Text to Sketch Generation with Multi-Styles

Tengjie Li, Shikui Tu*, Lei Xu*

{765127364, tushikui, leixu}@sjtu.edu.cn

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Abstract & Key Contributions

Problem:

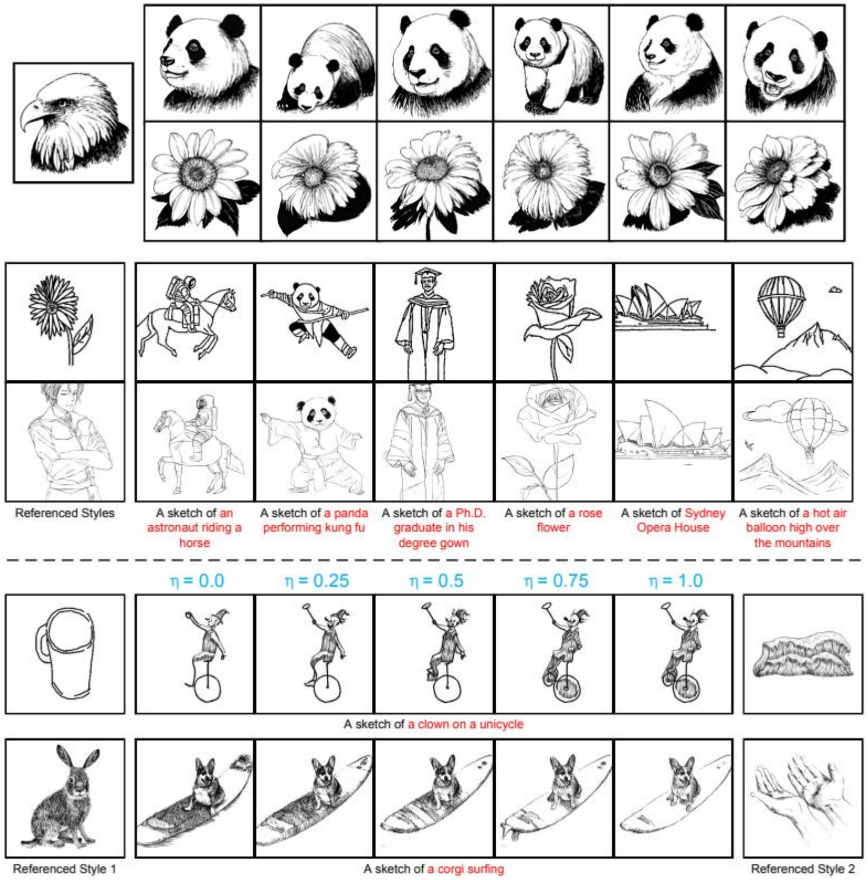
- Existing sketch generation methods lack precise style control mechanisms

Solution:

- M3S - Training-free framework based on diffusion models

Key Innovations:

- Reference feature injection with linear smoothing
- Style-content guidance mechanism
- Multi-style fusion via joint AdaIN modulation during denoising process



Introduction & Motivation

Sketching:

- Universal visual medium transcending cultural barriers

Challenges:

- Data acquisition difficulties for high-quality sketches
- Limited style controllability in existing methods
- Domain gap between natural images and sketches

Motivation:

