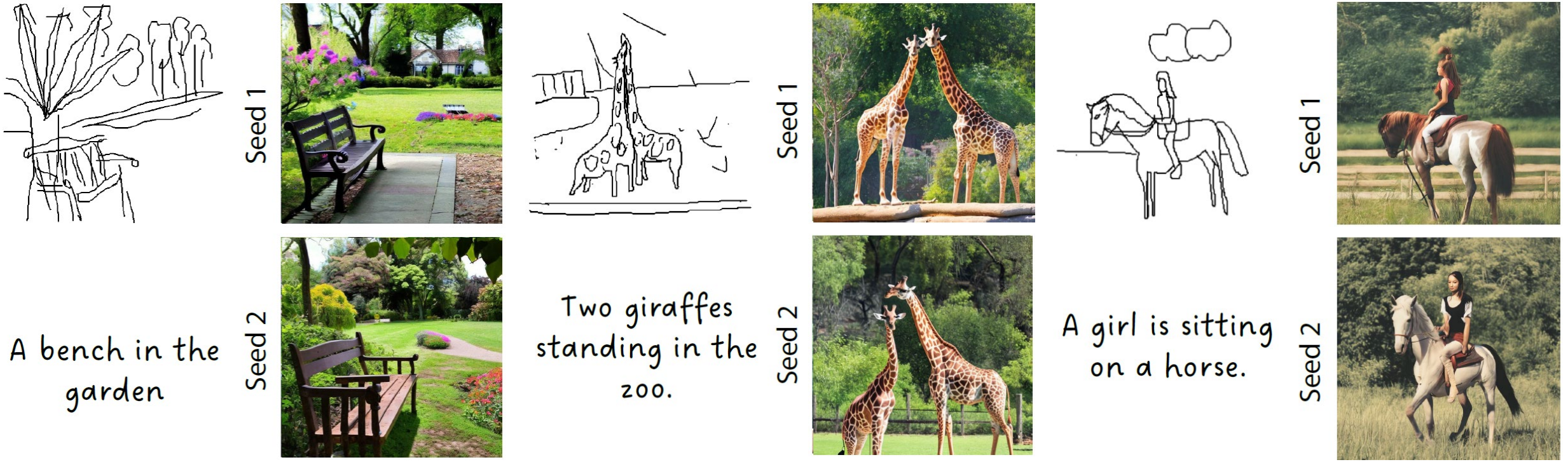
ICLR 2026

SketchingReality

From Freehand Scene Sketches to Photorealistic Images

Ahmed Bourouis[1] · Mikhail Bessmeltsev[2] · Yulia Gryaditskaya[1,3]

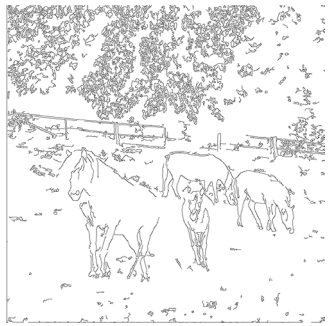
[1]University of Surrey [2]Université de Montréal [3]Adobe Research



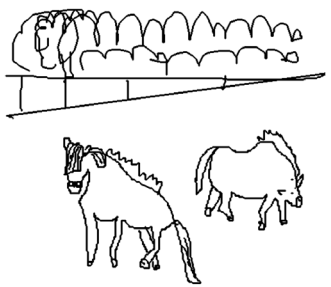
Motivation

Sketches are expressive. How to condition image generation on them?

Reference image



Edge map



Freehand sketch

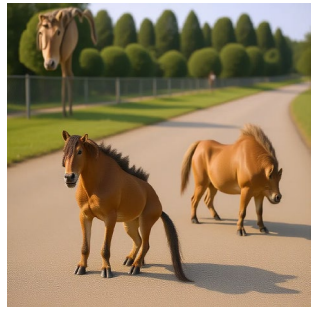
Challenge: Real freehand sketches are **abstract, distorted & not pixel aligned** like edge maps.

Existing methods fail
unrealistic or **sketch-ignoring** outputs.

