WACV 2026
TUCSON, AZ 3/6 - 3/10

CraftSVG: Multi-Object Text-to-SVG Synthesis via Layout Guided Diffusion





Ayan Banerjee *Nityanand Mathur* *Josep Lladós* *Umapada Pal* *Anjan Dutta*

CVC Centre de Visió per Computador

UAB Universitat Autònoma de Barcelona

Indian Statistical Institute — Unity in Diversity

University of Surrey

Motivation & Objective For SVG Generation

Conceptualization

Sketches are better to convey the concept than text description.

Text Description

In a village, a river flows between two men, each standing on opposite banks. Men are dressed in traditional clothing, and appear to be wearing turbans or head coverings. Also, the background includes trees, bushes, and some hills.

Corresponding Sketches

What Multi Object Scenarios Are?

Where existing techniques failed?

1. Enumeration

"Two giraffes and three elephants in a forest"

DiffSketcher, (NeurIPS, 2023)

VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)

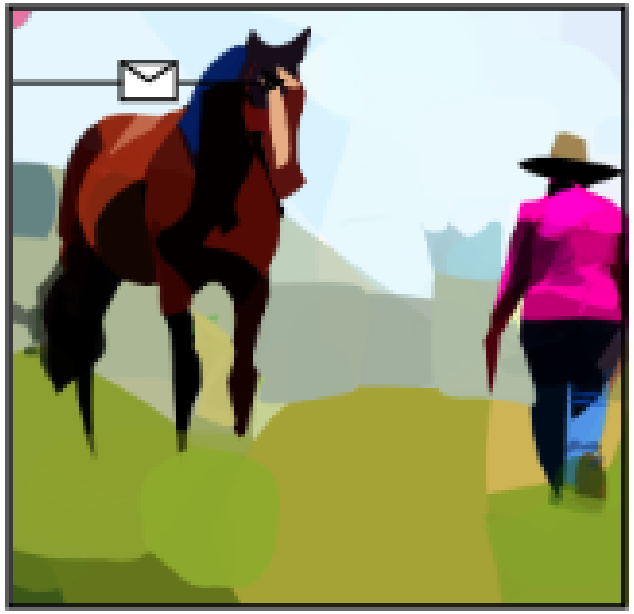
What Multi Object Scenarios Are?

Where existing techniques failed?

2. Spatial Relationship

"A man walks alongside a horse ridden by a woman"

DiffSketcher, (NeurIPS, 2023)



VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)

Challenges with Existing Methods

- LLMs return bounding box coordinate sequentially from top-left to bottom right corner which unable to maintain spatial relationship.
- Denoise foreground and background together which leads to semantic leakage.
- Initialize strokes with CLIP based cross attention or U-Net based self-attention which doesn't consider the semantics.
- Doesn't control opacity leads to artificial sketches than artistic effects/depth effects.

Our Contributions

1. Iterative Layout correction through in-context learning

A man and a woman stroll in the rain, sharing an umbrella, with the woman holding a bag

iter *1*

Reconstructed prompt: A man and a woman walking in the rain, carrying umbrellas, and the man is carrying bags

iter *2*

Reconstructed prompt: A man and a woman walking in the rain, and the man is carrying umbrella and bag

iter *3*

Reconstructed prompt: A man and a woman walking in the rain, sharing the umbrella and a bag is in the road