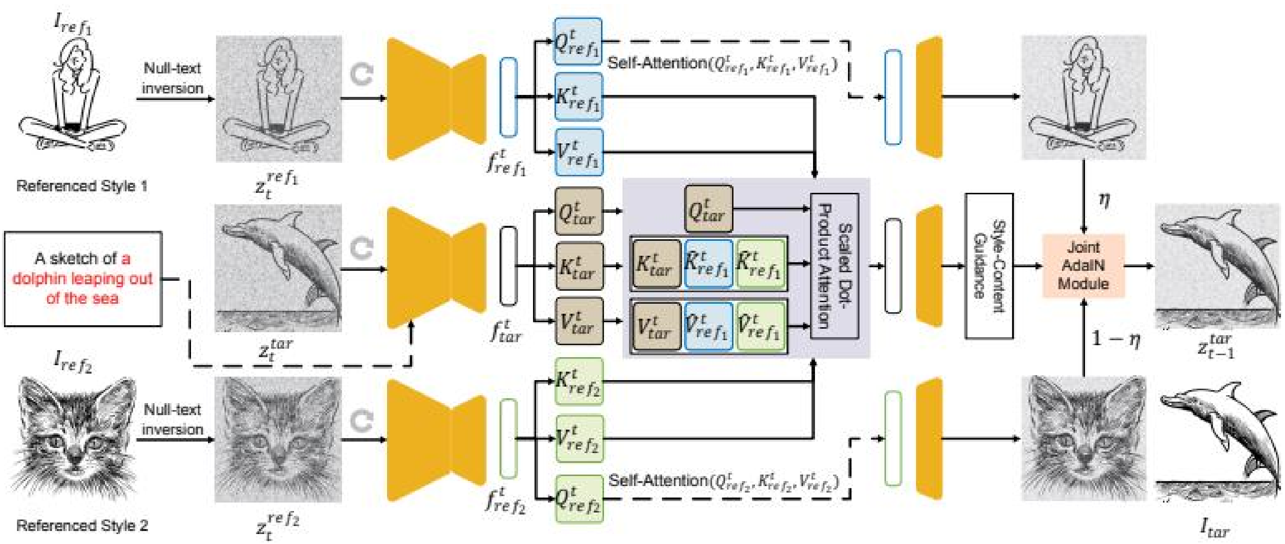
- Zero-shot style transfer to overcome data limiatations
- Leveraging pre-trained knowledge of text to image diffusion models

Methodology Overview

Framework: M3S pipeline for single and multi-style generation

Core Components:

- Style feature injection in self-attention layers, linear blending to mitigate content leakage
- Joint AdaIN for style tendency control
- Improved classifier-free guidance for style-content guidance balancing



Feature Injection Mechanism

Previous Limitations: Direct K/V substitution causes content leakage

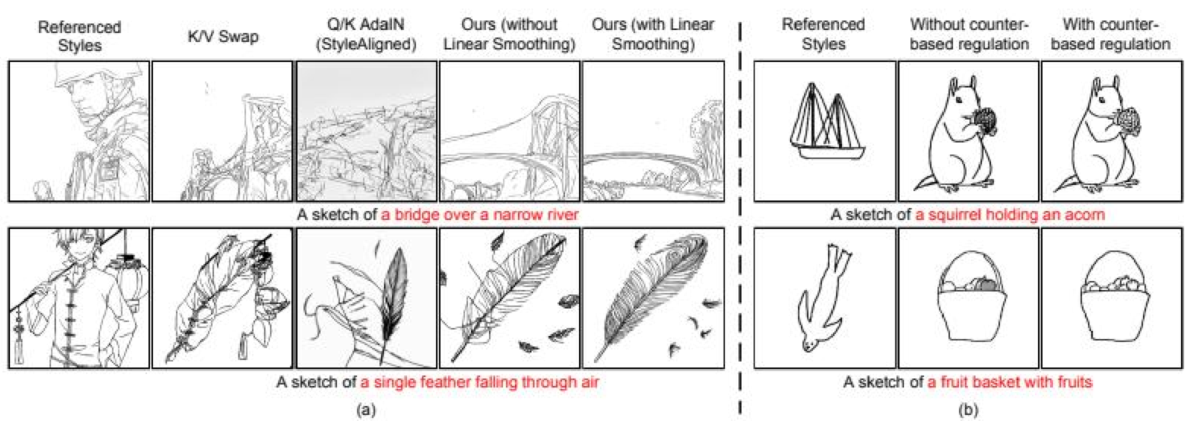
Our Approach: Concatenation strategy with linear smoothing

- Vanilla attention: $Attention(Q_{tar}, K_{tar}, V_{tar}) = softmax\left(\frac{Q_{tar}K_{tar}^T}{\sqrt{d}}\right)V_{tar}.$
- M3S single-style sketch generation:

$$Attention(Q_{tar}, \begin{bmatrix} K_{tar} \\ \hat{K}_{ref1} \end{bmatrix}, \begin{bmatrix} V_{tar} \\ \hat{V}_{ref1} \end{bmatrix}), \quad \begin{matrix} \hat{K}_{ref1} = \lambda K_{tar} + (1-\lambda)K_{ref1} \\ \hat{V}_{ref1} = \lambda V_{tar} + (1-\lambda)V_{ref1} \end{matrix}.$$