ControlNet SD2.1

ControlNet SDXL

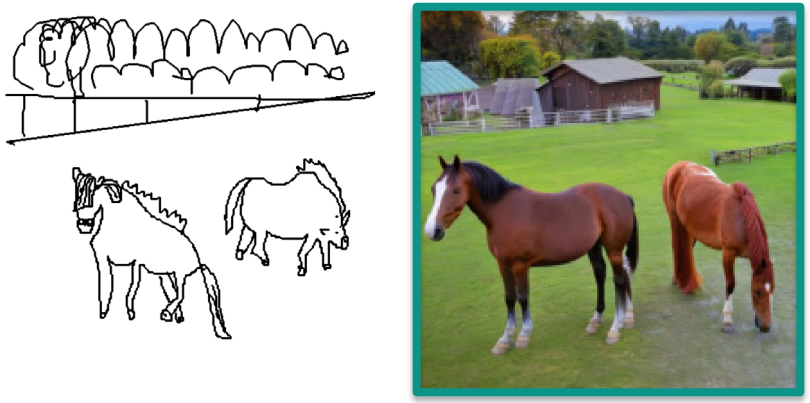
T2I Adapter



FLUX Kontext

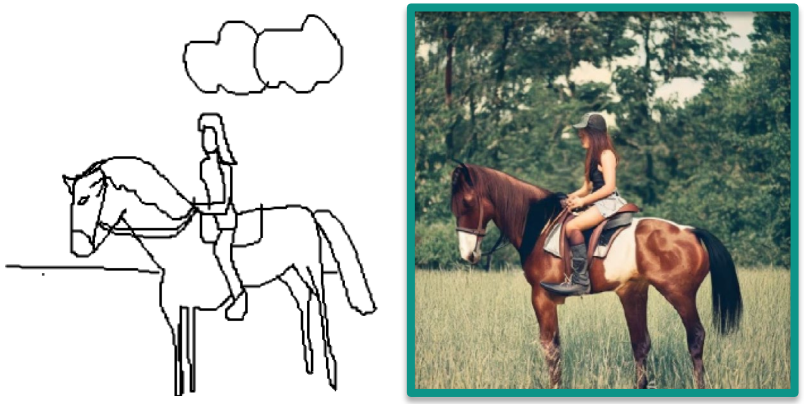
Key Insight

Interpret sketches **semantically**.
Understand ***what*** objects are and ***where*** they are.



"***Horses*** are in the ***garden*** area"

"A ***bench*** in the ***garden***"



"A ***girl*** is sitting on a ***horse***"

Our Solution

① Semantic Sketch Encoder

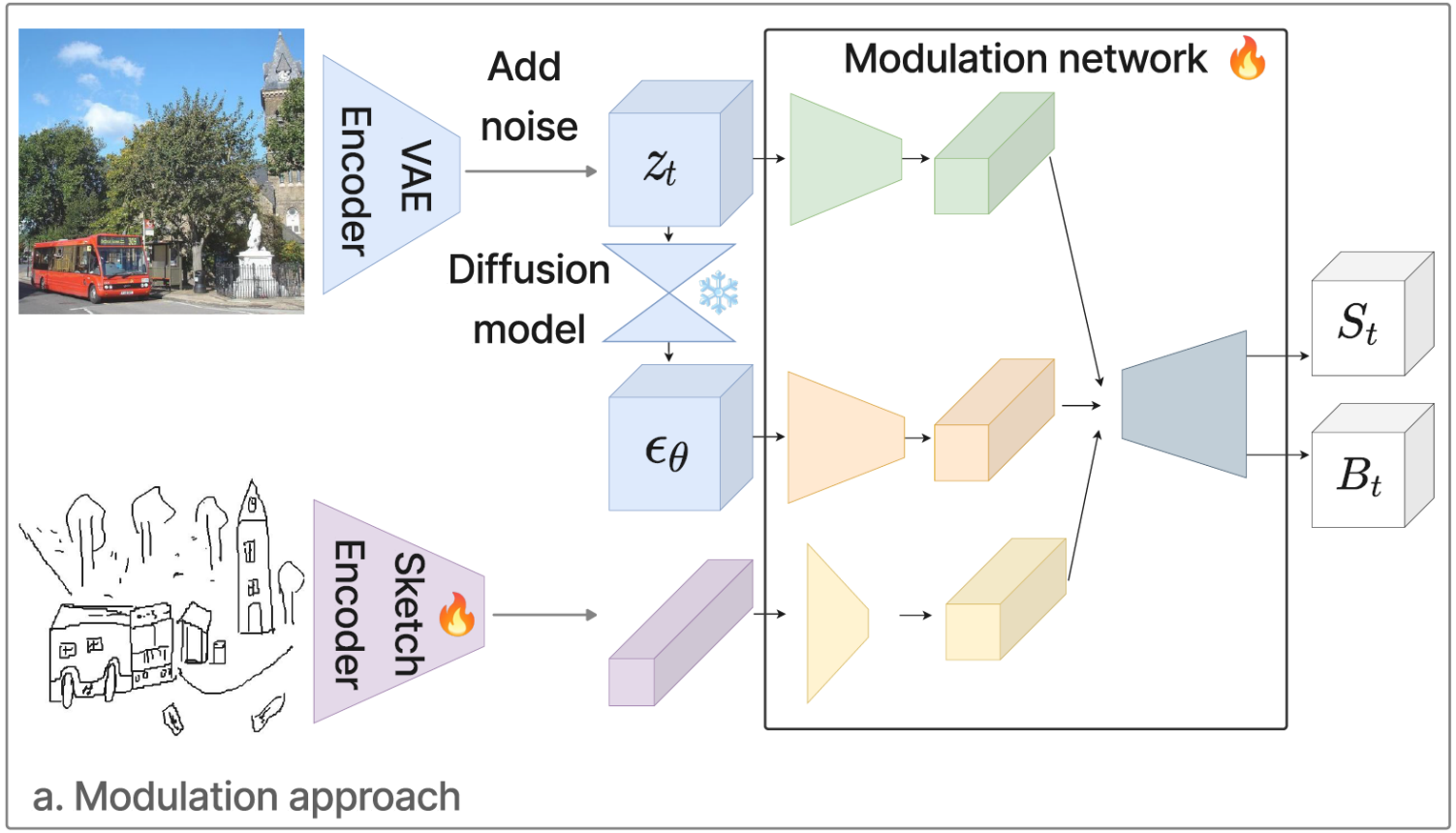
Captures ***what*** and ***where*** from abstract and distorted scene sketches.