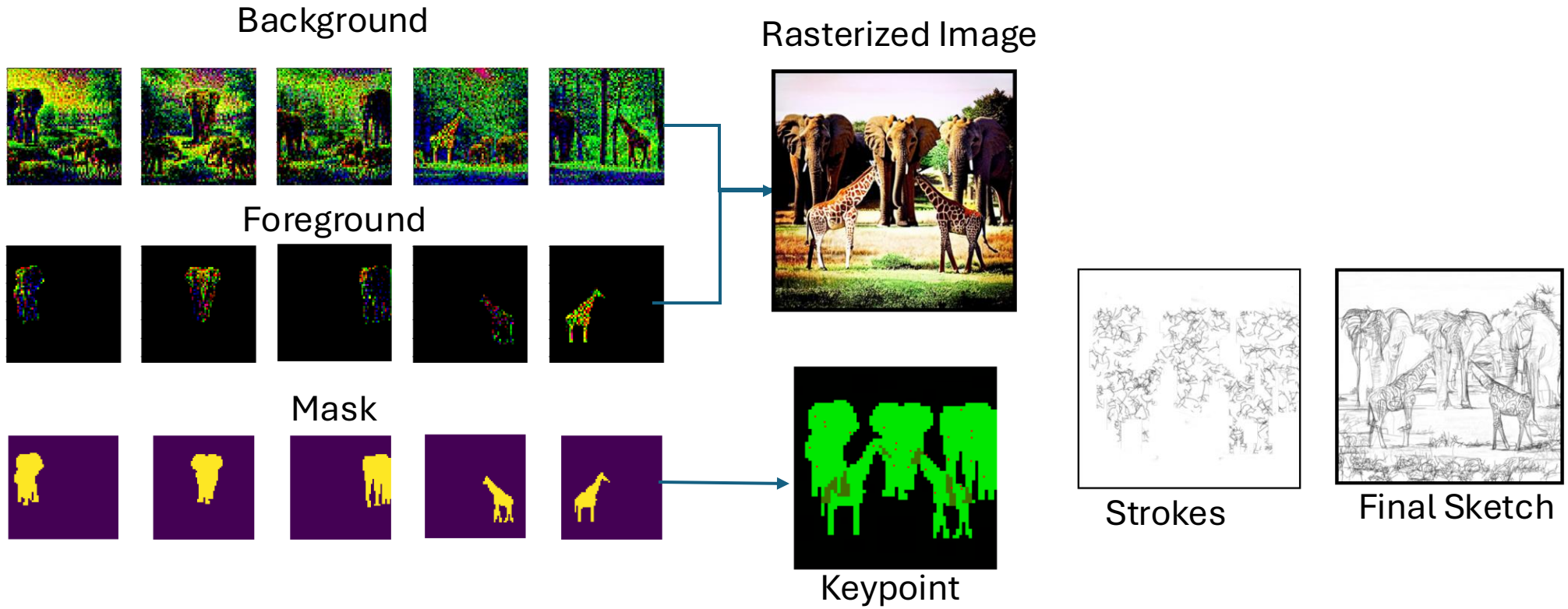
iter *N*

Reconstructed prompt: A man and a woman walking in the rain, sharing the umbrella and the woman is carrying a bag

$$\delta_{\text{rec}} = 1 - [\lambda_1 \text{sim}_{\cos}(T_p, T_r) + \lambda_2 \text{sim}_{\text{jac}}(T_p, T_r) - \lambda_3 \text{sim}_{\text{edit}}(T_p, T_r)$$

Our Contributions

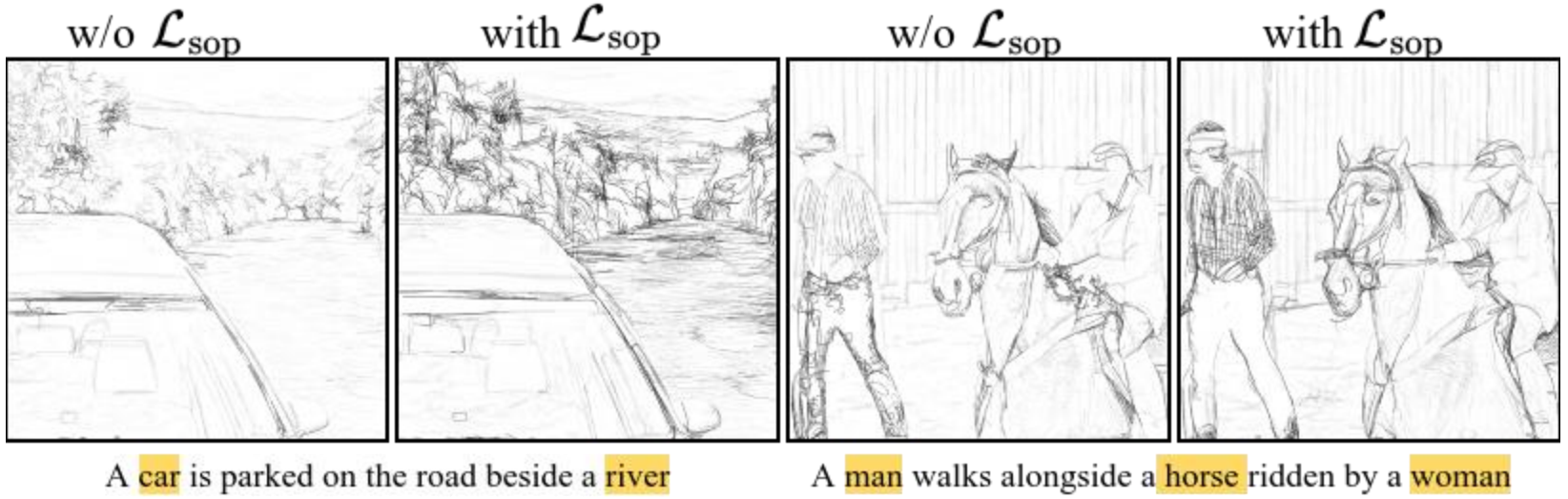
2. Per box mask-latent based canvas initialization



