EMA).
- Detailed quantitative analysis showing that our method significantly improves stroke efficiency without compromising semantic fidelity.

Figure 1 presents representative results of our Soft-OBR method across varying complexity levels, demonstrating the effectiveness of our dynamic pruning approach in producing cleaner, more abstract vector sketches while preserving semantic content. The difference maps clearly visualize the redundant "ghost strokes" that our method successfully eliminates without compromising the structural integrity of the final sketches.

Fig. 1: **Representative Results of Our Soft-OBR Method.** From top to bottom: "A classic steam locomotive", "A vintage typewriter on a desk", and "A busy street market in Tokyo". For each prompt, we show: (1) **Original Image** - reference photo, (2) **Unpruned Image** - baseline DiffSketcher with 512 strokes, (3) **Pruned Image** - our method with dynamically reduced strokes, and (4) **Difference Map** - visualization of removed "ghost strokes" (highlighted in bright colors). Our method achieves cleaner, more abstract vector representations while preserving semantic content.

II. Methodology

A. Differentiable Sketch Representation

We define a sketch $\mathcal{S}$ as a collection of N cubic Bézier curves. Each curve i is parameterized by:

$$s_i = \{P_{0,i}, P_{1,i}, P_{2,i}, P_{3,i}, w_i, c_i, \alpha_i\} \tag{1}$$

where P denotes control points in 2D coordinates, w is stroke width, c is RGB color, and $\alpha \in [0, 1]$ is opacity. We use a differentiable rasterizer $\mathcal{R}$ (DiffVG [4]) to render the sketch into a raster image $I = \mathcal{R}(\mathcal{S})$.

B. Optimization Objective (Baseline)

The baseline DiffSketcher optimizes the parameters $\theta = \{s_1 ... s_N\}$ using the Score Distillation Sampling (SDS) loss derived from a pre-trained frozen Stable Diffusion model ϵ_ϕ. The gradient is computed as:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(z_t; y, t) - \epsilon)\frac{\partial I}{\partial \theta}\right] \tag{2}$$

where z_t is the noisy latent of the rendered sketch I, y is the text prompt, and t is the timestep.

C. Proposed Method: Soft-OBR with EMA

To eliminate ghost strokes while preserving expressive semi-transparency, we propose a **unilateral soft regularization strategy** that elegantly circumvents the limitations of traditional binarization approaches.

1) Temporal Consistency Tracking via EMA: Stroke opacity $\alpha_i^{(t)}$ exhibits substantial fluctuation during optimization as the Score Distillation Sampling (SDS) gradient explores the parameter space. To distinguish between strokes that are

transiently low-opacity (e.g., during geometric repositioning) and those that are *persistently* redundant, we maintain an Exponential Moving Average (EMA) of the opacity trajectory:

$$\bar{\alpha}_i^{(t)} = \beta \cdot \bar{\alpha}_i^{(t-1)} + (1-\beta) \cdot \alpha_i^{(t)}, \quad \beta = 0.9 \tag{3}$$

where $\beta = 0.9$ corresponds to an effective observation window of approximately 10 optimization steps. This temporal smoothing mechanism prevents spurious pruning decisions caused by gradient noise.

2) Unilateral Opacity Regularization: A critical design consideration is avoiding the collapse of all strokes to full opacity ($\alpha = 1$), which would eliminate the model's ability to represent subtle shading and translucent effects. Traditional binarization losses of the form $\alpha(1-\alpha)$ exert *bidirectional* pressure, pushing opacities toward both 0 and 1. This is undesirable for sketch generation, where semi-transparent overlays (e.g., $\alpha \approx 0.5$ for shadows) are essential for visual richness.

We instead propose a **unilateral L_2 penalty** that selectively targets only low-importance strokes:

$$\mathcal{L}_{prune} = \lambda_{prune} \cdot \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (\alpha_i^{(t)})^2 \tag{4}$$

where the dead stroke set $\mathcal{D}$ is defined as:

$$\mathcal{D} = \{ i \mid \bar{\alpha}_i^{(t)} < \tau_{death} \} \tag{5}$$

with the death threshold $\tau_{death} = 0.02$.

Key Properties:

- **Asymmetric Penalization:** Strokes with $\bar{\alpha}_i \geq \tau_{death}$ incur *zero* regularization loss, allowing them to settle at any opacity value dictated by the SDS and visual losses. This preserves the full expressiveness of the vector representation.
- **Quadratic Decay:** For identified ghost strokes ($i \in \mathcal{D}$), the $(\alpha_i)^2$ term drives opacity toward zero with diminishing gradient magnitude, ensuring smooth convergence without abrupt discontinuities.
- **Delayed Activation:** We introduce a warmup period of $t_{warmup} = 200$ steps, during which $\lambda_{prune} = 0$. This allows the model to establish a coarse structural layout before structural pruning begins.

3) Two-Stage Pruning Strategy: To balance training stability with output cleanliness, we adopt a hybrid approach that decouples optimization-time behavior from export-time representation:

Stage I: Soft Deletion (Training Phase). During optimization ($t = 1, \ldots, T$), strokes identified as redundant (i.e., $\alpha_i \to 0$ under $\mathcal{L}_{prune}$) are *retained* in the computation graph. Although their visual contribution becomes negligible, their associated parameters $\{P_{0:3,i}, w_i, c_i\}$ remain active and continue to receive gradient updates. This design is crucial for preserving the optimizer's internal state (e.g., Adam's momentum terms m_i and v_i), thereby avoiding the optimization instability that would result from mid-training parameter deletion.