- M3S multi-style sketch generation:

$$Attention(Q_{tar}, [K_{tar}, \hat{K}_{ref1}, \hat{K}_{ref2}], [V_{tar}, \hat{V}_{ref1}, \hat{V}_{ref2}]).$$



Multi-Style Control and Style-Content Guidance

Multi-Style Control with Joint AdaIN Module:

$$z_t^{tar} = \eta * AdaIN(z_t^{tar}, z_t^{ref1}) + (1-\eta) * AdaIN(z_t^{tar}, z_t^{ref2})$$

Flexibility: Continuous interpolation between multiple styles

Style-Content Guidance-Dual Control Pathways:

- Content guidance (ω_1): Maintains text alignment
- Style guidance (ω_2): Ensures style consistency

Adaptive Scheduling: ω_2 linearly increases during denoising

Balancing: Optimal trade-off between fidelity and style expression

$$\begin{aligned}\tilde{\epsilon}_t = \epsilon_\theta(\mathbf{z}_t^{tar}, t, \emptyset) &+ \omega_1 \underbrace{\left(\epsilon_\theta^{\times}(\mathbf{z}_t^{tar}, t, text, K_{ref}, V_{ref}) - \epsilon_\theta(\mathbf{z}_t^{tar}, t, \emptyset)\right)}_{content\ guidance\ direction} \\ &+ \omega_2 \underbrace{\left(\epsilon_\theta^{\times}(\mathbf{z}_t^{tar}, t, \emptyset, K_{ref}, V_{ref}) - \epsilon_\theta(\mathbf{z}_t^{tar}, t, \emptyset)\right)}_{style\ guidance\ direction},\end{aligned}$$

Experimental Setup

Datasets: 6 diverse sketch styles

- 4 professional sketches dataset from 4SKST dataset
- 1 web-collected dataset
- 1 abstract dataset from Sketch dataset

Evaluation Metrics:

- CLIP-T: Text-sketch alignment
- DINO/VGG: Style consistency
- Human preference assessment

Baselines: StyleAligned, InstantStyle, CSGO, AttentionDistillation, etc.