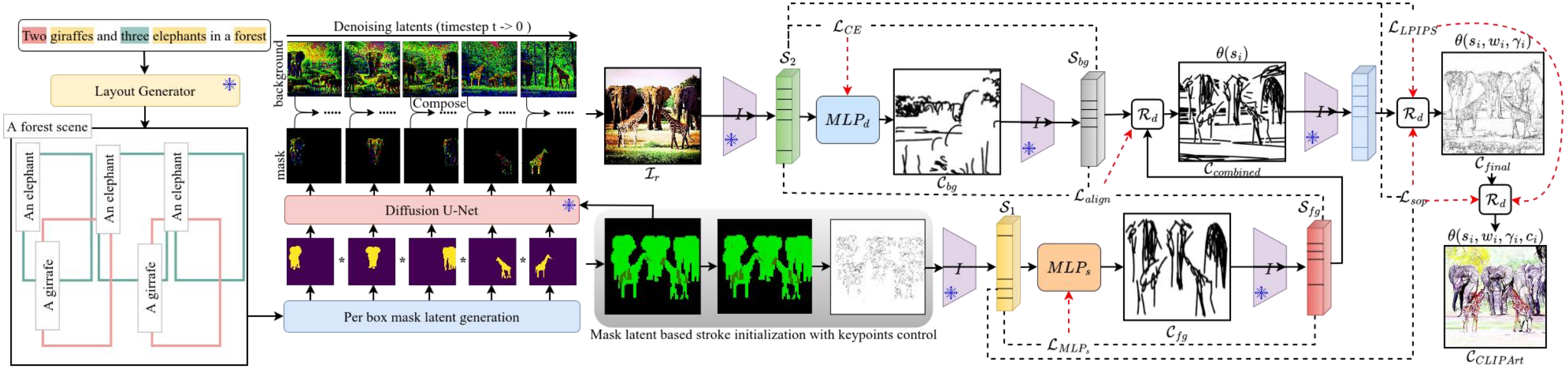
Our Contributions

3. Semantic aware opacity optimization

$$\mathcal{L}_{\text{sop}} = \left| 1 - \frac{\max(A_s \odot \mathcal{R}_d(\theta))}{\max(A_s \odot \mathcal{I}_r)} \right|$$



CraftSVG Architecture



Training Objective

$$\mathcal{L}_{MLP_s}(\mathcal{S}_1, \mathcal{S}_{fg}) = ||\mathcal{S}_1 - \mathcal{S}_{fg}||$$

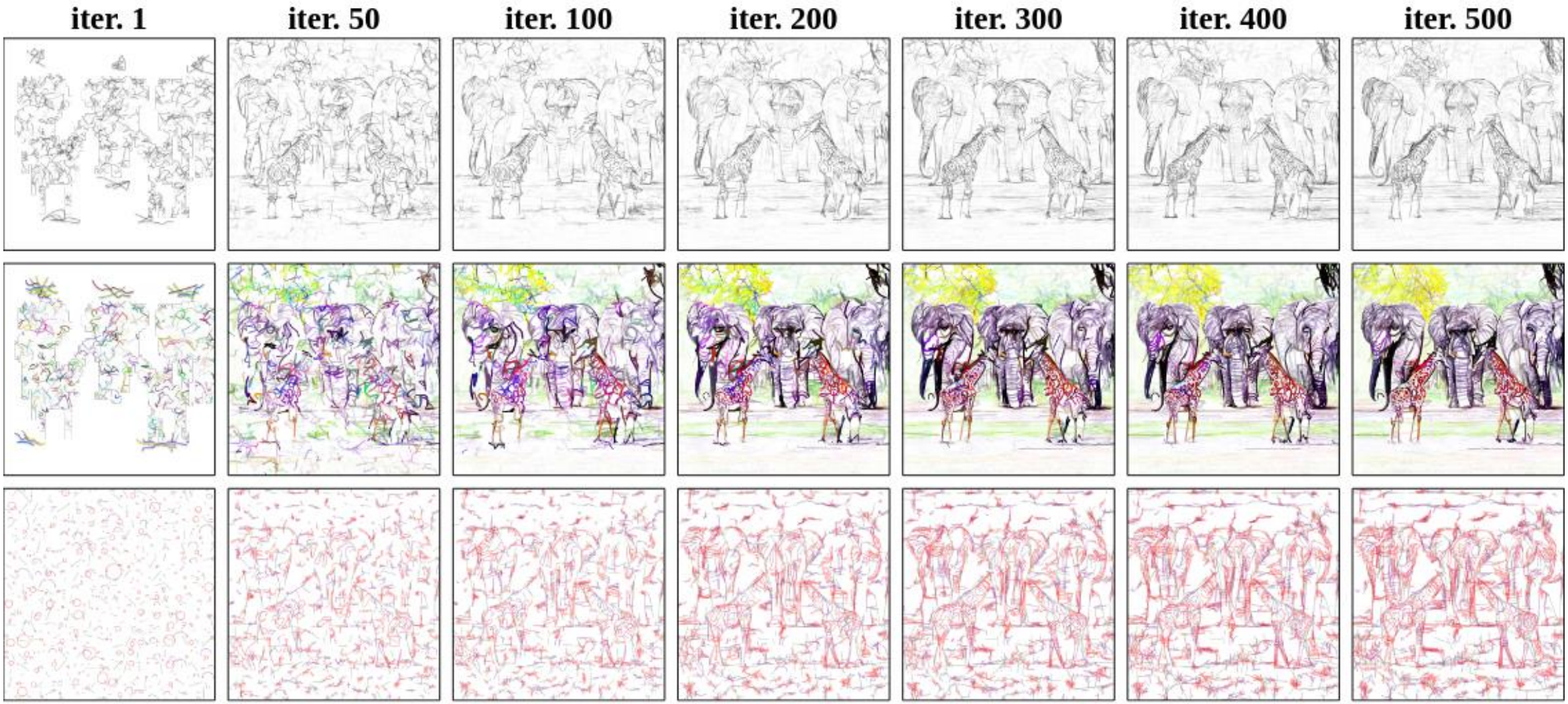
$$\mathcal{L}_{\text{align}}(\mathcal{S}_2, \mathcal{S}_{fg}, \mathcal{S}_{bg}) = \left| \frac{\mathcal{S}_2 - \mathcal{S}_{fg}}{\mathcal{S}_2 - \mathcal{S}_{bg}} \right|$$

$$\mathcal{L}_{\text{synth}} = \text{LPIPS}(\mathcal{I}_r, R_d(\theta)) + \lambda_{\text{sop}}\mathcal{L}_{\text{sop}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$$

