LOTS of Fashion!

Multi-Conditioning for Image Generation via Sketch-Text Pairing

Federico Girella, Davide Talon, Ziyue Liu, Zanxi Ruan, Yiming Wang, Marco Cristani

ICCV 2025 Oral

Presenter: XuShenghan
2025.10.19

LOTS of Fashion: Task Description

In fashion design, designers need to express their abstract inspirations through forms that are natural to humans, e.g., sketches or natural language.

LOTS of Fashion: Task Description


效果图
设计元素: □民族 □军警 □简约 □前卫 □乡村
□都市 □传统 □经典 ☑优雅 □其它
面布: ☑已到 □未到,预计
辅料: ☑已到 □未到,预计
☑弹性 □无弹性
□补款 □套款 □双面呢
□中华情 □重工款 □陈列款 □客人特定
□04.18 □05.09 □05.23 ☑06.06 □06.11
☑头版 □套色版
贴花边
隐拉

款号:	款式:
设计:	提交日期: 2018.5.24
纸样:	车位号:
审核:	完成日期: 2018.5.29
尺寸:	袖长: 6分
衣长:(后中)	裙长:
洋装长:(后中) 及膝	膝围:
裤长:	脚围:
肩棉:	不需要□

工艺说明: 打版前请与设计师沟通,谢谢!

工艺处理:
印花( ) 钉珠(√) 绣花( ) 车花(√)
对丝( ) 烫钻( ) 染色( ) 压褶( )
吊染/扎染( )