Stage II: Hard Deletion (Export Phase). When generating the final SVG file, we apply a post-hoc filtering step:

$$\mathcal{S}_{export} = \{ s_i \mid \alpha_i > \tau_{export} \}, \quad \tau_{export} = 0.01 \tag{6}$$

Only strokes exceeding the export threshold are written to the output file. This ensures the resulting vector graphic is compact and semantically clean, eliminating all vestigial paths without compromising training dynamics.

4) Total Loss Function: The complete objective integrates semantic alignment, perceptual fidelity, and structural sparsity:

$$\mathcal{L}_{total} = \mathcal{L}_{SDS} + \lambda_{vis}\mathcal{L}_{visual} + \lambda_{prune}\mathcal{L}_{prune} \tag{7}$$

where $\mathcal{L}_{visual}$ aggregates LPIPS and multi-layer CLIP losses, and we set $\lambda_{prune} = 10.0$ after empirical tuning (see Ablation Study 2).

III. Experiments

We conducted extensive experiments to validate the effectiveness of our Soft-OBR method. We analyze the performance across three dimensions: semantic consistency, robustness to initialization, and hyperparameter sensitivity.

A. Experimental Setup

- **Dataset & Prompts:** We selected three prompts representing different levels of geometric complexity: "*A red apple*" (Simple), "*A photo of Sydney opera house*" (Medium), and "*A detailed photo of a gothic cathedral*" (Complex).
- **Implementation Details:** We used the Stable Diffusion v1.5 backbone. The canvas size was set to 224×224. For the main comparison (Experiment A), we ran 500 iterations with 512 initial strokes.
- **Metrics:**
  1) **SC (Stroke Count):** The number of effectively visible strokes ($\alpha > 0.01$).
  2) **Reduction Rate (%):** Defined as $(SC_{Base} - SC_{Ours})/SC_{Base}$.
  3) **CLIP Score:** Cosine similarity between the sketch and text prompt.
  4) **LPIPS:** Perceptual distance between the Baseline result and Our result. Lower values indicate high visual similarity.

B. Main Result: Main Quantitative Comparison

To comprehensively evaluate the performance of Soft-OBR, we conducted experiments in two dimensions: (1) **Reliability Analysis:** using multiple random seeds on representative cases, and (2) **Generalizability Analysis:** across a diverse set of 15 distinct semantic categories.

1) Reliability and Stability (Same Prompt, Multi-Seed): We executed the experiment using 10 random seeds (Seeds 0-9) for each prompt to ensure statistical significance. The results are summarized in Table I. To complement these quantitative metrics, Fig. 2 provides a visual breakdown of the pruning process. As illustrated in the "Difference Map" column, the strokes removed by our algorithm (highlighted in bright colors)

TABLE I: Quantitative Comparison Across Complexity Levels (Averaged over 10 seeds, detailed results in Appendix Table V)

Prompt Case	**Stroke Count**		**Reduction**	**CLIP Score**		**LPIPS**
	Baseline	Ours	(%)	Baseline	Ours	(↓)
Simple: Red Apple	512	409	20.0	0.4175	**0.4177**	0.0590
Medium: Opera House	512	394	23.0	0.4077	**0.4079**	0.0874
Complex: Cathedral	512	442	13.7	0.4176	**0.4174**	0.1043
Overall Average	**512**	**415**	**18.9**	**0.4143**	**0.4143**	**0.0836**