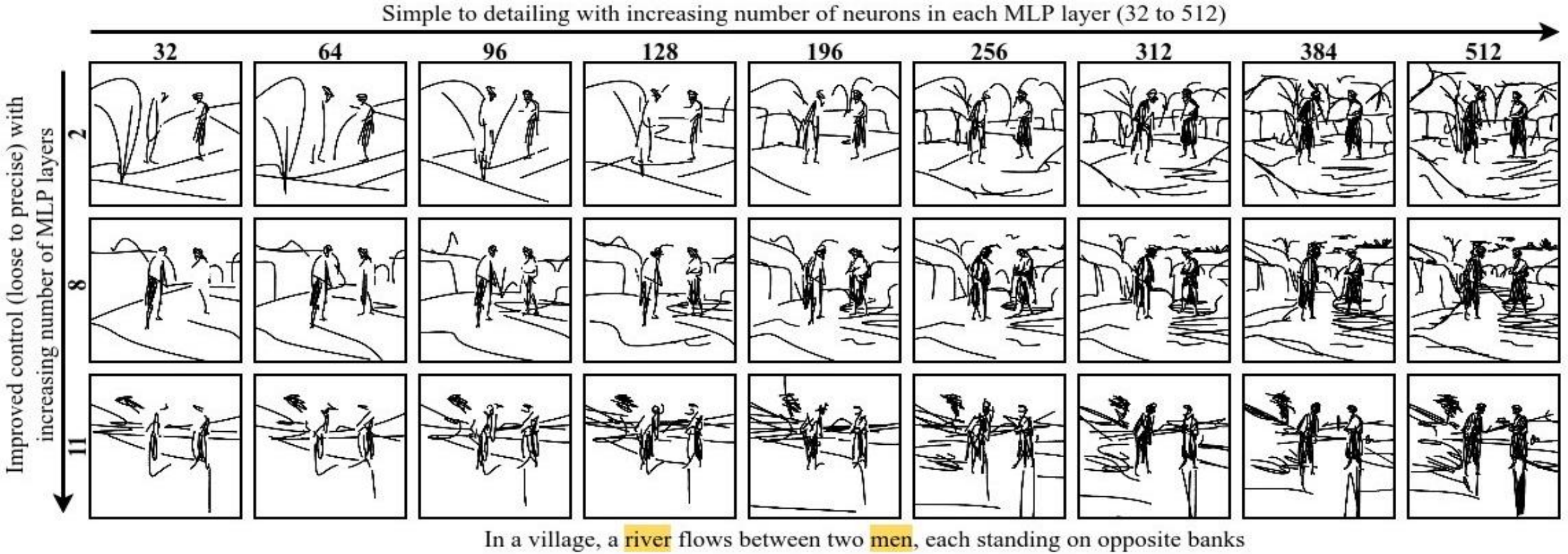
CraftSVG: Abstract Sketch to CLIPArt Generation

1st row depicts the abstract SVG (\#strokes = 64) obtained via the two parallel MLPs. 2nd row, the (\#strokes) = 1024 via optimizing opacity, and semantic awareness. 3rd row further optimizes the color to produce CLIPArt.

Evolution of Bezier Curves



Strokes Abstraction



Increasing the no. of layers and neurons enhances canvas control, detail, and aesthetics, while fewer neurons and layers maintain simplicity and recognizability (No. of strokes in MLP-based abstraction is in the range of [32, 128]).

Can we solve the problem?

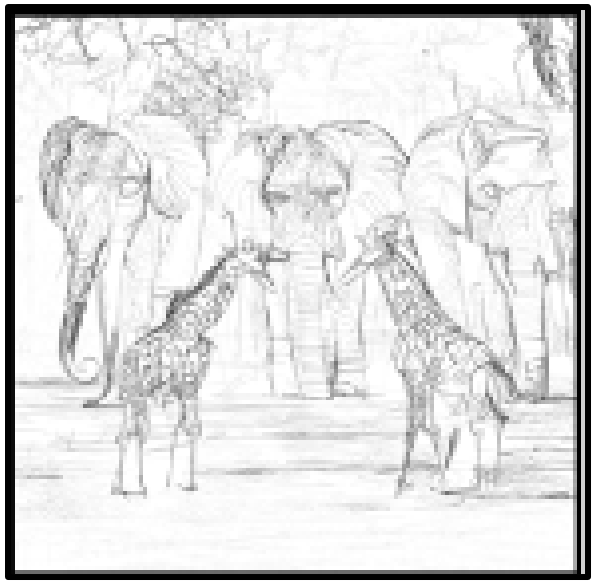
1. Enumeration

"Two giraffes and three elephants in a forest"

DiffSketcher, (NeurIPS, 2023)

VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)



CraftSVG

Can we solve the problem?