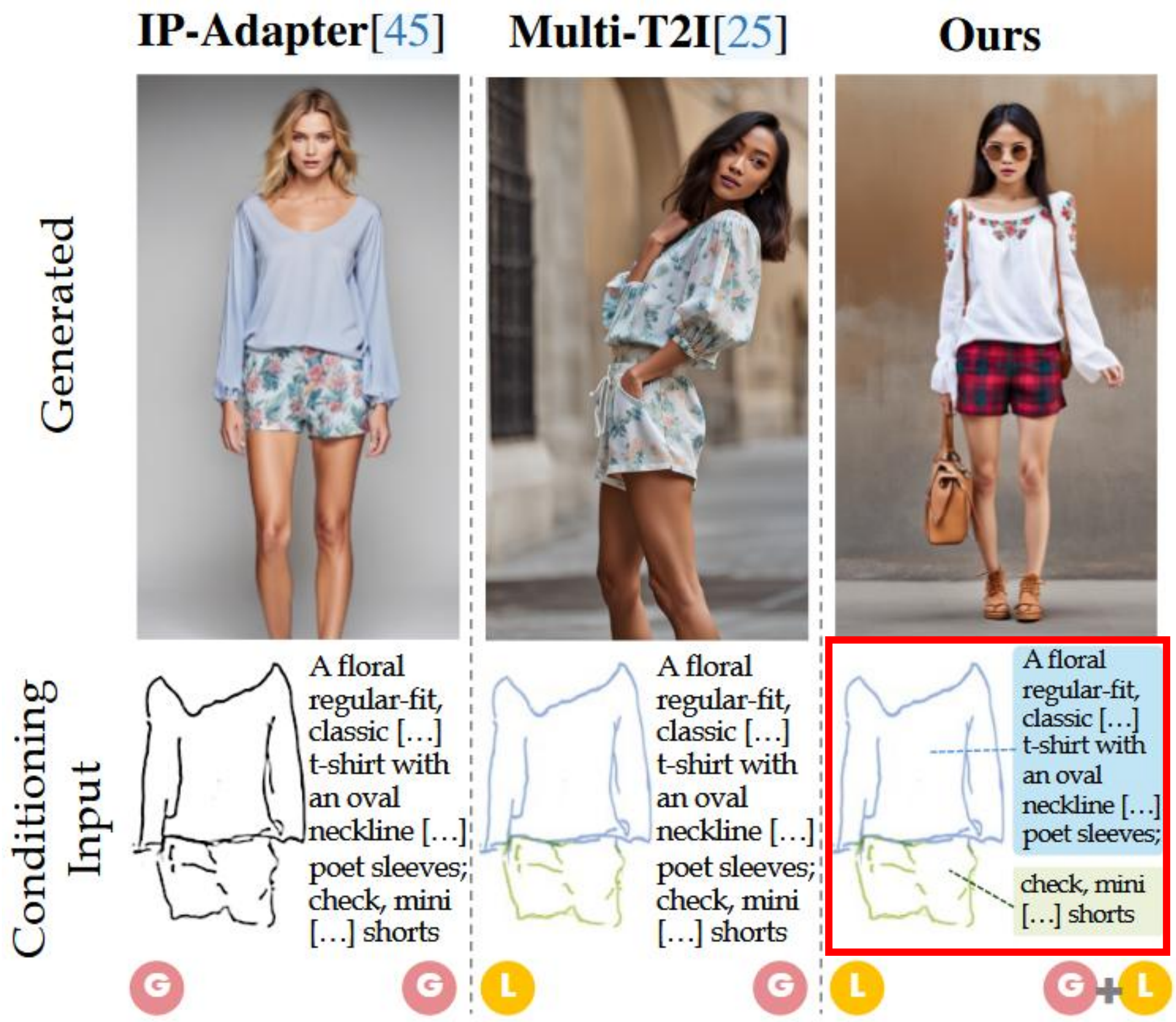
辅料版样及说明

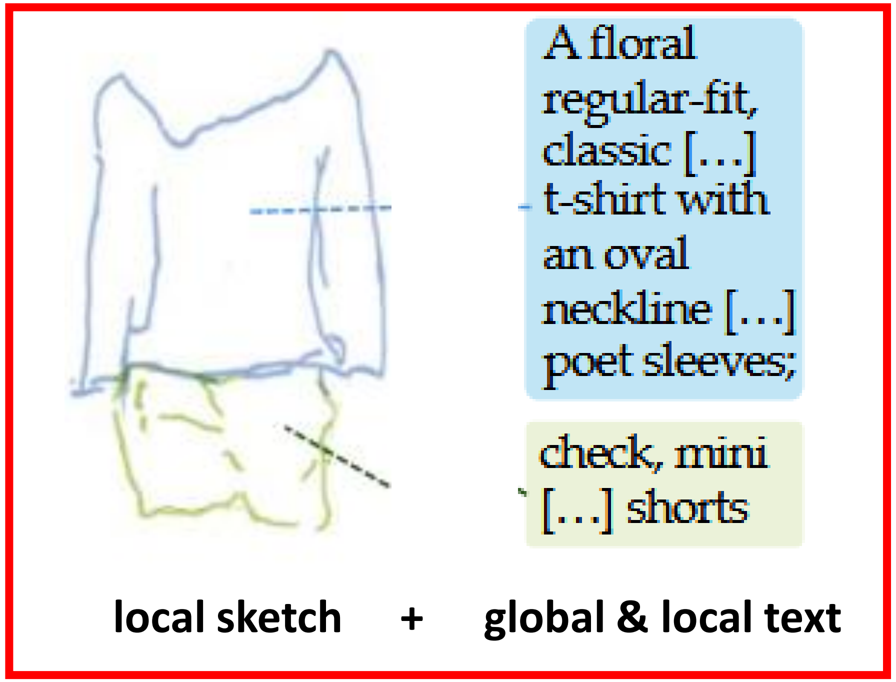
腰带/饰品	厂家:	编号:
	型号:	颜色:
	数量:	位置:
纽扣	厂家:	编号:
	型号:	颜色:
	数量:	位置:
拉链	厂家:	编号:
	型号:	颜色: 配色隐拉
	数量:	位置:
花边/织带	厂家:	编号:
	型号:	颜色:
	数量:	位置:
其它	□风衣链 白呖色□ 枪色□ 金色□	☑对钩 白呖色☑ 枪色□ 金色□

主料-A	样版	配布-B	样版
配布-C	样版	里布	样版

Multiple sketch-text pair is essential in describing a complete fashion design.
Each description pair specifies a localized part of the design, in terms of silhouette shapes, materials, and textual details, allowing fine-grained localized control over the generation.

LOTS of Fashion: Task Description

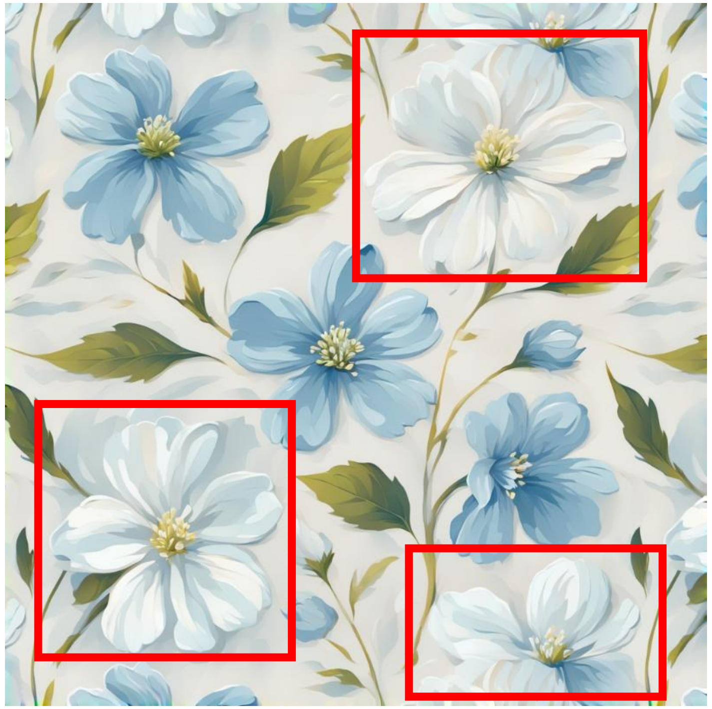




LOTS of Fashion: Task Description

It seems like ControlNet already provided a complete process on Multi-Conditioning for Image Generation. However, prior works mainly focus on global control rather than localized controlling via various forms of information.

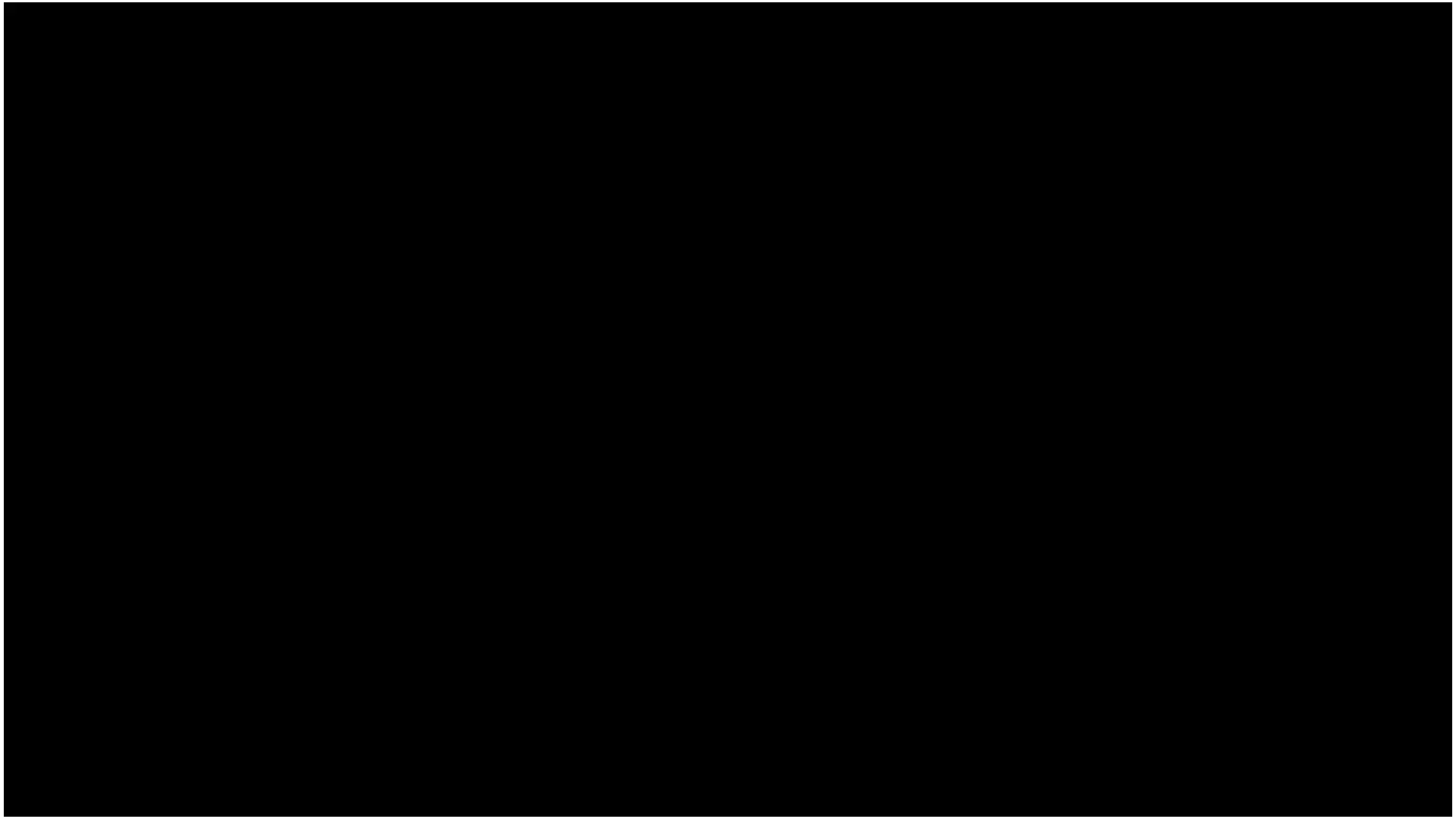
blue flower.