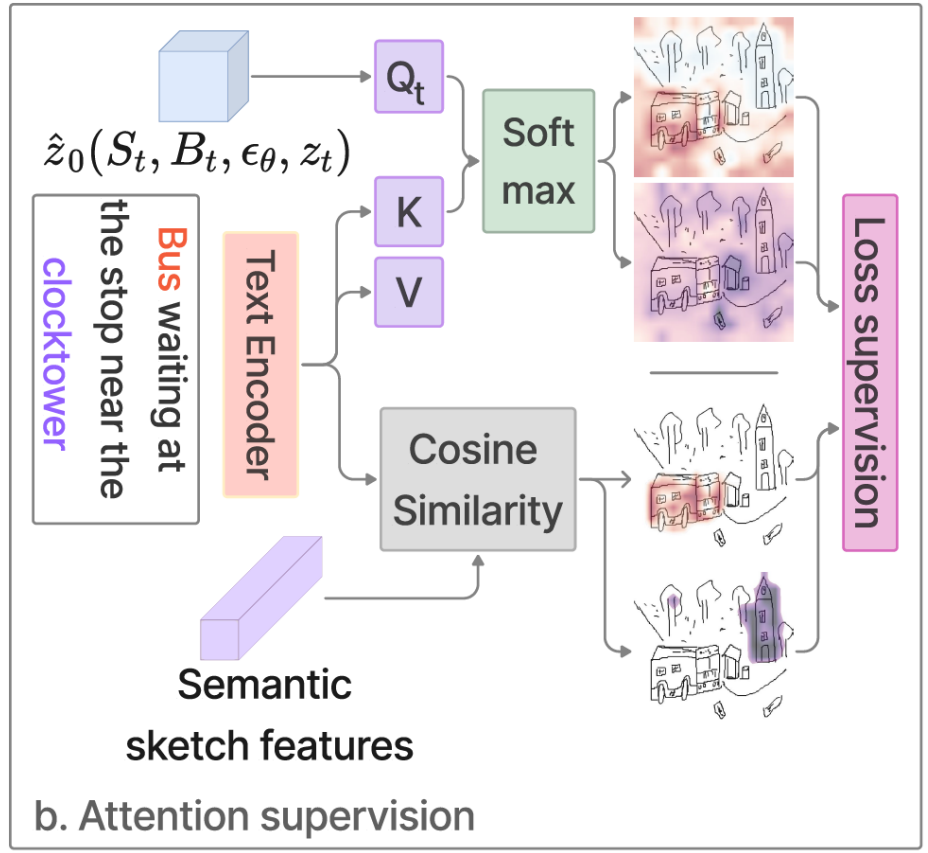
② Modulation Network

Injects ***sketch semantics*** via scale & shift, no backbone changes.

③ Attention Supervision Loss

Guides cross-attention using sketch semantics to ***align*** output layout semantically without pixel-aligned GT.





Our Solution

① Semantic Sketch Encoder

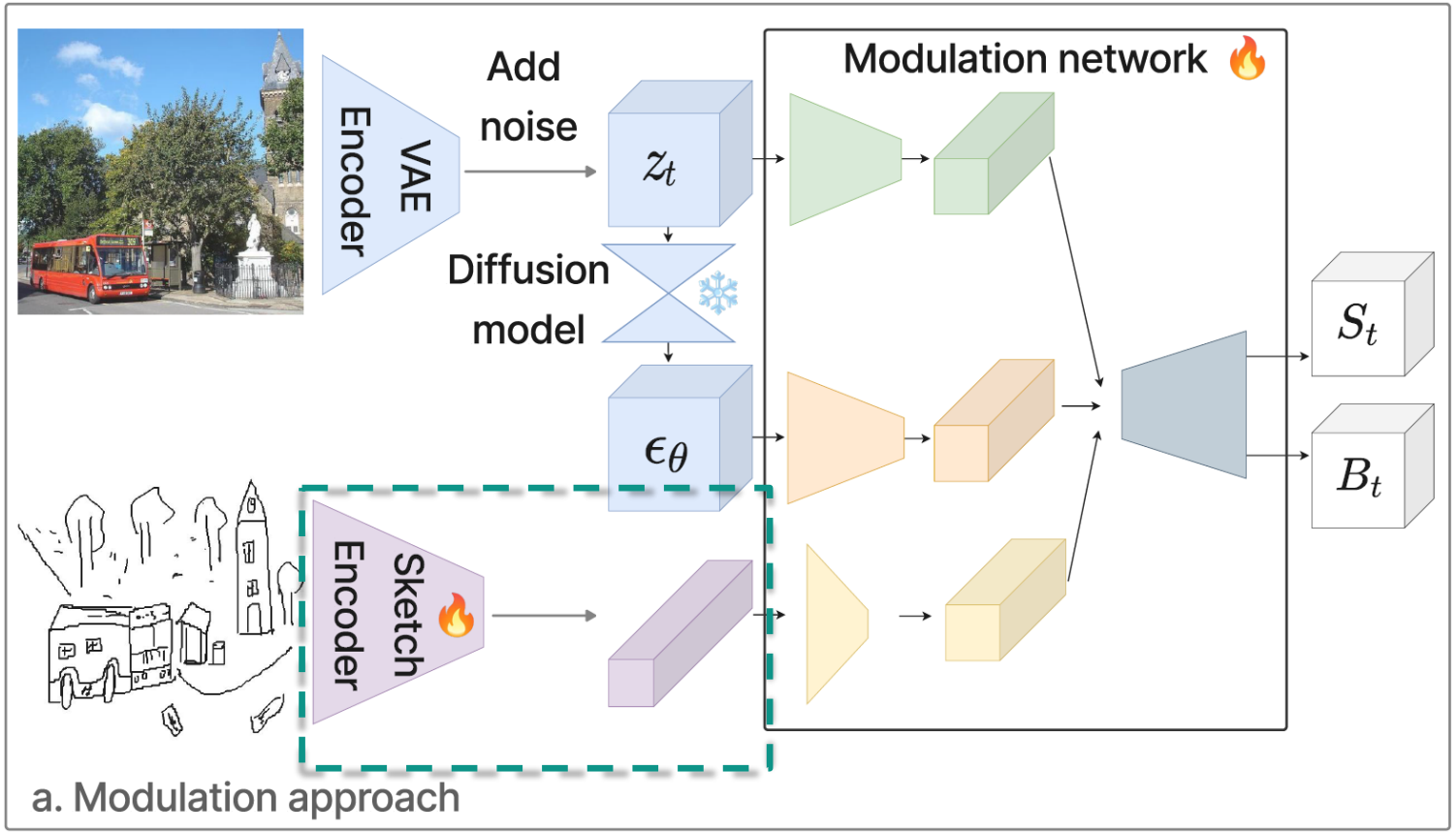
Captures ***what*** and ***where*** from abstract and distorted scene sketches.