2. Spatial Relationship

"A man walks alongside a horse ridden by a woman"

DiffSketcher (NeurIPS, 2023)

VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)

CraftSVG

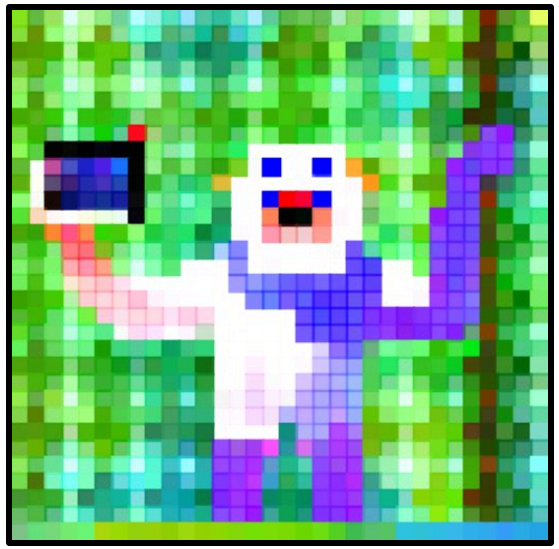
Can we solve the problem?

WACV 2026 TUCSON, AZ 3/6 - 3/10

3. Abstract Concept

"Yeti taking a selfie"

DiffSketcher (NeurIPS, 2023)



VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)

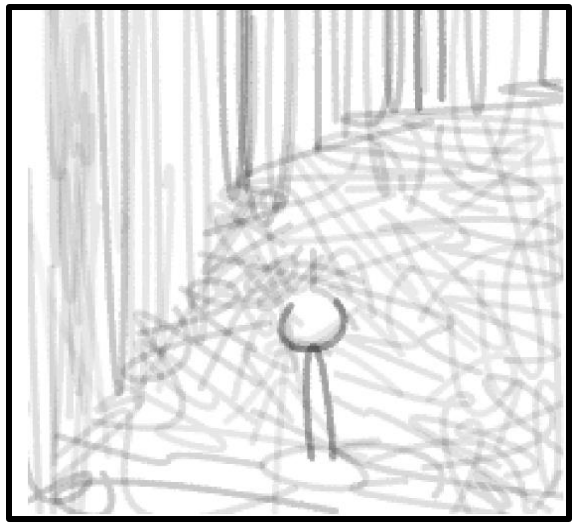
CraftSVG

Can we solve the problem?

WACV 2026 TUCSON, AZ 3/6 - 3/10

4. Scene Text Writing

"A stop sign is in grass behind a fence"



DiffSketcher (NeurIPS, 2023)

VectorFusion (CVPR, 2022)

SVGDreamer (CVPR, 2024)

CraftSVG

Quantitative Evaluation

WACV 2026 TUCSON, AZ 3/6 - 3/10

Table 1. Evaluation on conventional SVG synthesis techniques.

Method / Metric	CS ↑	FID ↓	PSNR ↑	CLIP-T ↑	BLIP ↑	Aes. ↑	HPSv2 ↑	Conf. ↑	Mem. (GB) ↓	Hit % ↑
CLIPDraw (b/w)	0.2882	172.67	4.13	0.1886	0.2672	3.2803	0.1883	0.27	**1.4**	6.8
CLIPDraw (CLIPArt)	0.2911	171.17	4.66	0.1952	0.2704	2.9182	0.1892	0.27	**1.4**	7.1
CLIPDrawX	0.3276	146.12	5.92	0.2102	0.3031	3.2144	0.1997	0.32	8.4	17.2
VectorFusion (Sketch)	0.4211	127.44	8.14	0.2719	0.2991	1.1671	0.2007	0.17	16.2	7.7
VectorFusion (CLIPArt)	0.5117	97.94	11.27	0.3710	0.4421	5.5312	0.2617	0.33	16.2	28.3
DiffSketcher (Sketch)	0.3629	120.04	8.38	0.2855	0.3987	4.2468	0.2411	0.60	12.7	27.6
DiffSketcher (CLIPArt)	0.3538	121.77	10.12	0.2892	0.3921	4.6481	0.2422	0.64	12.7	27.8
SVGDreamer (Abstract)	0.4412	67.12	11.67	0.3001	0.4623	4.8432	0.2685	0.46	32.7	17.6
SVGDreamer (Sketch)	0.5214	61.67	11.87	0.3211	0.4712	5.1423	0.2718	0.37	32.7	42.7
SVGDreamer (CLIPArt)	0.5962	59.13	14.54	0.3440	0.4972	6.5432	0.2888	0.31	32.7	48.8
CraftSVG (Abstract)	0.6176	51.42	15.98	0.4563	0.5223	5.9873	0.3167	**0.66**	12.1	62.1
CraftSVG (Sketch)	0.6342	48.42	16.07	0.4667	0.5432	6.7832	0.3276	**0.66**	12.1	65.2
CraftSVG (CLIPArt)	**0.7091**	**39.87**	**17.15**	**0.5013**	**0.5783**	**7.0779**	**0.3523**	0.61	12.1	**66.7**
CraftSVG (Primitive)	0.6019	55.76	15.51	0.4031	0.5038	5.079	0.3032	0.54	12.1	61.7