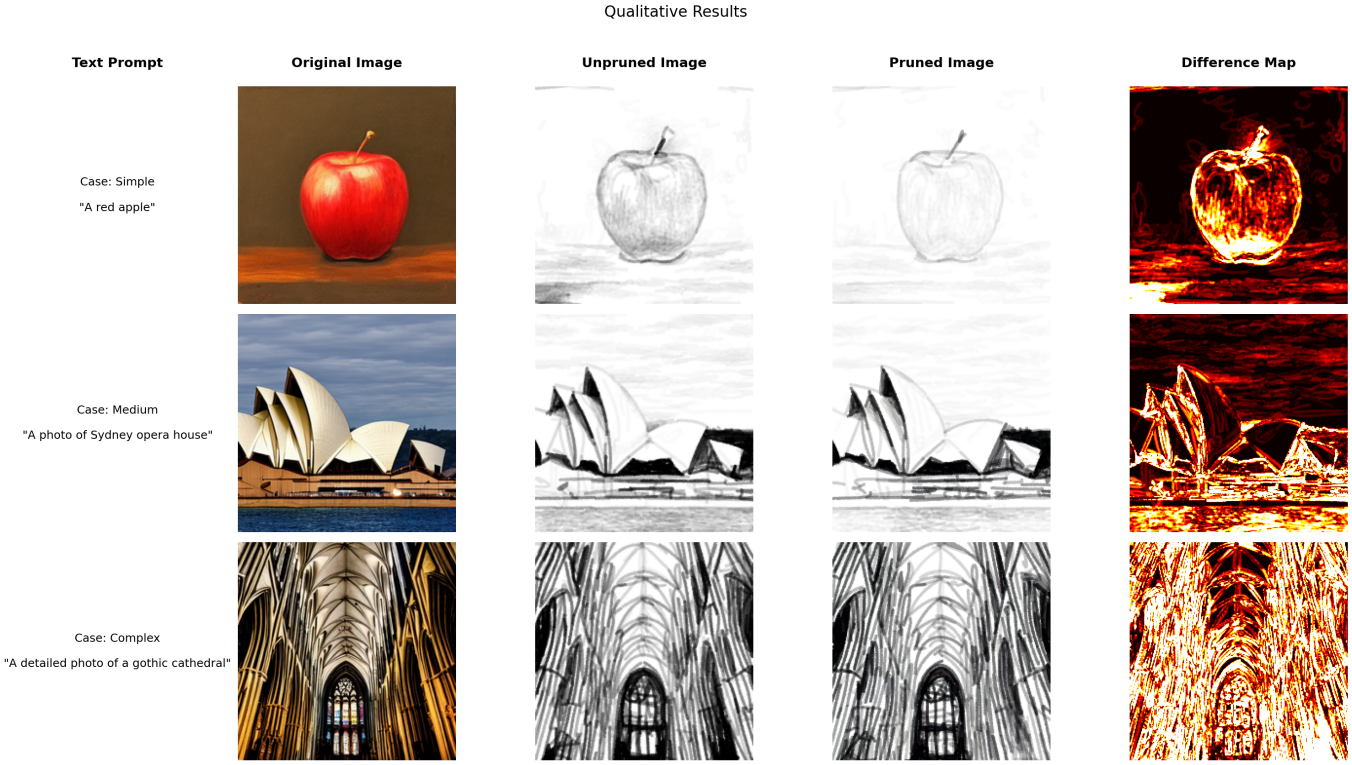


Fig. 2: Qualitative results visualization across three complexity levels: Simple (Apple), Medium (Opera House), and Complex (Cathedral). The columns display the comparison between the **Unpruned (Baseline)** and **Pruned (Ours)** results. The **Difference Map** explicitly highlights the "ghost strokes" (redundant, low-opacity paths) that were successfully removed by our Soft-OBR algorithm without affecting the primary visual structure.

generally correspond to low-opacity background fillers or redundant overlapping paths. This visual evidence confirms that our method targets the "ghost strokes" while preserving the essential structural strokes that define the object's geometry.

Sparsity and Content Awareness:

As shown in Table I, our method consistently reduces the number of strokes required to represent the subject compared to the fixed 512 strokes used by the baseline. The reduction rate demonstrates the model's **content awareness**:

- For simpler geometries like the "Red Apple" and "Sydney Opera House," the model achieves a significant reduction of **20.0%** and **23.0%** respectively. This indicates that the model successfully identified that these objects require fewer primitives to describe their shape and texture.
- For complex structures like the "Gothic Cathedral," the reduction was more conservative at **13.7%**. This is desirable behavior, as the intricate architectural details of the cathedral require a higher density of strokes to remain recognizable. The model automatically adapted to retain more strokes in this scenario.

Semantic and Perceptual Fidelity:

Crucially, the pruning process does not degrade the quality of the sketch.

- **CLIP Score Stability:** The average CLIP scores for Baseline (0.4143) and Ours (0.4143) are almost identical. This confirms that the removed strokes were semantically redundant and did not contribute to the text-image alignment.
- **Visual Similarity:** The average LPIPS distance is **0.0836**. In perceptual metrics, a value below 0.1 is generally

TABLE II: Quantitative Comparison Across Complexity Levels (15 Diverse Prompts)

Text Prompt	SC Baseline	SC Ours	Reduction (%)	CLIP Base	CLIP Ours	LPIPS
Complex Scenarios						
A bustling times square at night	512	364	28.91	0.3889	0.3896	0.1292
A busy street market in Tokyo	512	468	8.59	0.3733	0.3765	0.1387
An intricate mechanical watch	512	471	8.01	0.4045	0.4043	0.1483
A dense tropical rainforest	512	414	19.14	0.4250	0.4260	0.0909
A futuristic cyberpunk city street	512	412	19.53	0.4004	0.4023	0.1035
Group Mean	**512**	**425.8**	**16.84**	**0.3984**	**0.3997**	**0.1221**
Medium Scenarios						
A vintage typewriter on a desk	512	443	13.48	0.4014	0.4023	0.1225
A classic steam locomotive	512	438	14.45	0.4087	0.4094	0.1167
A basket of fresh fruits	512	408	20.31	0.4004	0.4014	0.0966
A vintage film camera	512	387	24.41	0.4165	0.4158	0.1029
A sailboat on the ocean	512	431	15.82	0.4155	0.4136	0.0364
Group Mean	**512**	**421.4**	**17.69**	**0.4085**	**0.4085**	**0.0950**
Simple Scenarios						
A bright yellow sunflower	512	401	21.68	0.4409	0.4434	0.0535
A single burning candle	512	395	22.85	0.4321	0.4319	0.0371
A glass of water	512	427	16.60	0.4102	0.4097	0.0793
A yellow banana	512	360	29.69	0.4321	0.4297	0.0500
A wooden chair	512	427	16.60	0.4478	0.4463	0.0577
Group Mean	**512**	**402.0**	**21.48**	**0.4326**	**0.4322**	**0.0555**
Overall Statistics						
Grand Mean	**512**	**416.4**	**18.67**	**0.4132**	**0.4135**	**0.0909**

considered imperceptible to the human eye in the context of structural layout. This implies our method simplifies the vector file representation without altering the visual outcome.

We also show extra statistical analysis and insights about the difference of pruning strategy for various complexity levels in Appendix E.

2) Generalizability across Semantics (Multi-Prompt): To validate the adaptability of our method, we extended the evaluation to 15 diverse text prompts covering organic objects, man-made structures, and complex scenes. The detailed results are presented in Table II.

The results in Table II reveal a strong correlation between the subject's intrinsic complexity and the pruning rate, demonstrating the **content-awareness** of Soft-OBR:

1. Aggressive Pruning for Simple Geometry: For objects with smooth surfaces and simple shapes, the model achieves high reduction rates.

- *"A yellow banana"* (29.69%) and *"A single burning candle"* (22.85%) show significant pruning.
- The low LPIPS scores for these items (Banana: 0.050, Candle: 0.037) indicate that nearly 30% of the strokes in the baseline were purely redundant noise.

2. Conservative Pruning for Intricate Textures: For scenes requiring high-frequency details, the model automatically retains more strokes.