② Modulation Network

Injects ***sketch semantics*** via scale & shift, no backbone changes.

③ Attention Supervision Loss

Guides cross-attention using sketch semantics to ***align*** output layout semantically without pixel-aligned GT.



Our Solution

① Semantic Sketch Encoder

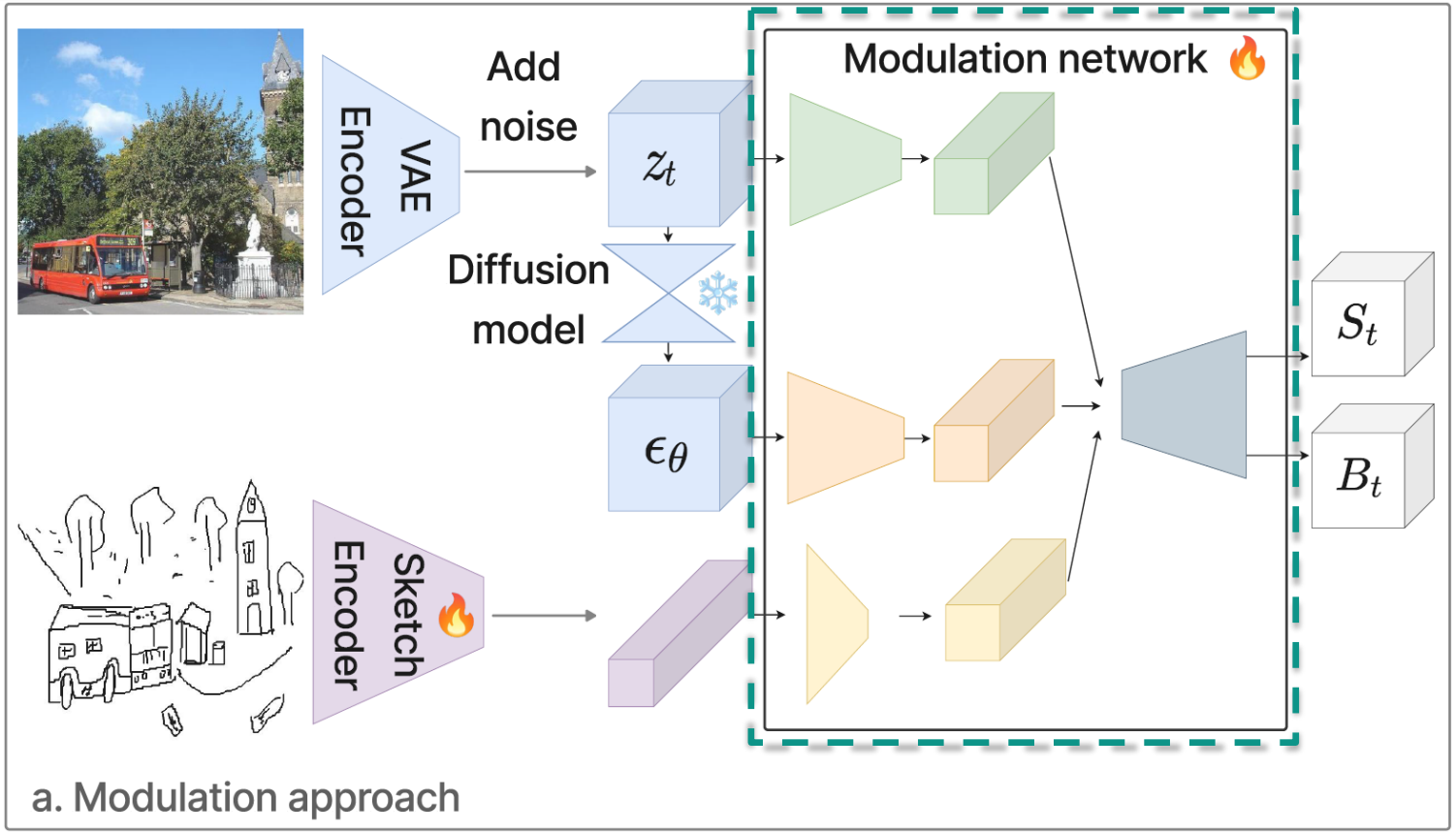
Captures ***what*** and ***where*** from abstract and distorted scene sketches.