blue flower. green leaves. white background.

LOTS of Fashion: Task Description

北京大學 PEKING UNIVERSITY



By using localized sketch-text pairing input, generate overall harmonious and detailed fashion images.

Background

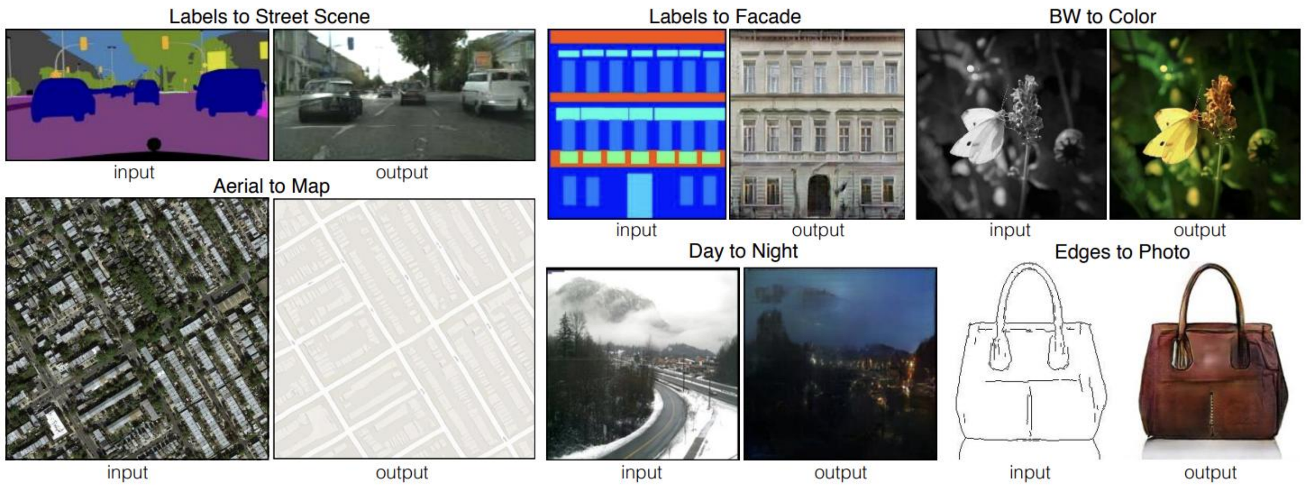
In order to achieve Multi-Conditioning for Image Generation via Sketch-Text Pairing, we need:

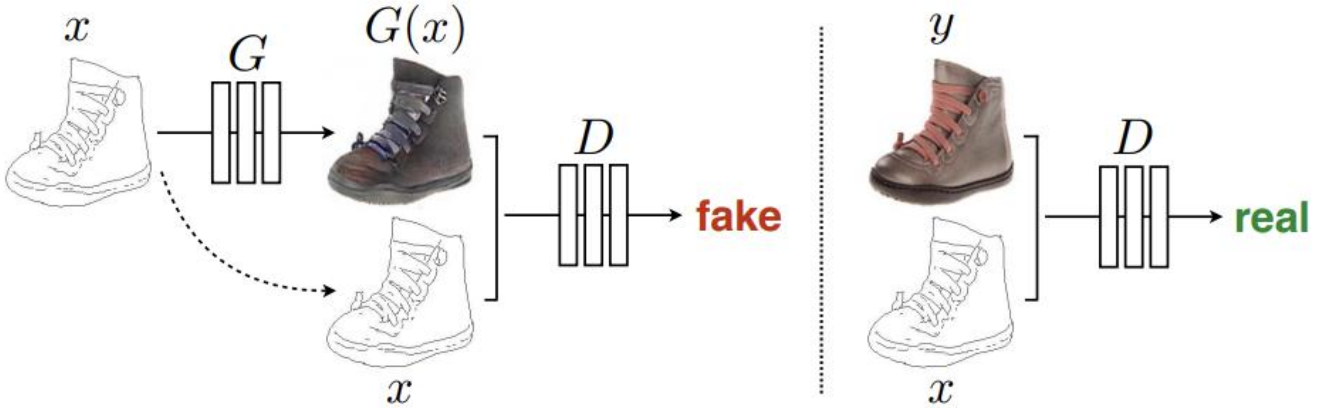
1. Text-to-Image Generation

2. Sketch-to-Image Generation

3. Controllable diffusion-based generation

Background: GAN based Sketch-to-Image Generation





Phillip Isola, Junyan Zhu et al., Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017

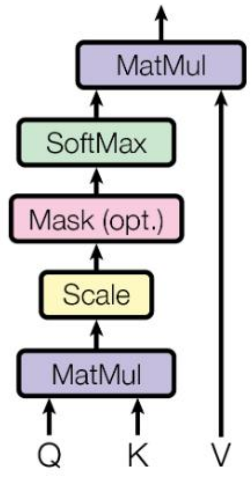
Background: Attention Mechanism

Calculate similarity in QK^T, and weight sum using V.