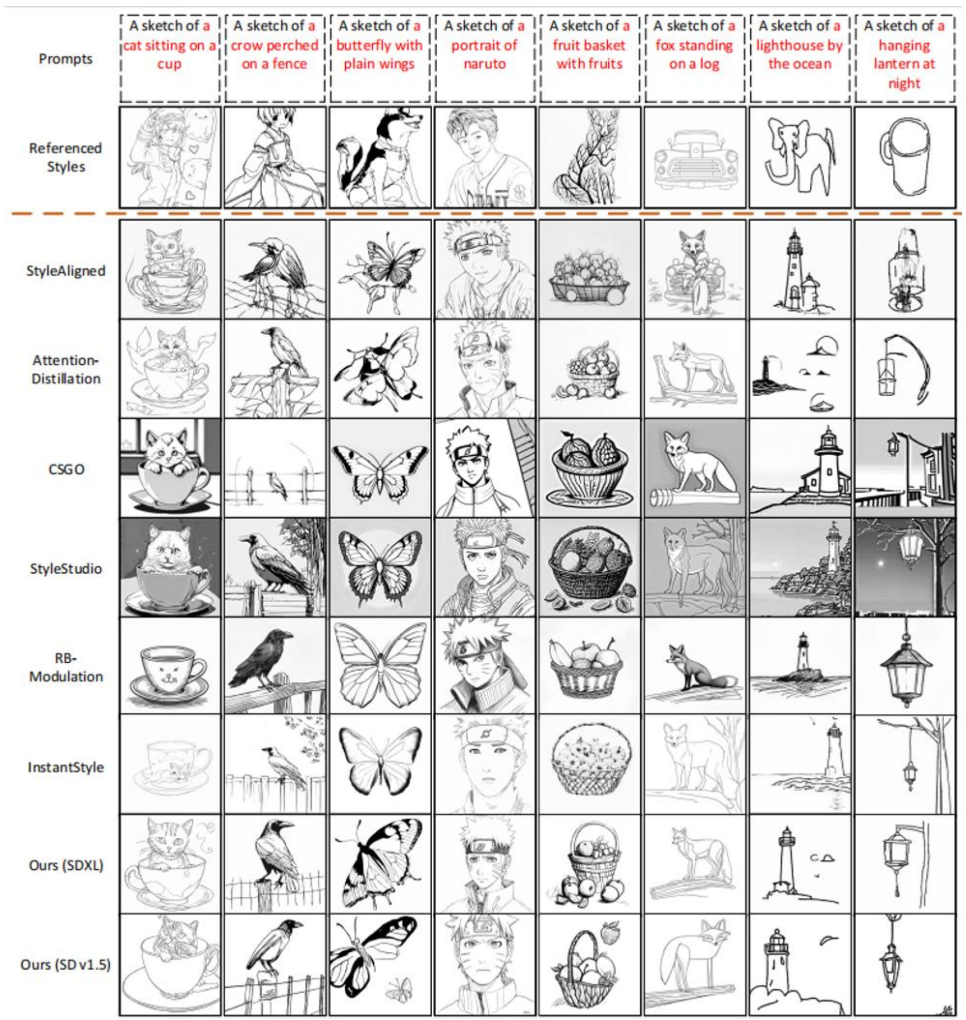
Qualitative Results

Superior Performance:

- M3S achieves better style consistency without content leakage
- Well balance between style consistency and text aligment

Cross-domain Synthesis:

- Effective even with structurally divergent references



Multi-Style Generation Examples

Style Fusion:

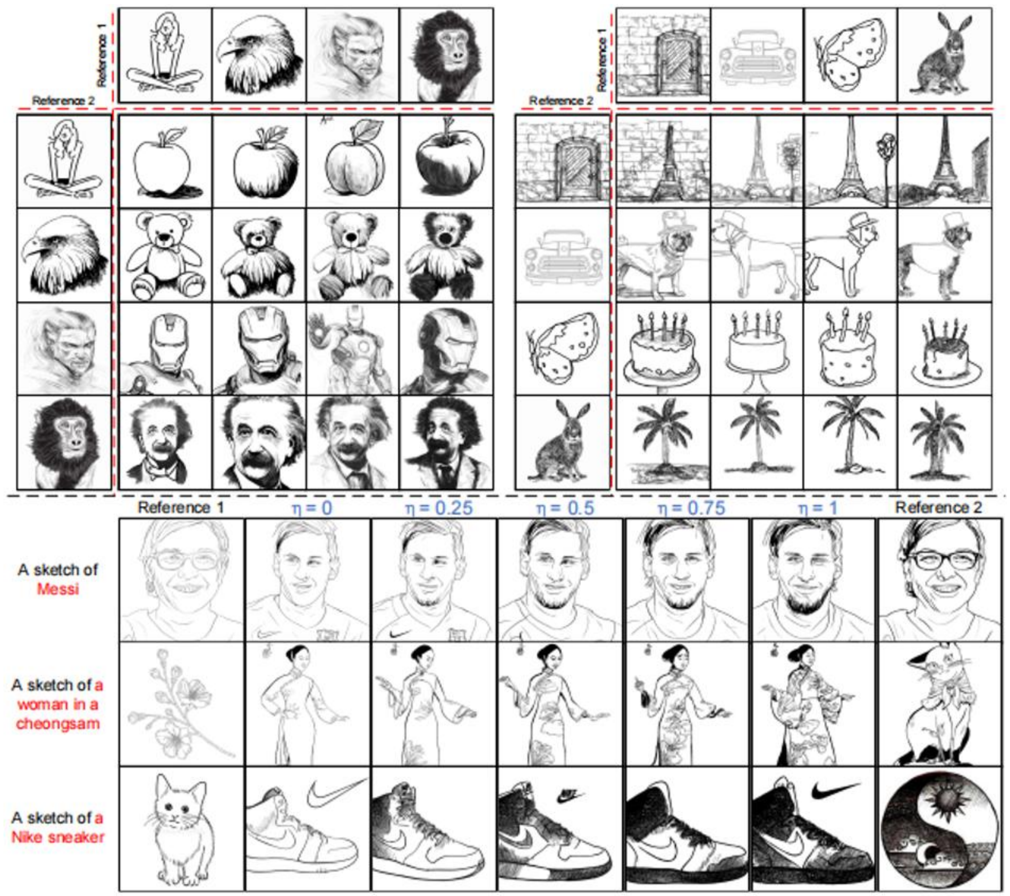
- Combining contour clarity from one reference with texture patterns from another

Controllable Interpolation:

- Smooth transition between styles via η parameter

Creative Applications:

- Enables novel artistic expressions



Quantitative Analysis

CLIP Scores:

- M3S(SDXL) achieves best text alignment (0.3514)

Style Consistency:

- Competitive DINO and VGG metrics

Human Evaluation:

- Highest preference ratings (6.19/8.0)

Statistical Significance:

- p-value = 1.06×10^{-5} against strongest baseline

Table 1: Sketch-text alignment and style consistency performance comparison across styles. 'Ours (SDXL*)' denotes that the parameters of our method are set to $\omega_1 = 7.5, \omega_2 = 20$, and $\lambda = 0.0$.