② **Modulation Network**

Injects ***sketch semantics*** via scale & shift, no backbone changes.

③ Attention Supervision Loss

Guides cross-attention using sketch semantics to ***align*** output layout semantically without pixel-aligned GT.



Our Solution

① Semantic Sketch Encoder

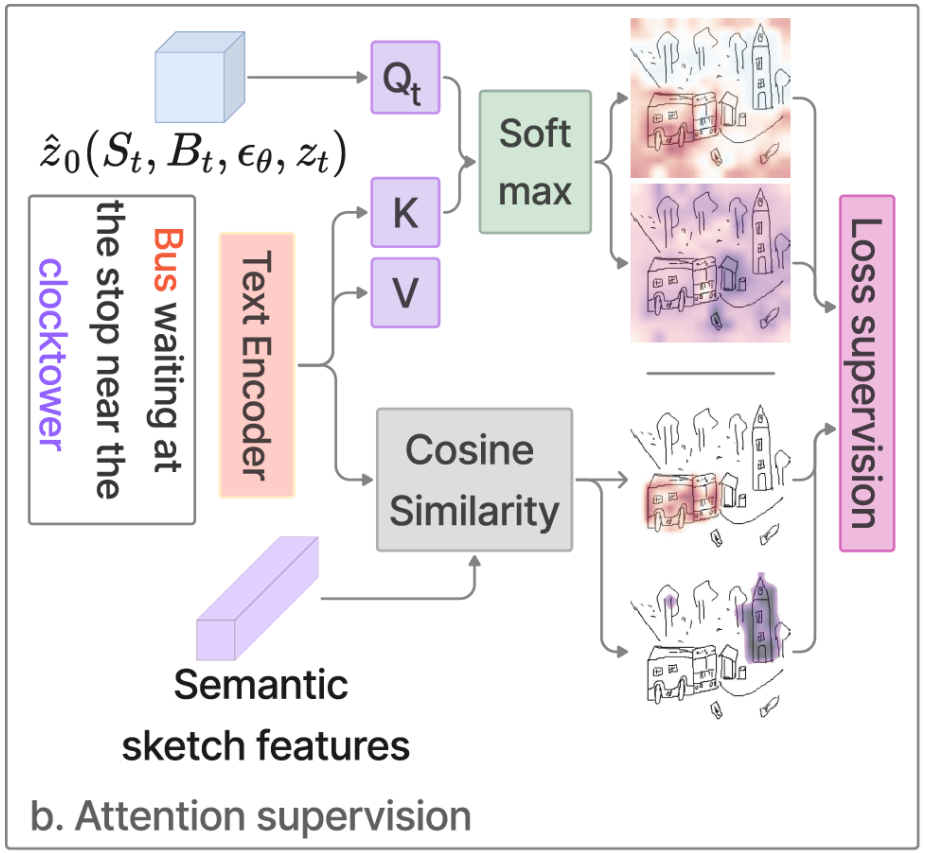
Captures ***what*** and ***where*** from abstract and distorted scene sketches.

② Modulation Network

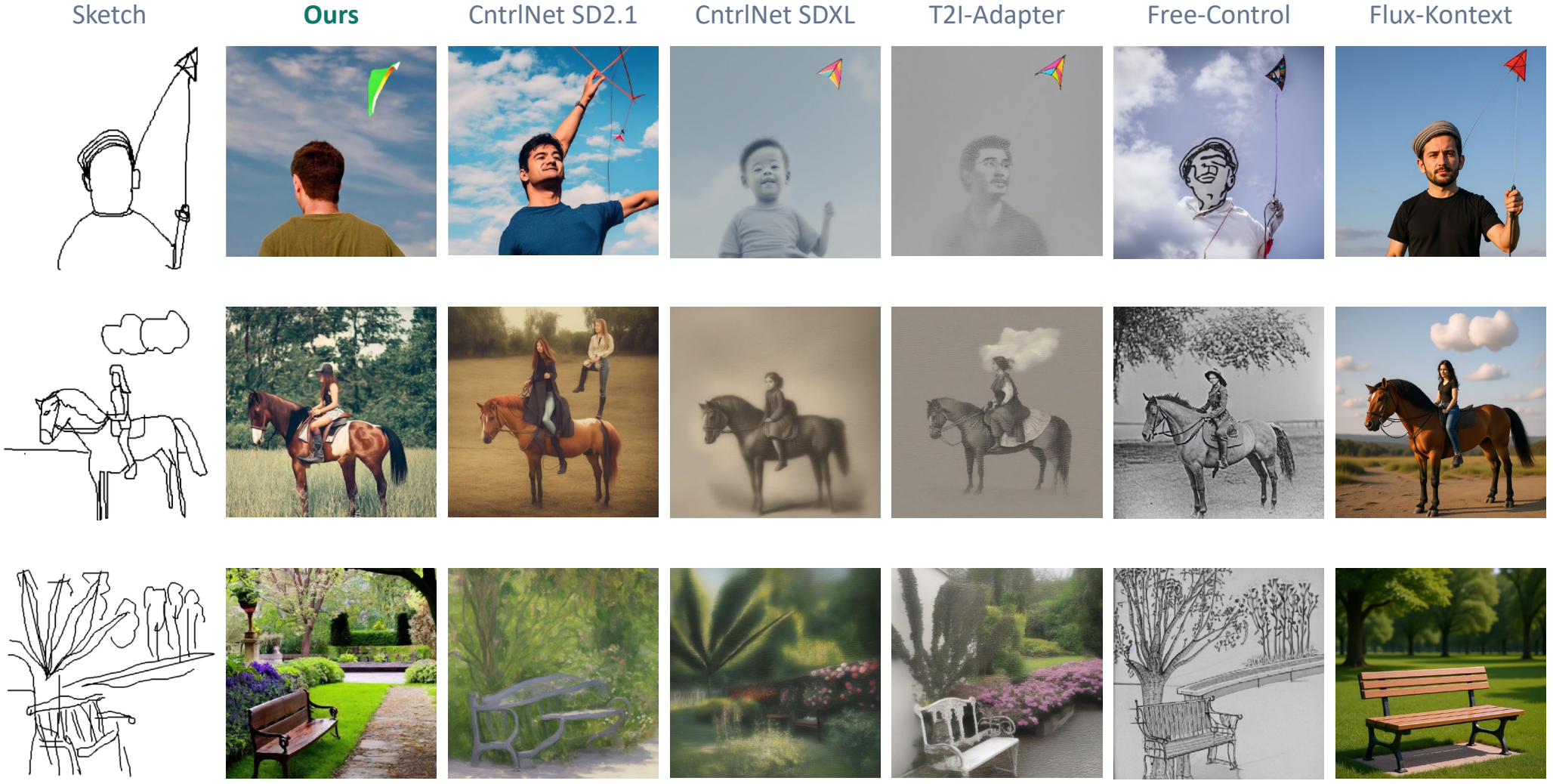
Injects ***sketch semantics*** via scale & shift, no backbone changes.