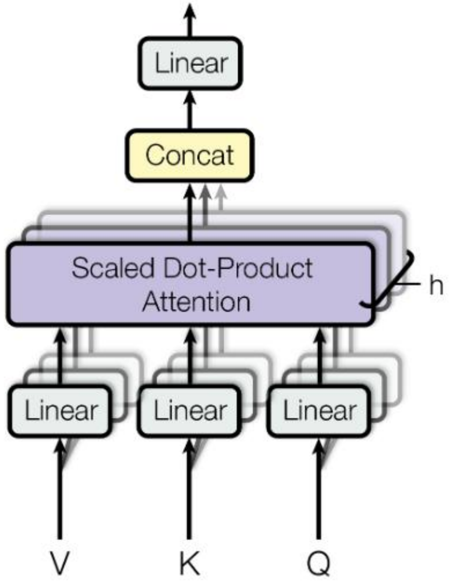
- Long-range dependency and dynamic weight
- Global information capturing

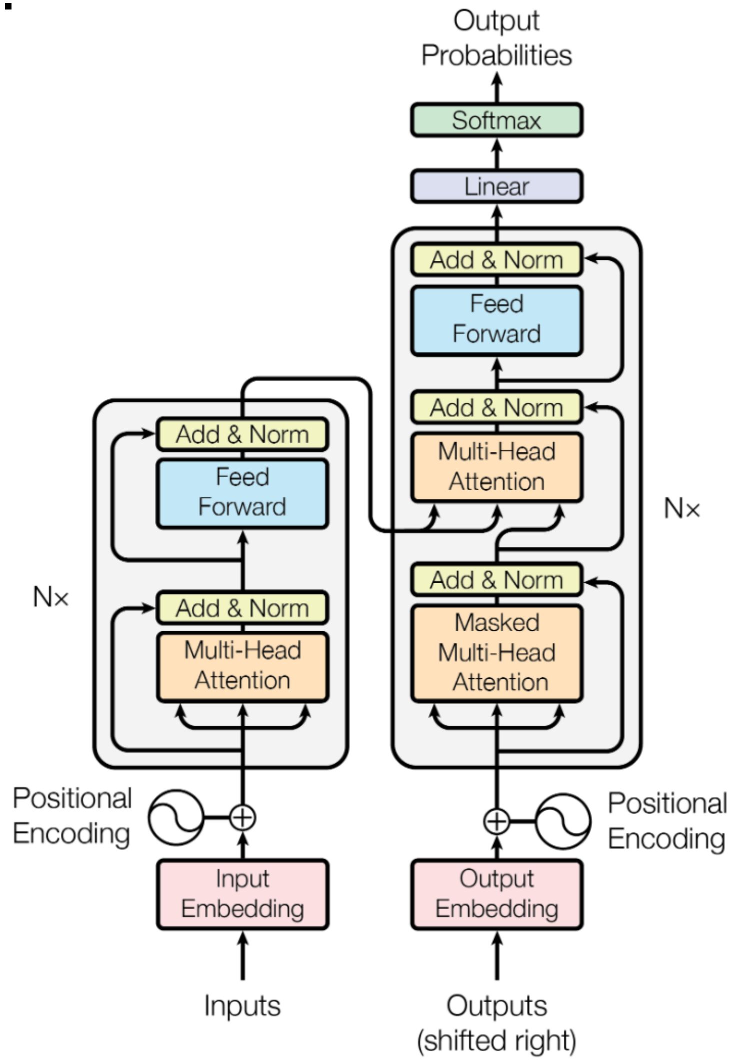
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})\,V$$

Scaled Dot-Product Attention



Multi-Head Attention



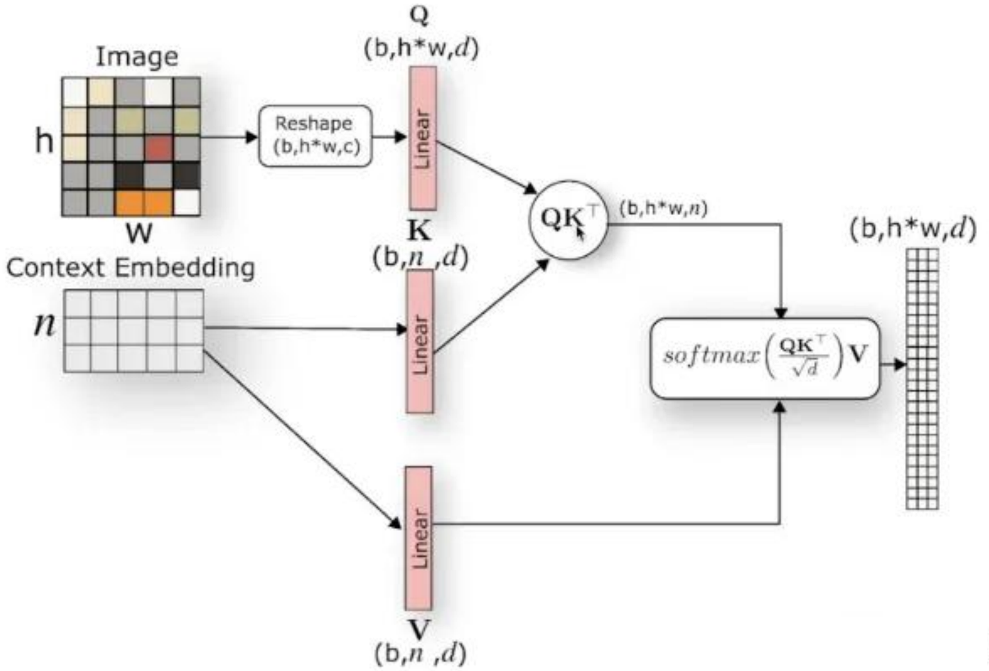


Background: Cross Attention

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

$$Q \in \mathbb{R}^{m \times d_k}, K \in \mathbb{R}^{n \times d_k}, V \in \mathbb{R}^{n \times d_v}$$

Q comes from the image, while K and V come from the conditional control.