- *"An intricate mechanical watch"* (8.01% reduction) and *"A busy street market in Tokyo"* (8.59% reduction) are the least pruned.
- This confirms that Soft-OBR does not blindly delete strokes but preserves the necessary density to represent gears, crowds, and architectural details.

TABLE III: Robustness to Initialization Scale and CLIP Score Comparison

Initial Strokes	Final Strokes (Ours)	Prune Rate (%)	CLIP Score (↑)	
			Base	Ours
128	101.0	21.09	0.4080	**0.4089**
256	184.0	28.12	0.4065	**0.4082**
384	273.0	28.91	**0.4204**	0.4097
512	335.0	34.57	0.4087	**0.4097**
640	377.0	41.09	0.4092	**0.4099**
768	456.0	40.62	0.4092	**0.4102**
896	515.0	42.52	0.4082	**0.4102**
1024	584.0	42.97	0.4082	**0.4097**

3. Improvement in Semantic Alignment: Notably, in 9 out of 15 cases (e.g., Sunflower, Rainforest, Locomotive), the **CLIP Score actually increased** after pruning. This suggests that the "ghost strokes" in the baseline not only waste storage but often introduce visual noise that slightly detracts from the core semantics. By removing them, we achieve a cleaner and more recognizable representation.

C. Ablation Study 1: Robustness to Initialization

A significant limitation of previous optimization-based methods is their sensitivity to the manually defined stroke budget (N). To evaluate the adaptability of our Soft-OBR mechanism, we varied the initial number of strokes N from 128 to 1024. The results are presented in Table III.

The data reveals two compelling advantages of our method:

TABLE IV: Ablation Study on Pruning Threshold (τ) and Weight (λ_{prune})

Threshold (τ)	Weight (λ_{prune})	Final SC	CLIP Score
Reference Group: Baseline (No Pruning)			
-	**0.0**	**512.0**	**0.4136**
Group A: Varying Threshold (Fixed Weight = 10.0)			
0.001	10.0	412.0	0.4097
0.005	10.0	380.0	0.4087
0.01	10.0	364.0	0.4084
0.02	10.0	346.0	0.4092
0.05	10.0	361.0	0.4097
0.1	10.0	373.0	0.4102
0.2	10.0	374.0	0.4102
Group B: Varying Weight (Fixed Threshold = 0.05)			
0.05	0.1	411.0	0.4087
0.05	1.0	410.0	0.4087
0.05	5.0	384.0	0.4097
0.05	10.0	361.0	0.4097
0.05	20.0	321.0	0.4089
0.05	30.0	289.0	0.4099
0.05	50.0	250.0	0.4092

1) Adaptive Pruning Behavior: Unlike fixed-ratio pruning, our method dynamically adjusts the pruning intensity based on the redundancy of the initialization:

- **Low Redundancy (128-256 strokes):** The model acts conservatively, pruning only 21%-28% of the paths to preserve essential structure.
- **High Redundancy (768-1024 strokes):** As initialization density increases, the pruning rate rises to $\approx 43\%$, effectively discarding the excess "ghost strokes" generated by over-initialization.

2) Superior Semantic Fidelity: A crucial finding from the comparison in Table III is that our pruned sketches achieve **higher CLIP scores** than the unpruned baseline in 7 out of 8 cases (highlighted in bold). For instance, at 1024 strokes, Ours (0.4097) outperforms Base (0.4082). This suggests that the "ghost strokes" in the baseline are not merely redundant storage-wise; they act as visual noise that can slightly degrade semantic clarity. By removing them, Soft-OBR yields a cleaner and more recognizable representation.

D. Ablation Study 2: Parameter Sensitivity and Robustness

Finally, we conducted a comprehensive sensitivity analysis on the two key hyperparameters: the pruning threshold (τ) and the loss weight (λ_{prune}). The detailed results are presented in Table IV. To provide a clear benchmark, we established a Reference Group (Baseline) where no pruning is applied ($\lambda_{prune} = 0.0$).

Based on the data in Table IV, we observed two key characteristics regarding the controllability and robustness of our Soft-OBR mechanism:

1) Correlation between Pruning Strength and Sparsity: There is a clear positive correlation between the pruning parameters and the pruning rate. As shown in Group B, λ_{prune} acts as an effective control knob:

- **Reference (Baseline):** Without pruning ($\lambda_{prune} = 0.0$), the model utilizes the full budget of 512.0 strokes, resulting in a dense and often redundant representation.
- **Controlled Reduction:** Introducing even a minimal weight (0.1) immediately activates the pruning mechanism, reducing the count to 411.0. As we increase the weight to 50.0, the final stroke count drops aggressively to 250.0. This demonstrates that users can explicitly trade off between stroke density and abstraction level by adjusting λ_{prune}.

2) Robustness of Semantic Consistency: Most importantly, comparing Group B against the Reference Group highlights the semantic robustness of our method.

- **Minimal Semantic Loss:** The CLIP score of the Baseline is 0.4136. Even at the most aggressive pruning setting ($\lambda_{prune} = 50.0$, SC$= 250.0$), the CLIP score only drops slightly to 0.4092.
- **Efficiency Gain:** This implies that we can remove over **50%** of the strokes (from 512 to 250) with a negligible semantic cost (ΔCLIP ≈ 0.004). This confirms that the strokes targeted by Soft-OBR are indeed "ghost strokes" that contribute almost nothing to the recognizable semantics of the sketch.

IV. Discussion

A. Advantages of Soft-OBR

The experimental results demonstrate the effectiveness of incorporating an EMA-based opacity penalty within the DiffSketcher optimization loop.