③ **Attention Supervision Loss**

Guides cross-attention using sketch semantics to ***align*** output layout semantically without pixel-aligned GT.

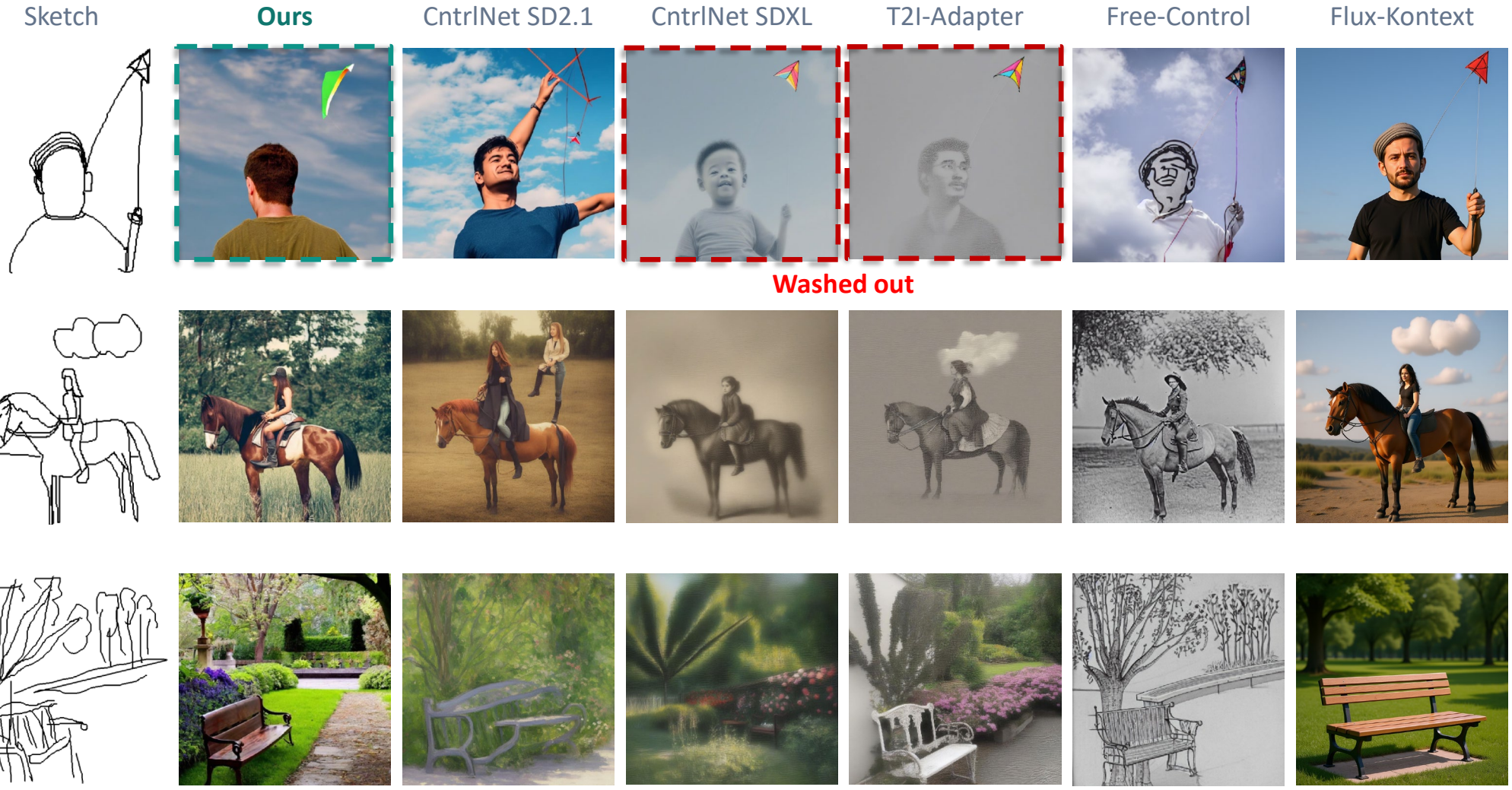