Q : Specifies the image’s **structure and layout**

K : **Compact** representation of the generated image

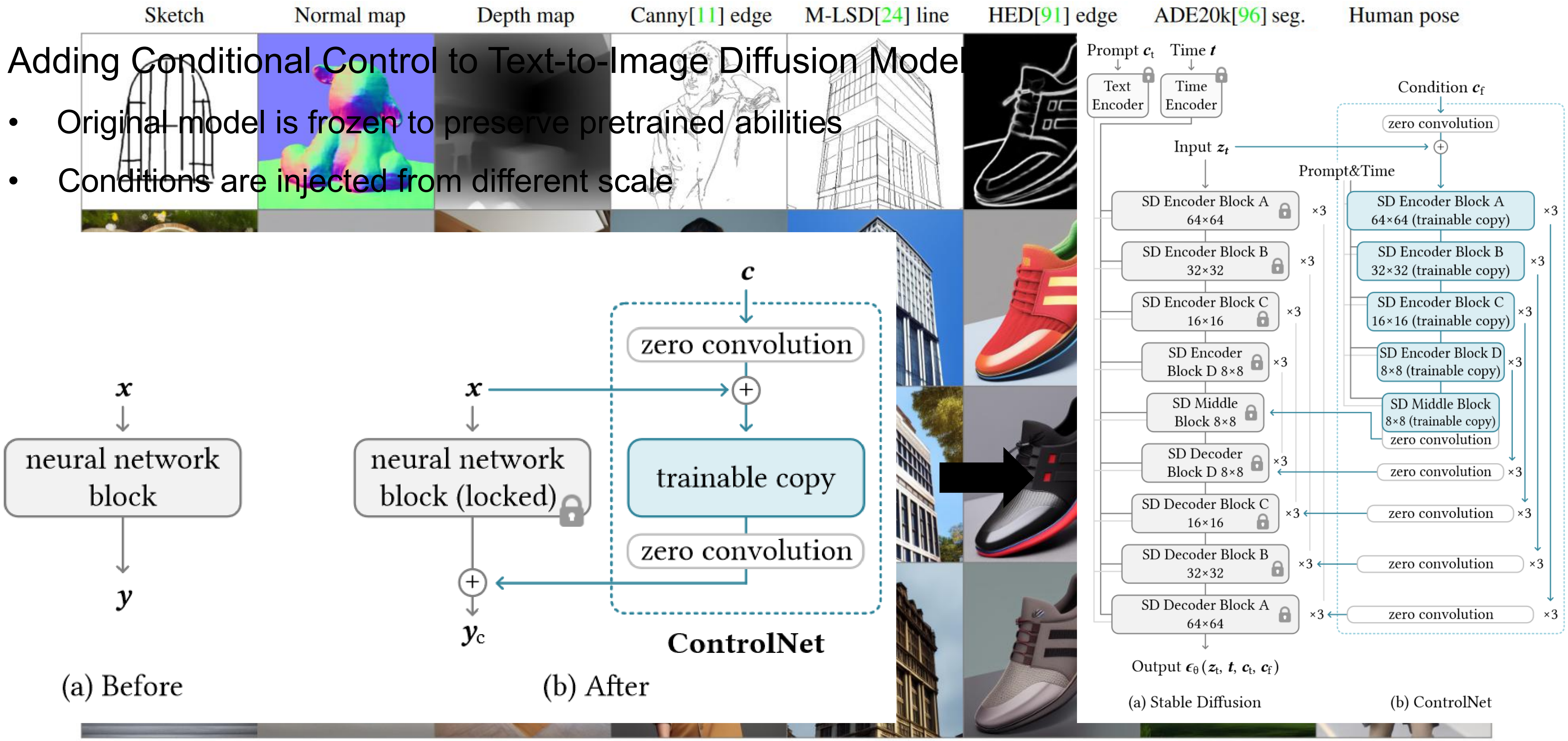
V : Injects **detailed appearance** information into the output

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar AverbuchElor, and Daniel Cohen-Or. Cross-image attention for zeroshot appearance transfer. In ACM SIGGRAPH, 2024.
Mingdeng Cao. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In IEEE International Conference on Computer Vision (ICCV), 2023.

Background: ControlNet

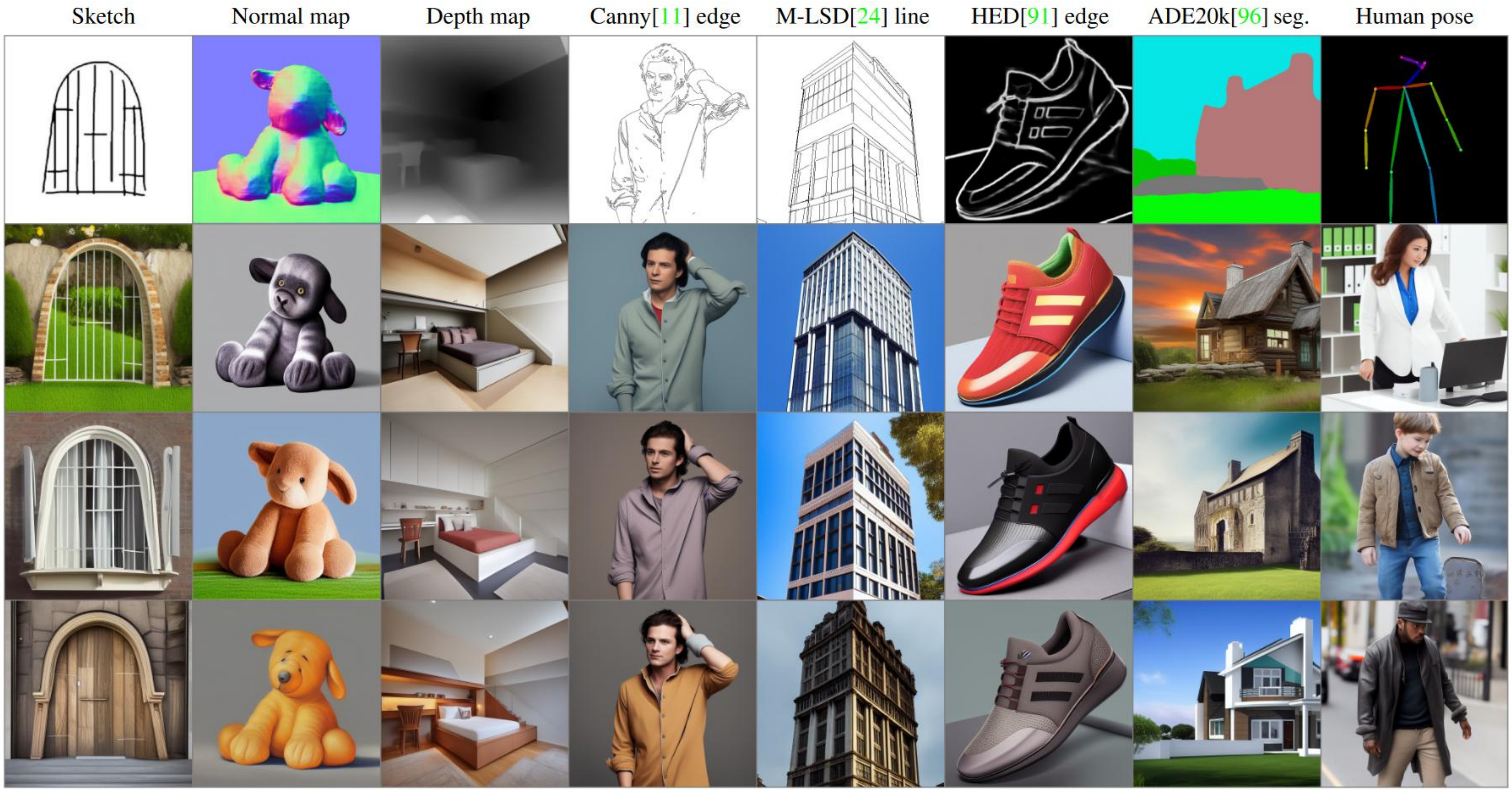
Adding Conditional Control to Text-to-Image Diffusion Model