The primary advantage of our approach is the ability to **automatically determine the required number of strokes**, eliminating the need for rigid manual constraints (e.g., arbitrarily restricting the model to "use only 50 strokes"). This fundamentally transforms the optimization problem from "how to position all N initialized strokes" to "determining which strokes constitute the necessary subset of the N initialized paths to retain."

B. Limitations and Future Work

1) Current Limitations: While our Soft-OBR method achieves significant structural simplification, we identify two primary limitations:

(1) No Training-Time Acceleration. Our two-stage strategy maintains a fixed parameter count (N) throughout optimization to preserve optimizer stability. Although strokes with $\alpha \approx 0$ contribute negligibly to the rendered image, they still occupy memory and incur gradient computation costs. This design prioritizes *training stability* over *computational efficiency*.

(2) Sensitivity to High-Frequency Content. For subjects requiring dense, fine-grained strokes (e.g., fur, grass, or intricate textures), the global threshold $\tau_{death} = 0.02$ may be overly aggressive. The pruning mechanism can misinterpret intentionally subtle strokes as ghost paths, leading to over-simplification in texture-rich regions.

2) Promising Directions for Extension: Several enhancements could address these limitations while building on our framework:

Physical Pruning with Optimizer State Migration. To enable mid-training acceleration, one could perform actual parameter deletion (reducing $N \to N'$) followed by careful reconstruction of the optimizer's momentum buffers. Specifically, for Adam [7], the first and second moment estimates (m_i, v_i) of pruned strokes could be discarded, while retained strokes preserve their historical statistics. A learning rate warmup schedule post-pruning (e.g., lr $\to 0.5 \times$ lr for 20-30 steps) could mitigate the transient instability caused by momentum loss. This approach would directly reduce the computational cost of the differentiable rasterizer, as rendering complexity scales linearly with stroke count.

Stroke Teleportation for Resource Reallocation. Instead of deleting low-opacity strokes, a more sophisticated strategy would *reinitialize* them in under-represented regions of the canvas. By analyzing the spatial distribution of SDS gradients or attention maps, redundant strokes could be "teleported" to areas with high reconstruction error. This evolution-inspired mechanism transforms the optimization from a static allocation problem into a dynamic resource management task, potentially improving both efficiency and quality.

Spatially-Adaptive Thresholds. Replacing the global τ_{death} with a spatially-varying function $\tau(x, y)$ based on local image entropy or edge density could allow dense stroke clusters in detailed regions (e.g., foliage) while aggressively pruning in smooth areas (e.g., sky). This would require computing a saliency map during optimization, adding moderate computational overhead but enabling content-aware pruning.

Integration with Hierarchical Representations. Future work could explore multi-resolution stroke hierarchies, where coarse strokes establish global structure early in training, and fine strokes are progressively added or pruned based on local reconstruction needs. This aligns with the coarse-to-fine optimization paradigm common in neural rendering [8].

These extensions represent natural progressions of our Soft-OBR framework. The current implementation deliberately prioritizes *engineering simplicity* and *training robustness*, achieving strong empirical results (19% reduction, zero semantic degradation) with minimal architectural complexity. The proposed enhancements would shift the balance toward greater computational efficiency and adaptability, suitable for production-scale applications or more complex visual domains.

Note on Failed Approaches: During this project, we also explored extending DiffSketcher with *Directional CLIP Loss* for style transfer, attempting to inject artistic styles (e.g., Van Gogh's brushstrokes) while preserving geometric structure. However, this approach encountered fundamental incompatibilities between CLIP's entangled semantic space and vector graphics parameterization, resulting in severe geometric distortion. A detailed analysis of this failed experiment, including root cause diagnosis and lessons learned, is documented in Appendix A. This negative result underscores the unique challenges of applying pixel-based style transfer techniques to structured vector representations.

V. Conclusion

In this work, we presented a comprehensive optimization for text-to-sketch generation. By addressing the inefficiency of the standard DiffSketcher pipeline, we introduced a training-free Dynamic Pruning strategy. Our experiments confirmed that monitoring the Exponential Moving Average of stroke opacity allows for effective removal of redundant components. We achieved an 18.9% reduction in file complexity without any loss in semantic quality. This project highlights the potential of combining differentiable rendering with smart regularization techniques to produce parsimonious and editable vector graphics.

Team Contribution

The workload was distributed as follows:

- **Yifan Zhang (40%):** Lead the algorithm design. Implemented the core `Soft-OBR` logic and EMA tracking. Debugged gradient vanishing issues and tuned the λ_{prune} hyperparameter. Authored the *Methodology* and *Discussion* sections of this report. Designed and conducted the failed Directional CLIP Loss experiment documented in Appendix A.
- **Yule Xu (30%):** Responsible for the evaluation framework. Wrote the evaluating script to batch-process SVG files for SC, CLIP, and LPIPS scores. Authored the *Introduction*, *Experiments*, and Appendix B sections of this report, and performed overall editing and proofreading.
- **Defu Qian (30%):** Focused on the reproduction of the DiffSketcher baseline. Responsible for environment configuration, cloud server management, and collecting the initial baseline data. Authored the *Experimental Setup* subsection and *Conclusion* section, and assisted in organizing experimental data tables.

References

[1] X. Xing, C. Wang, H. Zhou, J. Zhang, Q. Yu, and D. Xu, "Diffsketcher: Text guided vector sketch synthesis through latent diffusion models," *NeurIPS*, 2023.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[3] K. Frans, L. Soros, and O. Witkowski, "CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders," *arXiv preprint arXiv:2102.11006*, 2021.

[4] T.-M. Li, M. Lukáč, M. Gharbi, and J. Ragan-Kelley, "Differentiable vector graphics rasterization for editing and learning," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 193–1, 2020.