Method	Style1			Style2			Style3		
	CLIP-T(↑)	DINO(↑)	VGG(↓)	CLIP-T(↑)	DINO(↑)	VGG(↓)	CLIP-T(↑)	DINO(↑)	VGG(↓)
StyleAligned [12]	0.3130	0.6691	0.0308	0.3095	0.7064	0.0684	0.3013	0.6309	0.0621
AttentionDistillation [63]	0.3305	**0.7738**	**0.0930**	0.3320	**0.7724**	**0.0320**	0.3225	**0.7132**	**0.0305**
CSGO [52]	0.3336	0.5276	0.0571	0.3257	0.5409	0.1370	0.3232	0.5154	0.1018
StyleStudio [21]	0.3395	0.5164	0.1873	0.3351	0.5601	0.1954	0.3349	0.5337	0.1790
RB-Modulation [34]	0.3298	0.3624	0.0592	0.3300	0.3429	0.2085	0.3279	0.3453	0.1733
InstantStyle [48]	0.3512	0.4934	0.0417	0.3508	0.4929	0.1577	**0.3455**	0.4394	0.1321
Ours (SDXL)	**0.3607**	0.6545	0.0165	**0.3556**	0.6531	0.0674	0.3422	0.6041	0.0534
Ours (SD v1.5)	0.3507	0.6383	0.0200	0.3452	0.6846	0.0616	0.3416	0.6269	0.0571
Ours (SDXL*)	0.3480	0.7344	0.0122	0.3340	0.7356	0.0464	0.3319	0.6870	0.0371
	Style4			**Style5**			**Style6**		
	CLIP-T(↑)	DINO(↑)	VGG(↓)	CLIP-T(↑)	DINO(↑)	VGG(↓)	CLIP-T(↑)	DINO(↑)	VGG(↓)
StyleAligned [12]	0.3137	0.6407	0.0244	0.3004	0.5428	0.0445	0.2879	0.4445	0.0300
AttentionDistillation [63]	0.3222	**0.7572**	**0.0061**	0.3377	**0.6221**	**0.0173**	0.3289	0.7027	0.0190
CSGO [52]	0.3321	0.5134	0.0526	0.3298	0.4288	0.0972	0.3241	0.5012	0.0716
StyleStudio [21]	0.3402	0.5100	0.1595	0.3377	0.3539	0.1215	0.3338	0.3612	0.1434
RB-Modulation [34]	0.3178	0.3373	0.0465	0.3247	0.3233	0.0972	0.3221	0.2737	0.0780
InstantStyle [48]	0.3513	0.4494	0.0262	0.3480	0.4408	0.0601	0.3417	0.5130	0.0421
Ours (SDXL)	**0.3612**	0.6493	0.0115	0.3467	0.5332	0.0304	**0.3420**	0.6922	0.0259
Ours (SD v1.5)	0.3518	0.6337	0.0136	**0.3494**	0.5777	0.0272	0.3405	**0.7653**	**0.0170**
Ours (SDXL*)	0.3506	0.7212	0.0085	0.3383	0.6328	0.0191	-	-	-

Table 2: The average rating of different methods by the human preference assessment.

	StyleAligned [12]	AttentionDistillation [63]	CSGO [52]	StyleStudio [21]
Rating	2.77	4.28	3.83	4.22
	RB-Modulation [34]	**InstantStyle [48]**	**Ours(SD v1.5)**	**Ours (SDXL)**
Rating	4.20	5.08	5.44	**6.19**