- Original model is frozen to preserve pretrained abilities
- Conditions are injected from different scale



Lvmin Zhang, Anyi Rao, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In International Conference on Computer Vision (ICCV), 2023.

Background: ControlNet

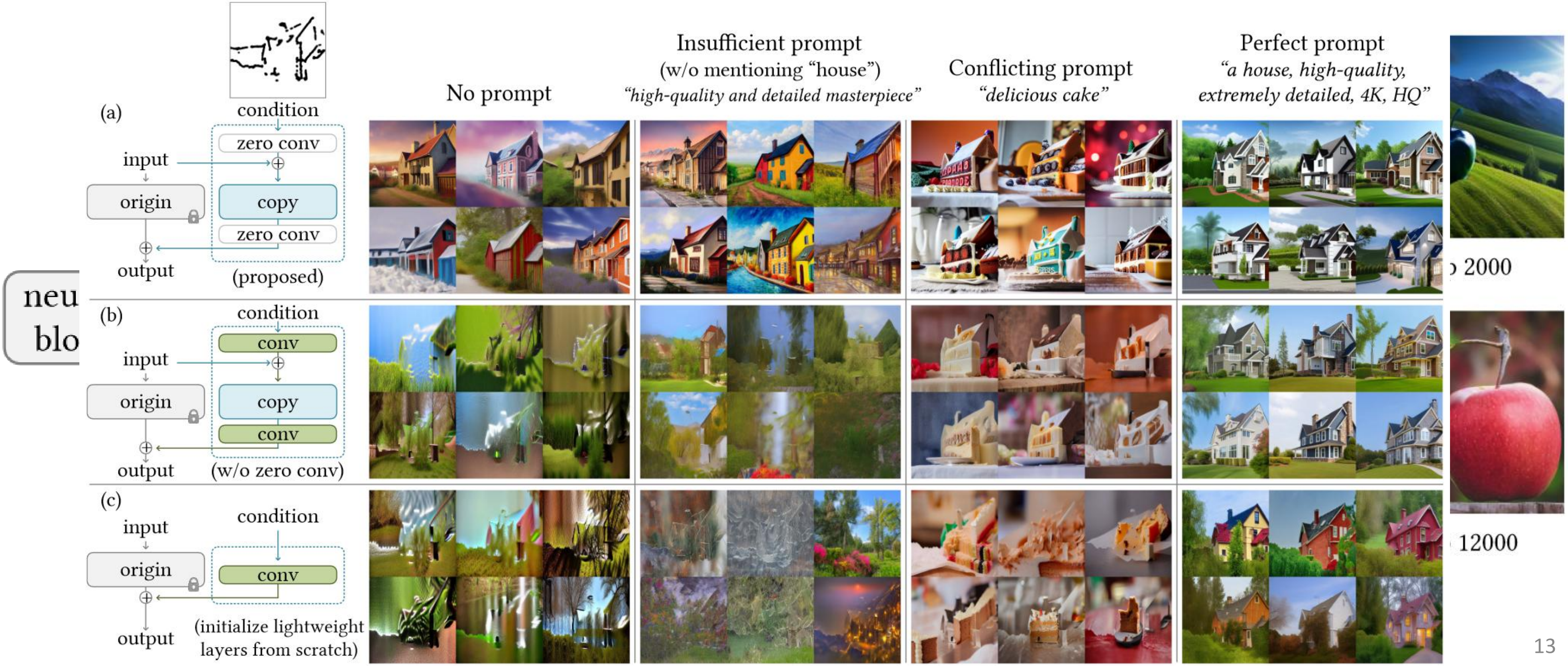


Lvmin Zhang, Anyi Rao, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In International Conference on Computer Vision (ICCV), 2023.

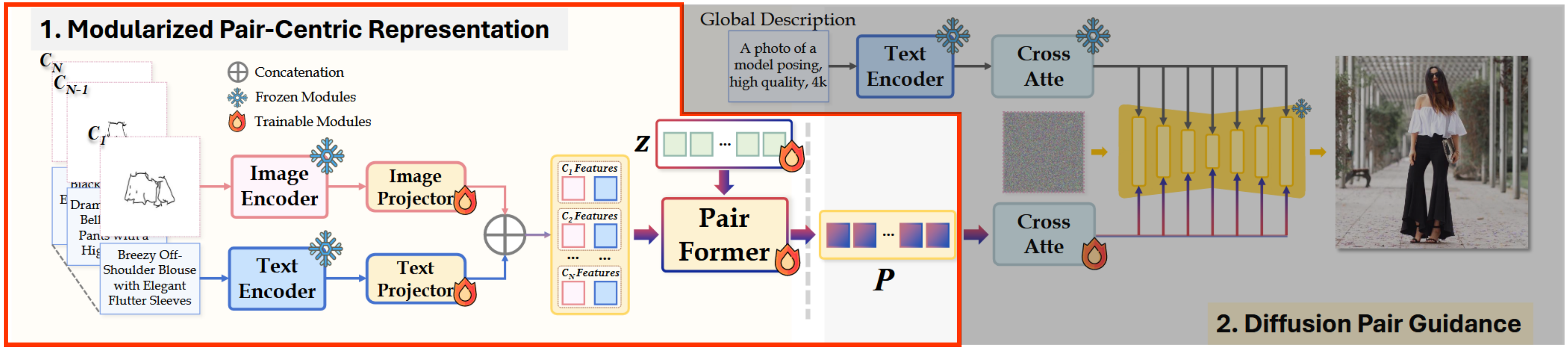
Background: ControlNet

Due to the **zero convolutions**, ControlNet always predicts high-quality images during the entire training.