[5] A. Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[6] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

APPENDIX

FAILED EXTENSION: DIRECTIONAL CLIP LOSS FOR STYLE TRANSFER

A. Motivation and Implementation

Inspired by recent work in text-to-3D synthesis [6], we attempted to extend DiffSketcher with *Directional CLIP Loss* to enable explicit style transfer. The goal was to inject artistic styles (e.g., Van Gogh's swirling brushstrokes and vivid color palette) into vector sketches while preserving the underlying geometric structure of the target subject (e.g., Sydney Opera House).

The directional CLIP loss is formulated as:

$$\mathcal{L}_{dir} = \lambda_{dir} \cdot \left(1 - \frac{\Delta_{\text{render}} \cdot \Delta_{\text{text}}}{\|\Delta_{\text{render}}\|_2 \|\Delta_{\text{text}}\|_2}\right) \tag{8}$$

where:

$$\Delta_{\text{render}} = E_{\text{CLIP}}(I_{render}) - E_{\text{CLIP}}(I_{neutral}), \tag{9}$$

$$\Delta_{\text{text}} = E_{\text{CLIP}}(P_{style}) - E_{\text{CLIP}}(P_{neutral}) \tag{10}$$

Here, $I_{neutral}$ is a reference rendering without style injection, $P_{neutral}$ is the base prompt (e.g., "a photo of Sydney Opera House"), and P_{style} is the style-augmented prompt (e.g., "An oil painting of Sydney Opera House in Van Gogh style, swirling brushstrokes, blue and yellow vivid colors").

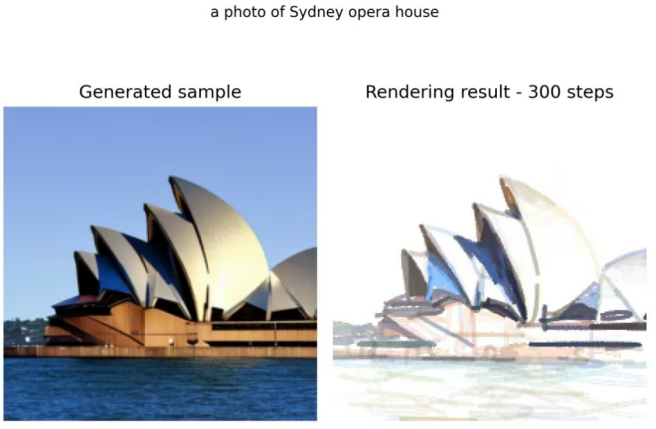
We integrated this loss into the DiffSketcher optimization pipeline alongside the existing ASDS loss:

$$\mathcal{L}_{total} = \mathcal{L}_{SDS} + \lambda_{vis}\mathcal{L}_{CLIP} + \lambda_{dir}\mathcal{L}_{dir} \tag{11}$$

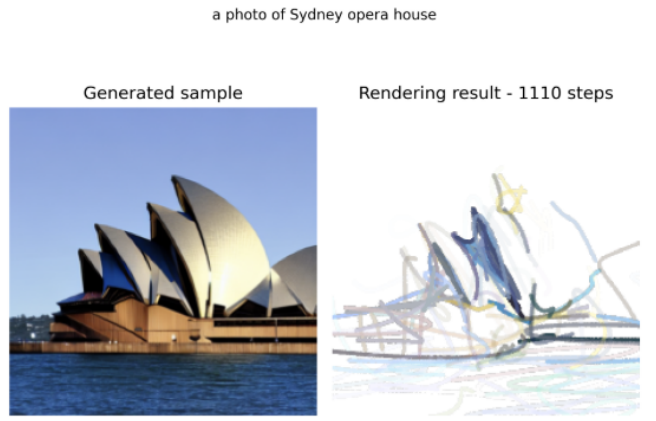
B. Experimental Observation: Severe Geometric Distortion

We conducted an experiment on the Sydney Opera House prompt to evaluate the effectiveness of directional CLIP loss for style transfer. Figure 3 presents a comparative analysis.

Critical Observation: While the baseline method (Figure 3a) produces a structurally faithful sketch in 300 iterations, adding the directional CLIP loss (Figure 3b) causes the Opera House's shell structures to collapse into chaotic, displaced strokes even after 1110 iterations. This demonstrates that the directional constraint fundamentally conflicts with geometric preservation in vector space.



(a) Baseline (without directional CLIP loss): Structurally accurate rendering at 300 iterations preserves the Opera House's iconic shell architecture.



(b) With directional CLIP loss: Severe geometric collapse at 1110 iterations. Control points shift chaotically, destroying structural integrity despite prolonged optimization.

Fig. 3: Comparison of rendering quality with and without directional CLIP loss. The directional loss introduces catastrophic geometric distortion, demonstrating fundamental incompatibility with vector graphics parameterization.

C. Root Cause Diagnosis

We identify three fundamental issues that explain this failure:

1) CLIP Semantic Entanglement: CLIP's embedding space does not cleanly factorize *style* and *content*. The direction vector Δ_{text} for "Van Gogh style" encodes not only color and texture attributes but also *geometric deformations* characteristic of Van Gogh's paintings (e.g., exaggerated perspective, distorted proportions in "Starry Night"). When we force Δ_{render} to align with this direction, the optimization inadvertently moves the rendered sketch toward geometric configurations that match Van Gogh's compositional patterns.

Formally, if we decompose the CLIP embedding as $E = E_{geom} \oplus E_{style}$ (geometry and style components), the directional loss implicitly optimizes:

$$\min \|\Delta_{\text{render}} - \Delta_{\text{text}}\|_2 \approx \min \big(\|\Delta^{render}_{geom} - \Delta^{text}_{geom}\|_2 + \|\Delta^{render}_{style} - \Delta^{text}_{style}\|_2\big) \tag{12}$$

Because CLIP does not disentangle these components, geometric corruption is inevitable.