Visual Comparisons



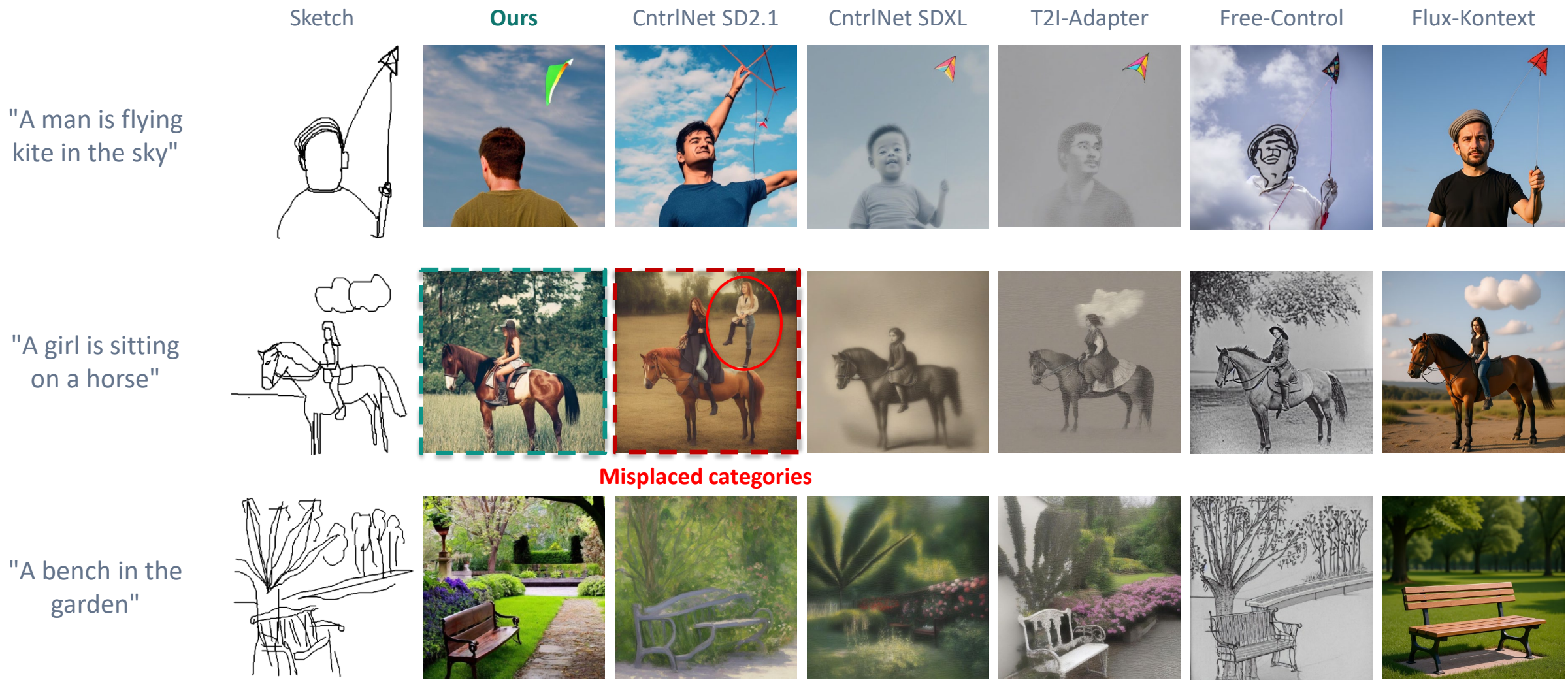
Freehand sketches from FS-COCO. Our method produces photorealistic images that follow the sketch layout.