At a certain step in the training process, the model suddenly learns to follow the input condition.

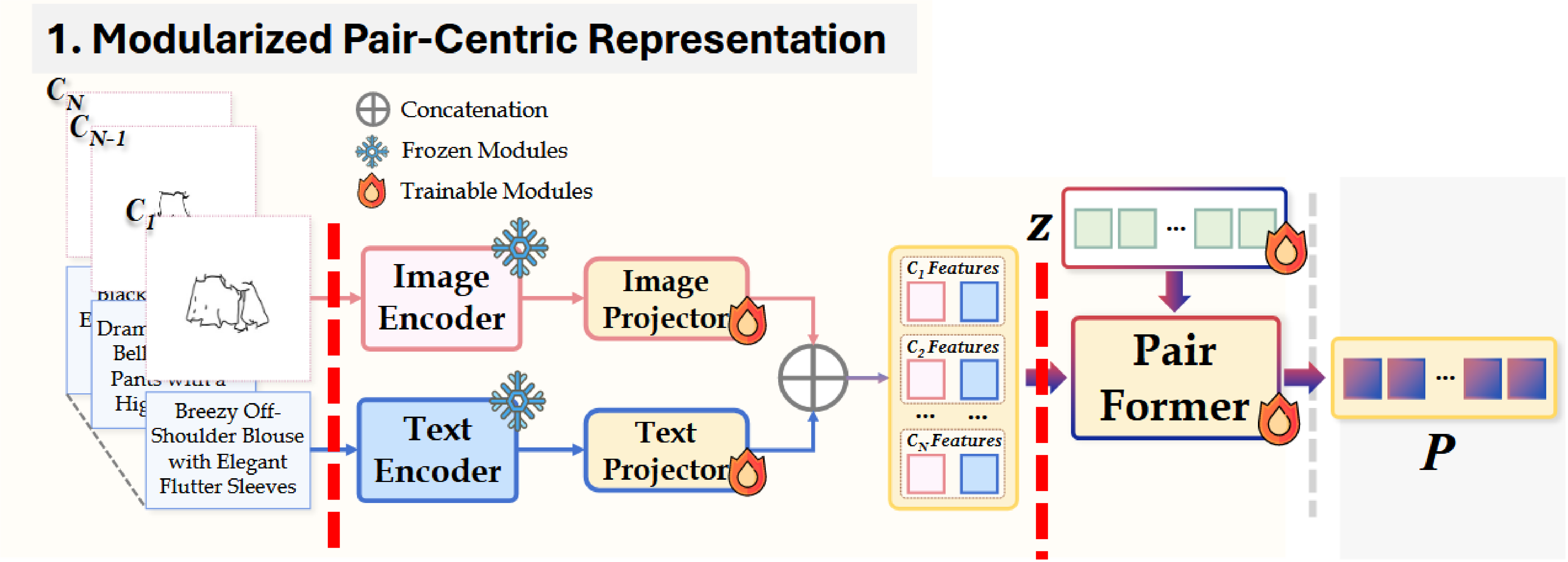


LOTS Method



1. Modularized Pair-Centric Representation
C_N
C_{N-1}
C_1
Breezy Off-Shoulder Blouse with Elegant Flutter Sleeves
Concatenation
Frozen Modules
Trainable Modules
Image Encoder
Image Projector
Text Encoder
Text Projector
C_1 Features
C_2 Features
C_N Features
Z
Pair Former
P
Global Description
A photo of a model posing, high quality, 4k
Text Encoder
Cross Atte
Cross Atte
2. Diffusion Pair Guidance