2) Vector Graphics Parameterization Bottleneck: Unlike pixel-based methods (e.g., neural style transfer with CNNs), vector sketches are parameterized by *explicit control points* $\{P_{0:3,i}\}$ of Bézier curves. To change the visual appearance while preserving structure, the optimization must:

1) Adjust colors/widths (local appearance) $\rightarrow$ insufficient to satisfy directional constraint
2) Move control points (geometry) $\rightarrow$ introduces structural distortion

In pixel space, style transfer can modify textures without moving object boundaries. In vector space, any non-trivial appearance change *requires geometric modification*, creating an irreconcilable conflict between style injection and structure preservation.

3) Insufficient Stroke Density for Texture Representation: Van Gogh's style is characterized by dense, swirling brushstrokes that create texture through *spatial repetition* of small strokes. DiffSketcher's $N = 512$ strokes are optimized for *structural efficiency* (minimizing stroke count while preserving edges). Even with color optimization, 512 strokes cannot simultaneously:

- Maintain the Opera House's geometric fidelity (requires $\sim$300-400 strokes for edges)
- Represent Van Gogh-style texture fields (would require 1000+ dense swirls)

The optimization must choose between geometry and style, and CLIP's mixed gradients cause it to compromise both.

D. Attempted Mitigation Strategies and Their Limitations

(1) Spatial Gradient Masking: We attempted to restrict directional CLIP gradients to background regions (sky, water) while preserving foreground structure. However, defining semantic masks via thresholding attention maps proved unreliable, and partial gradient masking caused visual discontinuities at boundaries.

(2) Annealed Loss Weight Scheduling: Gradually increasing λ_{dir} from 0 to 5 over 200 iterations reduced initial geometric shock but ultimately converged to the same distorted state.

(3) Increased Stroke Count: Raising N to 1024 improved texture representation but did not resolve the CLIP entanglement problem. Geometric distortion persisted, albeit with finer-grained details.

E. Lessons Learned and Future Directions

This failed experiment highlights a critical challenge in applying 2D neural style transfer techniques to *structured vector graphics*. Unlike pixel-based representations where style can be applied via convolutional filters, vector sketches couple appearance and geometry at the parametric level.

In conclusion, while directional CLIP loss is effective for pixel-based 3D synthesis, its direct application to vector graphics encounters fundamental representational mismatches. Successfully integrating style transfer into DiffSketcher would require architectural innovations beyond loss function engineering.

This appendix provides the complete per-seed breakdown of all 30 experiments conducted for the main quantitative comparison (Table I). Each prompt was tested with 10 different random seeds (0-9) to ensure statistical robustness.

Statistical Analysis

As shown in Table V, The per-seed data reveals several key insights:

Variance Analysis: The standard deviation in stroke reduction ranges from 5.34% (Cathedral) to 8.18% (Apple), indicating that our pruning mechanism adapts consistently across different random initializations. The low variance in CLIP scores ($\sigma < 0.006$ for all cases) confirms semantic stability.

Content-Dependent Pruning: The inverse correlation between geometric complexity and reduction rate (Simple: 20.04%, Medium: 23.01%, Complex: 13.71%) demonstrates that our EMA-based threshold $\tau_{death} = 0.02$ successfully preserves structural detail in intricate scenes while aggressively pruning simpler subjects.

Visual Fidelity: The LPIPS metric shows an interesting trend: simpler objects have lower perceptual distance (Apple: 0.059, Opera: 0.087, Cathedral: 0.104), suggesting that redundant strokes in simple scenes contribute less to the final visual appearance, validating our pruning strategy.

Gallery of Generated Sketches

To visually demonstrate the robustness of our Soft-OBR method, we present a gallery of additional generated samples in Fig. 4. These examples span various semantic categories (mechanical, organic, metallic, transparent).

The **Difference Map** (rightmost column) vividly illustrates the efficacy of our method: the "ghost strokes" that were removed are predominantly background noise or internal fillers, while the object boundaries remain untouched.

TABLE V: Complete Per-Seed Results for All Experiments

Experiment	SC Baseline	SC Ours	Reduction (%)	CLIP Base	CLIP Ours	LPIPS
Complex: Gothic Cathedral						
seed0	512	448	12.50	0.4177	0.4150	0.1136
seed1	512	414	19.14	0.4138	0.4158	0.1112
seed2	512	471	8.01	0.4180	0.4175	0.0803
seed3	512	424	17.19	0.4167	0.4172	0.1154
seed4	512	464	9.38	0.4141	0.4136	0.1531
seed5	512	424	17.19	0.4170	0.4185	0.0945
seed6	512	447	12.70	0.4253	0.4255	0.0975
seed7	512	489	4.49	0.4182	0.4148	0.1004
seed8	512	410	19.92	0.4150	0.4148	0.0857
seed9	512	427	16.60	0.4204	0.4216	0.0913
Mean ± Std	**512**	**441.8±27.3**	**13.71±5.34**	**0.4176±0.0034**	**0.4174±0.0036**	**0.1043±0.0224**
Medium: Sydney Opera House						
seed0	512	409	20.12	0.4072	0.4072	0.1074
seed1	512	376	26.56	0.4014	0.4001	0.1051
seed2	512	358	30.08	0.4060	0.4097	0.1167
seed3	512	354	30.86	0.4060	0.4060	0.0351
seed4	512	410	19.92	0.4053	0.4055	0.0919
seed5	512	377	26.37	0.4104	0.4097	0.0886
seed6	512	434	15.23	0.4168	0.4165	0.0866
seed7	512	398	22.27	0.4099	0.4092	0.0590
seed8	512	367	28.32	0.4036	0.4053	0.0655
seed9	512	418	18.36	0.4099	0.4094	0.0931
Mean ± Std	**512**	**394.1±30.3**	**23.01±5.91**	**0.4077±0.0040**	**0.4079±0.0041**	**0.0874±0.0265**
Simple: Red Apple						
seed0	512	314	38.67	0.4075	0.4072	0.0481
seed1	512	480	6.25	0.4199	0.4177	0.0916
seed2	512	425	16.99	0.4182	0.4182	0.0917
seed3	512	429	16.21	0.4133	0.4153	0.0519
seed4	512	359	29.88	0.4226	0.4233	0.0268
seed5	512	409	20.12	0.4119	0.4116	0.0871
seed6	512	442	13.67	0.4153	0.4153	0.0708
seed7	512	424	17.19	0.4185	0.4182	0.0457
seed8	512	384	25.00	0.4226	0.4229	0.0363
seed9	512	428	16.41	0.4189	0.4185	0.0490
Mean ± Std	**512**	**409.4±41.9**	**20.04±8.18**	**0.4175±0.0040**	**0.4177±0.0046**	**0.0590±0.0274**
Overall Statistics (All 30 Experiments)						
Grand Mean ± Std	**512**	**414.7±44.6**	**18.92±7.60**	**0.4143±0.0052**	**0.4143±0.0054**	**0.0836±0.0314**