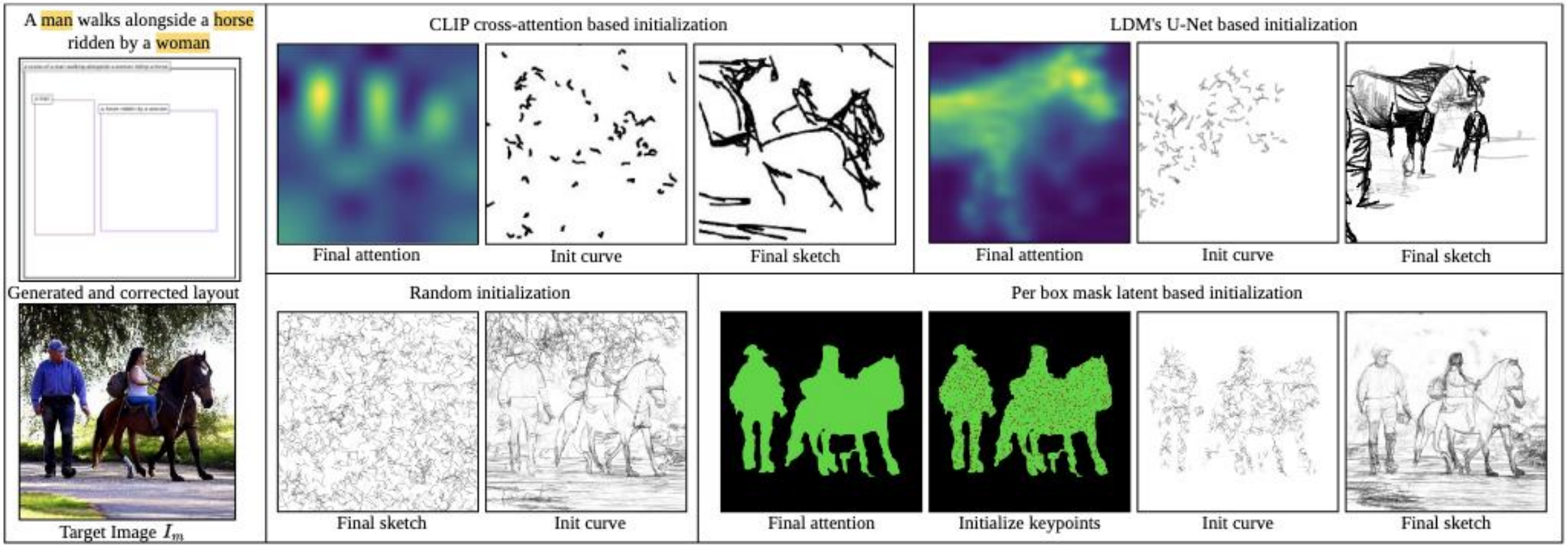
Evaluation of LLMs for Layout Generation

Table 2. Layout generation with LLMs.

Methods	METEOR ↑	ROUGE ↑	CIDEr ↑	SPICE ↑	Average ↑
Text2Scene	0.189	0.446	0.601	0.123	0.339
Llama-7B	0.232	0.573	0.648	0.214	0.416
Vicuna-13B	0.274	0.656	0.721	0.222	0.468
GPT-3.5	0.271	0.711	0.753	0.254	0.497
GPT-4o	0.289	**0.737**	0.797	**0.382**	0.511
Claude-2	0.291	0.717	0.782	0.291	0.538
Claude-3.5 Sonnet	**0.302**	0.728	**0.801**	0.322	**0.544**

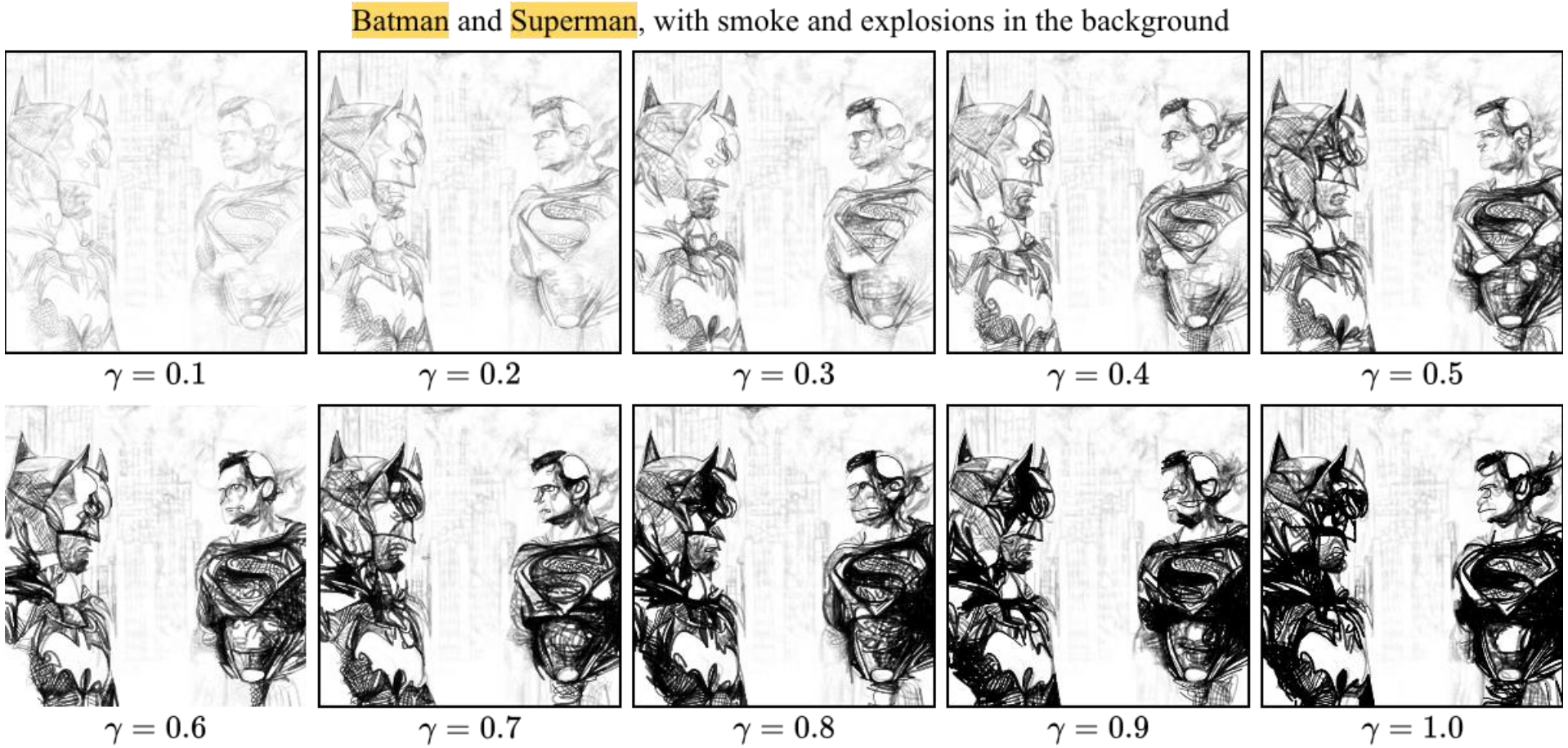
The performances are based on 50 unique prompts tested 5 times each, without layout correction, to determine the best LLM for our work. We started testing the Text2Scene module, which struggled with objects outside the MS-COCO dataset. We then explored LLMs, starting with Llama-7B, which improved the baseline by 8%. Vicuna-13B further refined Llama-7B, offering a 5% improvement, and GPT-3.5 surpassed Vicuna-13B by 3%. Between GPT-4o and Claude-3.5 Sonnet, with a 2:3 ratio favoring Claude-3.5 Sonnet, we chose the latter for our experiments.