Visual Comparisons



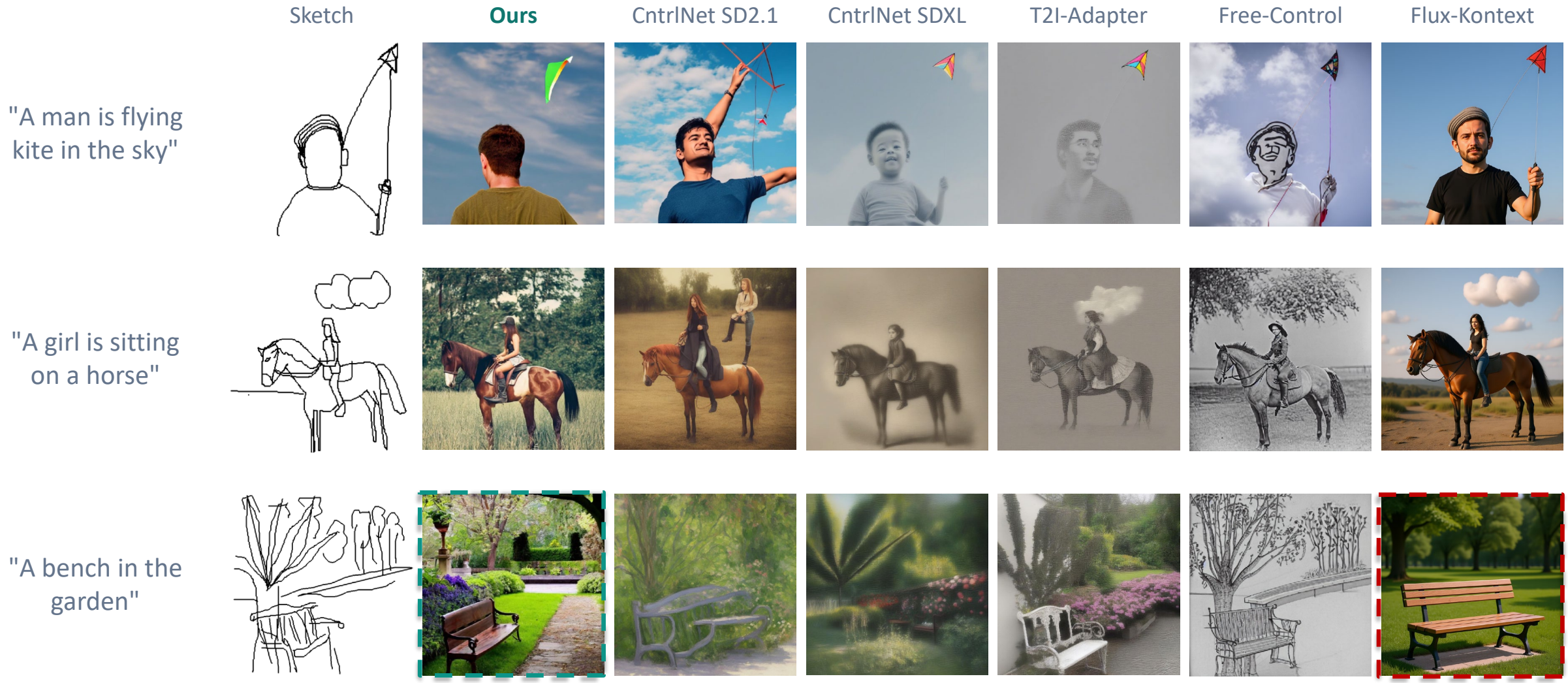
Freehand sketches from FS-COCO. Our method produces photorealistic images that follow the sketch layout.

Visual Comparisons



Freehand sketches from FS-COCO. Our method produces photorealistic images that follow the sketch layout.

Visual Comparisons



Freehand sketches from FS-COCO. Our method produces photorealistic images that follow the sketch layout.

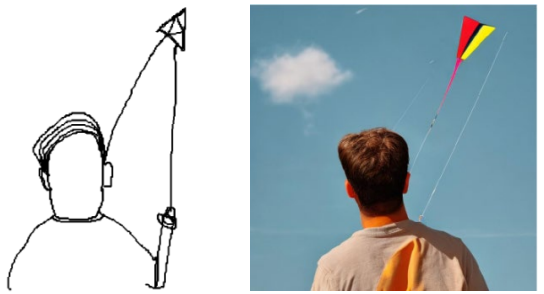
Quantitative Results

Method	Setup	FID↓	CLIP↑	LPIPS↓
CntrlNet SD2.1 (Zhang et al., 2023)	Zero-shot	135.595	1.136	0.773
	$\mathcal{L}_{\text{noise}}$ only (syn)	136.821	1.141	0.771
	$\mathcal{L}_{\text{noise}}$ only (free)	139.872	1.042	0.789
	$\mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{attn}}$	135.891	1.196	0.768
T2I Adapter $s = 0.8\ \tau = 0.2$ (Mou et al., 2023)	Zero-shot	144.329	-0.203	0.813
	$\mathcal{L}_{\text{noise}}$ only	138.479	0.318	0.779
	$\mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{attn}}$	137.982	0.391	0.774
T2I Adapter $s = 0.8\ \tau = 0.4$ (Mou et al., 2023)	Zero-shot	159.816	0.213	0.819
	$\mathcal{L}_{\text{noise}}$ only	151.736	0.426	0.781
	$\mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{attn}}$	139.568	0.454	0.778
CntrlNet SDXL (Zhang et al., 2023)	Zero-shot	174.462	0.027	0.825
CntrlNext SDXL (Peng et al., 2024)	Zero-shot	134.094	0.909	0.774
SG (Voynov et al., 2023)	Zero-shot	137.381	1.043	0.782
FreeControl (Mo et al., 2024)	Inference-time	141.632	1.089	0.793
Ours	Full	**121.973**	**1.291**	**0.739**

More Results

"Train going on the track"

"Two girls are playing with frisbee disc"

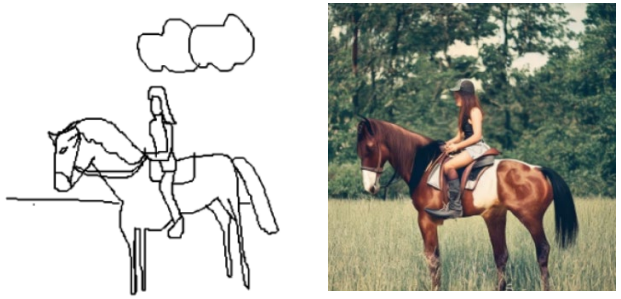


"A man is flying kite in the sky"

"A bench in the garden"

"Airplane is standing on the airport"

"Three people walking holding umbrellas"



"A girl is sitting on a horse"

"A building with a clock on it"

ICLR 2026

Project Page | Paper