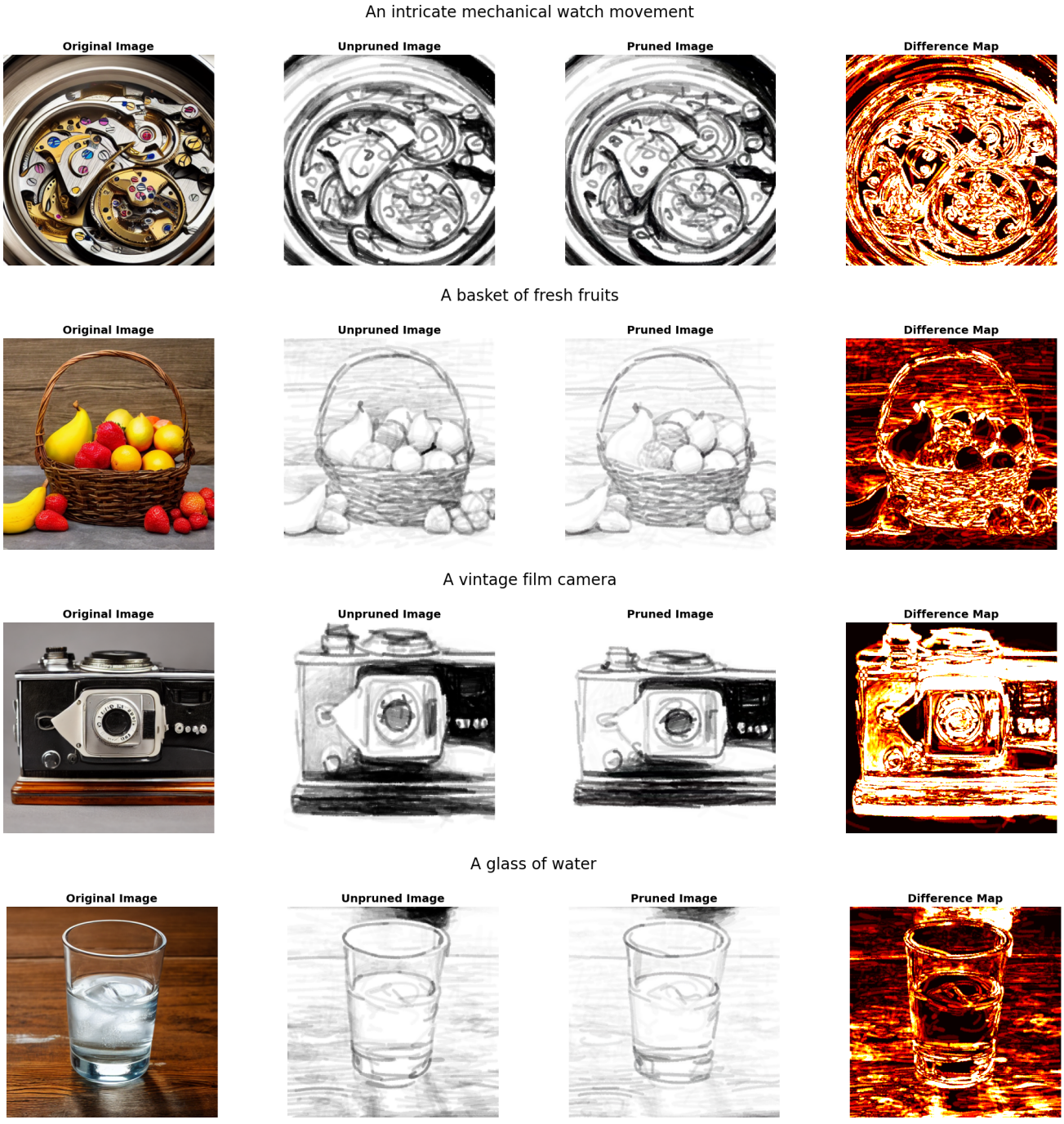


Fig. 4: **Additional Qualitative Results.** From top to bottom: ”An intricate mechanical watch movement”, ”A basket of fresh fruits”, ”A vintage film camera”, and ”A glass of water”. The **Difference Map** confirms that the pruned information consists primarily of redundant ghost strokes, leaving the visual semantics intact.