Ablation on Canvas Initialization



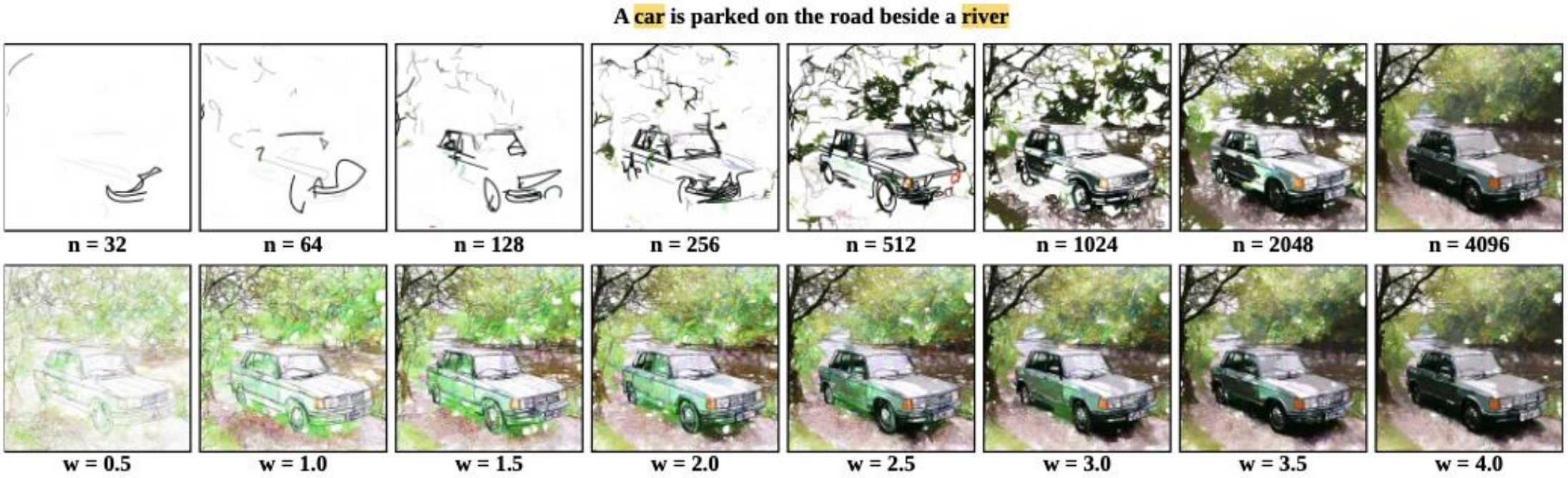
LDM and CLIP attention fail to capture object-level detail, while random init adds clutter. Per-box mask latent attention ensures complete, detailed canvas coverage.