Ablation Studies

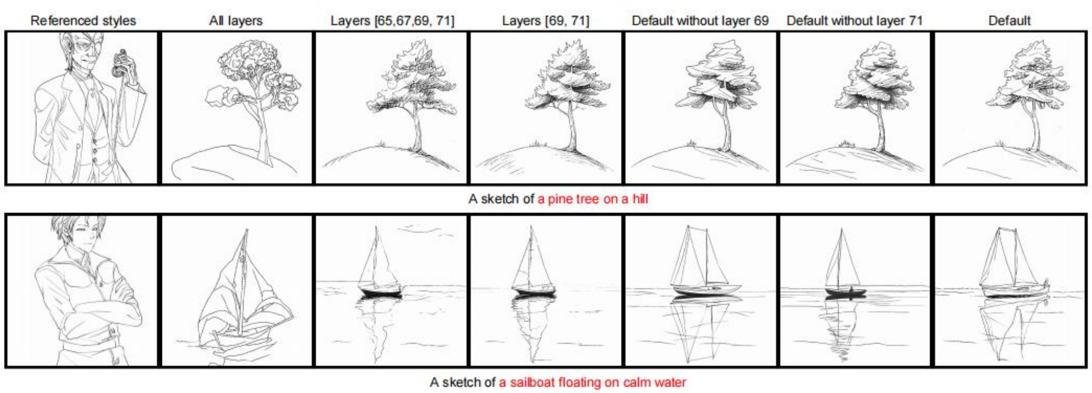
Parameter Analysis: λ=0.1 provides optimal balance

Table 3: Multi-style sketch generation performance under different η values.

M3S Imp.	Ref. style	$\eta=0$ CLIP-T(↑)	$\eta=0$ DINO-ref1(↑)	$\eta=0$ DINO-ref2(↑)	$\eta=0.25$ CLIP-T(↑)	$\eta=0.25$ DINO-ref1(↑)	$\eta=0.25$ DINO-ref2(↑)	$\eta=0.5$ CLIP-T(↑)	$\eta=0.5$ DINO-ref1(↑)	$\eta=0.5$ DINO-ref2(↑)
SDXL	S5-S5	0.3442	0.3936	0.4944	0.3514	0.4180	0.4821	0.3495	0.4408	0.4556
SD v1.5	S5-S5	0.3465	0.3850	0.4776	0.3453	0.4215	0.4597	0.3499	0.4469	0.4509
SDXL	QD-S5	0.3426	0.3051	0.4724	0.3455	0.3266	0.4622	0.3457	0.3330	0.4397
SD v1.5	QD-S5	0.3434	0.3630	0.4339	0.3417	0.3948	0.4236	0.3452	0.4102	0.4057

M3S Imp.	Ref. style	$\eta=0.75$ CLIP-T(↑)	$\eta=0.75$ DINO-ref1(↑)	$\eta=0.75$ DINO-ref2(↑)	$\eta=1$ CLIP-T(↑)	$\eta=1$ DINO-ref1(↑)	$\eta=1$ DINO-ref2(↑)
SDXL	S5-S5	0.3499	0.4578	0.4221	0.3470	0.4693	0.3975
SD v1.5	S5-S5	0.3478	0.4528	0.4257	0.3528	0.4626	0.3825
SDXL	QD-S5	0.3447	0.3409	0.4209	0.3396	0.3617	0.3916
SD v1.5	QD-S5	0.3440	0.4250	0.3938	0.3468	0.4381	0.3766

Layer Selection: Strategic feature injection in specific UNet layers (SDXL)