Ablation on Opacity



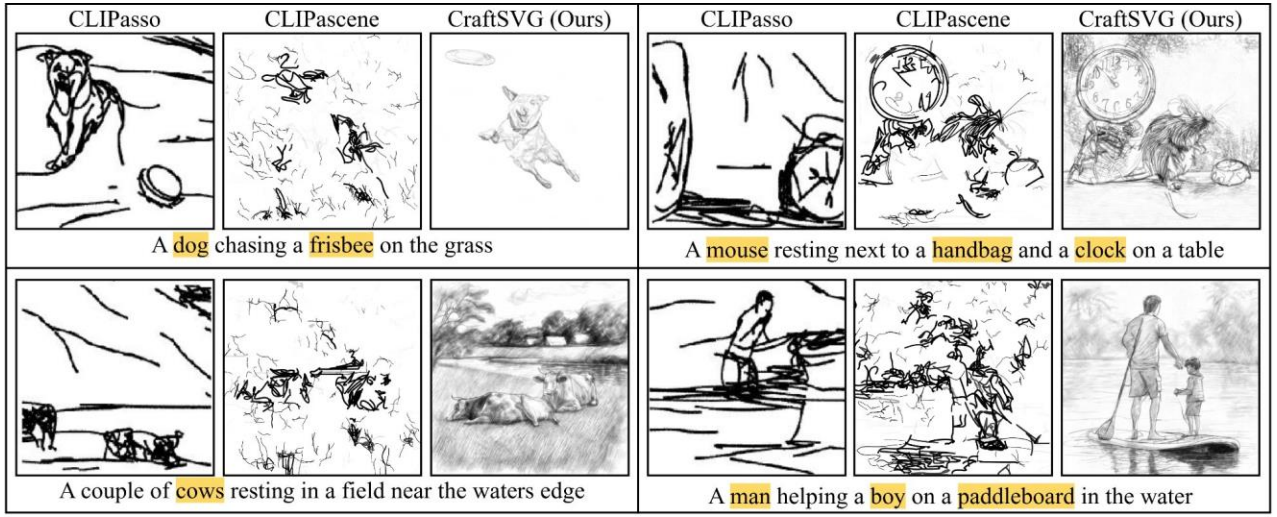
Initializing strokes with higher gamma values enhances foreground-background contrast (n: 1024, w: 4.0).

Ablation on Strokes



With an increasing number of parameters, CLIPArt gets a realistic look.

Ablation on Strokes

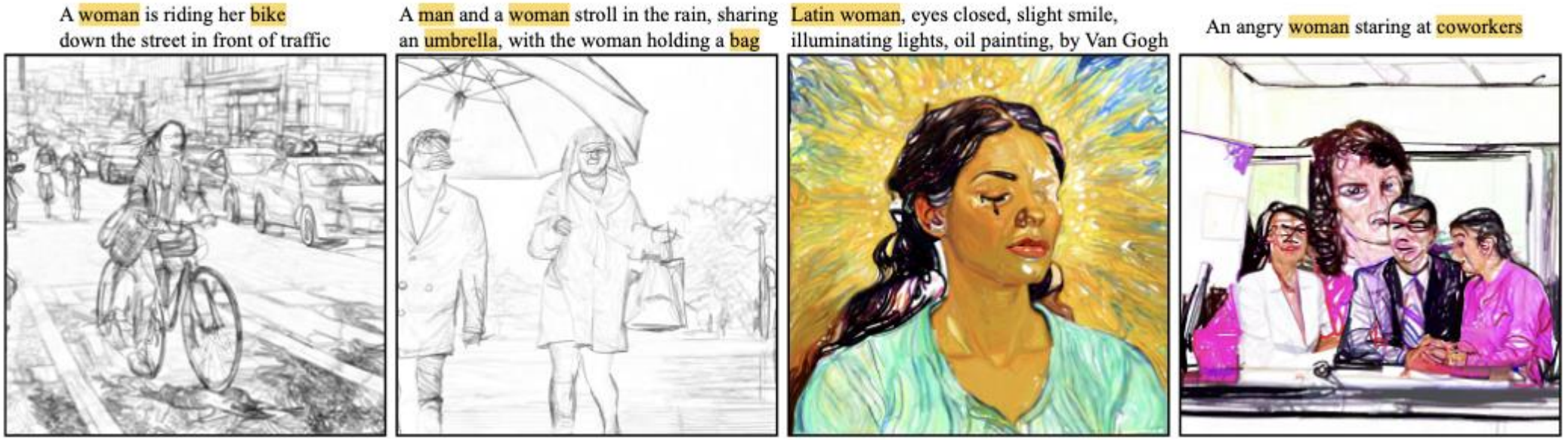


We have used the same rasterized images generated by the proposed diffusion method as a starting point. (no. of strokes = 1024).

Table 3. Comparison with image abstraction techniques.

Method / Metric	CS ↑ [48]	FID ↓ [49]	PSNR ↑ [50]	CLIP-T ↑ [2]	BLIP ↑ [51]	Aes. ↑ [36]	HPSv2 ↑ [37]
CLIPasso [18]	0.3517	98.12	9.34	0.2122	0.3106	3.9523	0.1991
CLIPascene [19]	0.3216	101.08	8.82	0.2003	0.2937	3.0143	0.1712
CraftSVG (Sketch)	**0.6342**	**48.42**	**16.07**	**0.4563**	**0.5223**	**6.7832**	**0.3167**

Failure Cases



CraftSVG is unable to synthesize detailed faces while maintaining enumeration and spatial relationship.

Conclusions & Future Scope

- ✓ Generate multi-object SVGs by maintaining enumeration and spatial relationship through per-box mask latent mechanism.
- ✓ Generate abstract to detail SVG through two parallel MLPs and Bezier curve evolution.
- ✓ Iteratively correct layout without finetuning or further retraining of the LLMs.
- ✓ We will try to improve the detailed face drawing through face-consistent attention mechanism.


Thank
you!