Ablation Studies

Diffrent Control Strength of Syle and Content

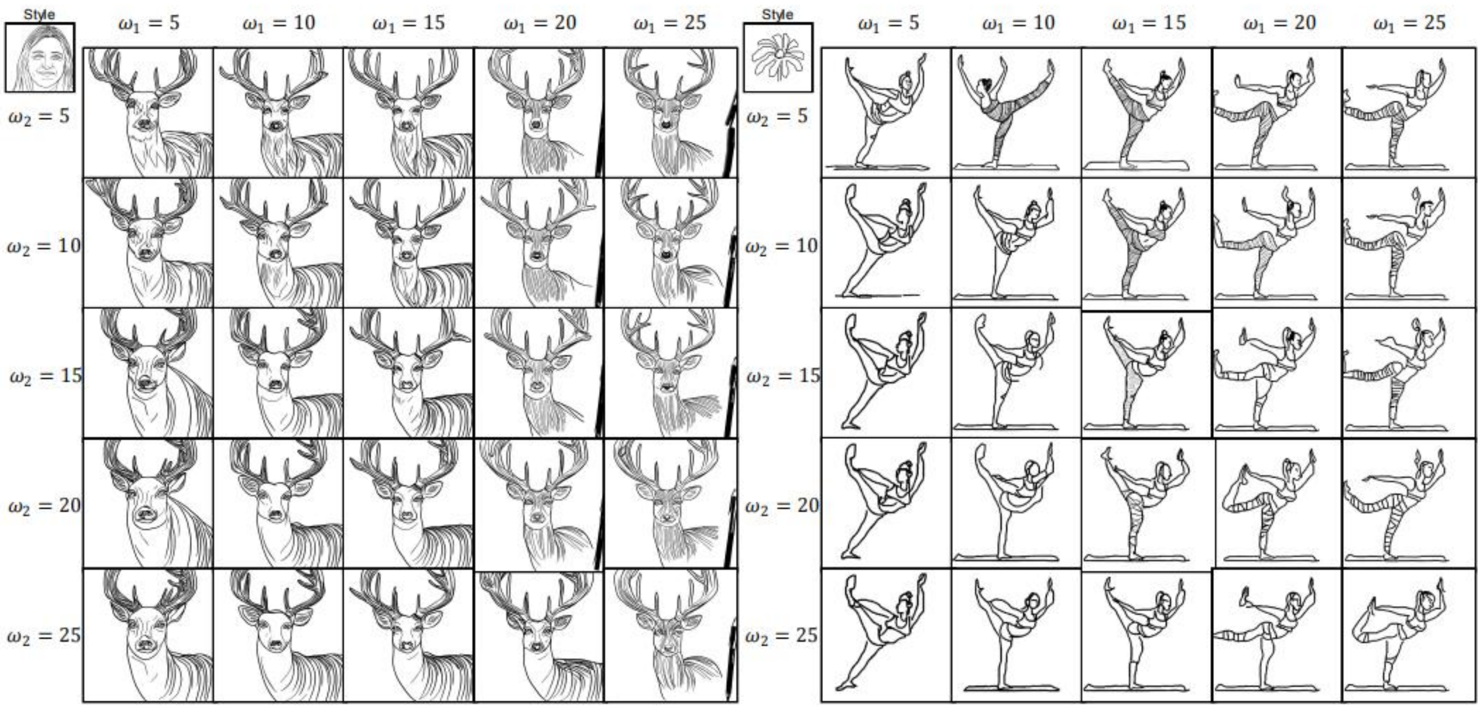


Figure 9: The generated results with different content guidance scale ω_1 and style guidance scale ω_2. Left: "a sketch of a deer". Right: "a sketch of a person doing yoga".

Limitations & Future Work

Current Limitations:

- Challenges with extremely sparse references

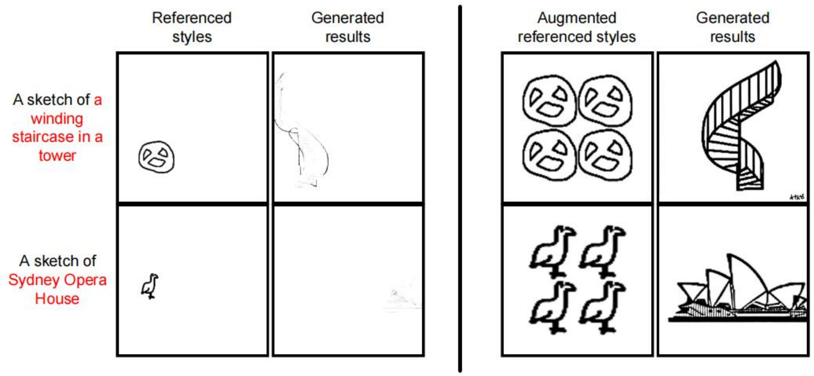


Figure 15: Left: Failure cases of M3S. When the referenced sketches are too small or sparse, M3S is difficult to produce meaningful results. Right: A potential resolution through image augmentation.

Future Directions:

- Localized style control for specific regions
- Enhanced handling of abstract sketches
- Real-time generation optimization

Conclusion & Resources

Summary:

- M3S enables training-free, controllable multi-style sketch generation

Contributions:

- Novel feature injection, adaptive style control, extensive validation

Availability:

- Code and models are open-sourced at https://github.com/CMACH508/M3S

Acknowledgement:

- Supported by Shanghai Municipal Science and Technology Major Project

END