LOTS Method: Modularized Pair-Centric Representation

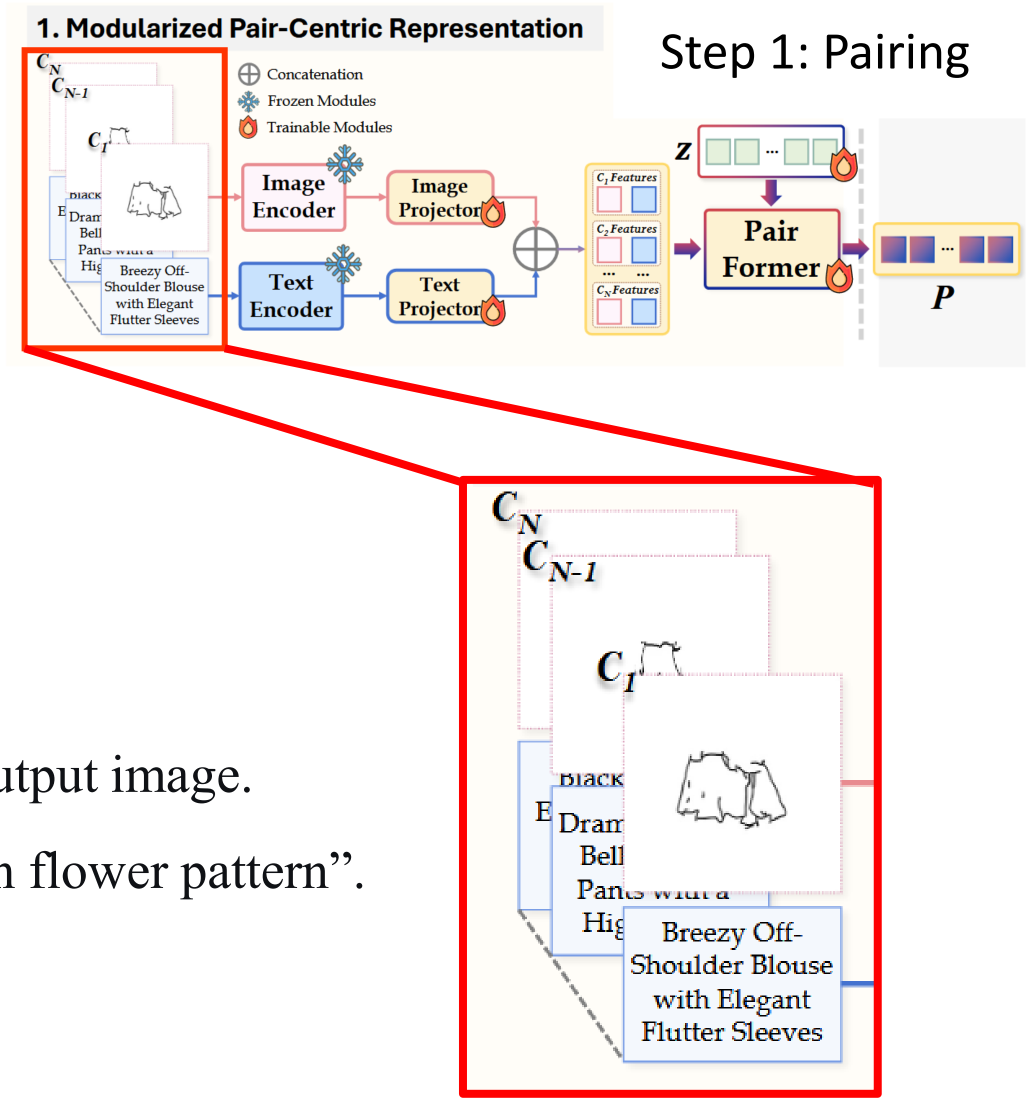


Step 1: Pairing

Step 2: Encoding

Step 3: PairFormer

LOTS Method: Modularized Pair-Centric Representation



Use **Sketch-Image Pairs** $\mathbf{C_i} = (\mathbf{S_i}, \mathbf{T_i})$ as input.

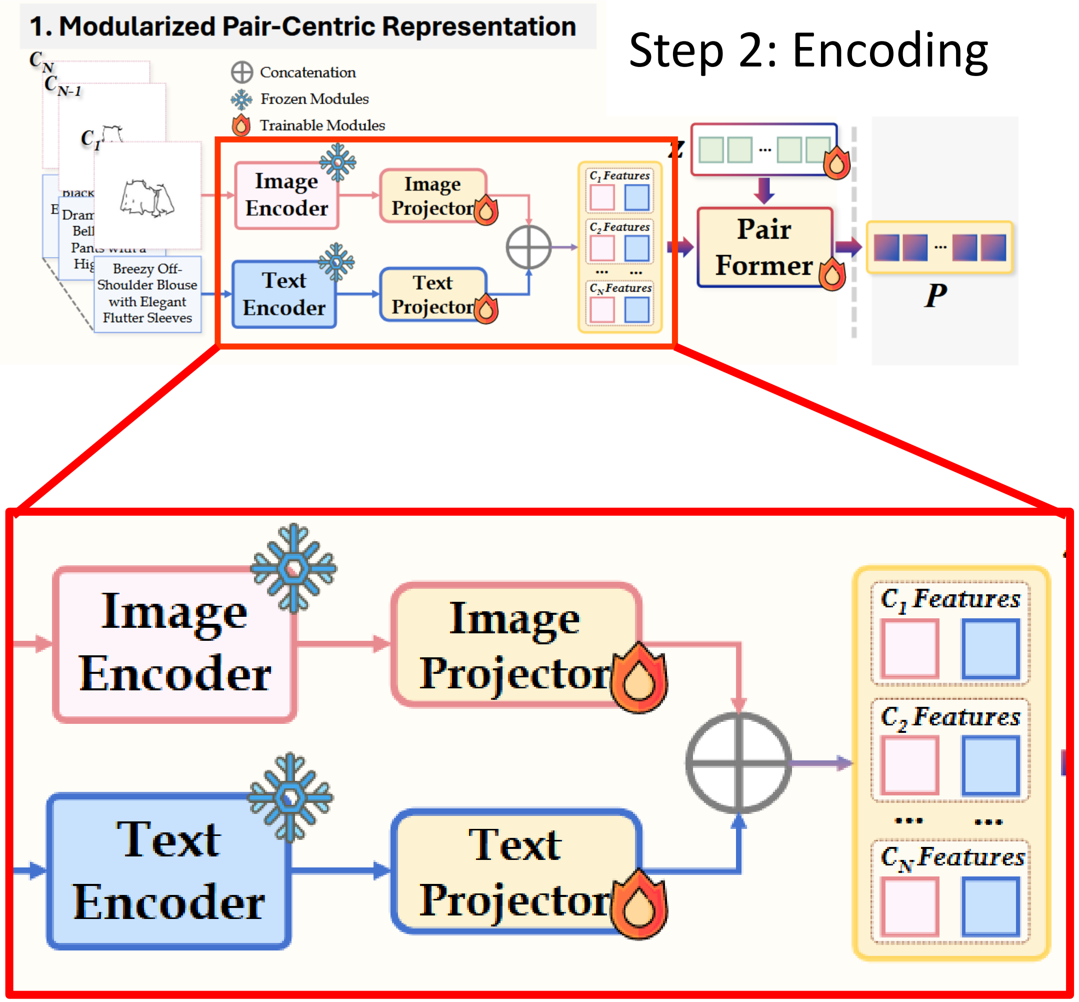
S_i is a binary sketch array, sharing same size with the target output image.

T_i is a text description in natural language such as "a shirt with flower pattern".

LOTS Method: Modularized Pair-Centric Representation

北京大学 PEKING UNIVERSITY

Step 2: Encoding



For each Sketch-Text Pair C_i :

Image Encoder f^S :

DINOv2 is employed to processes the sketch S_i into h_i^S.

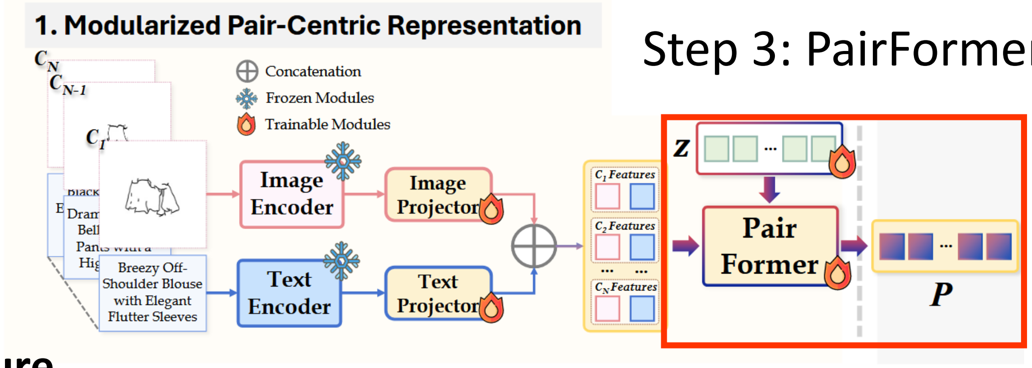
Text encoder f^T :

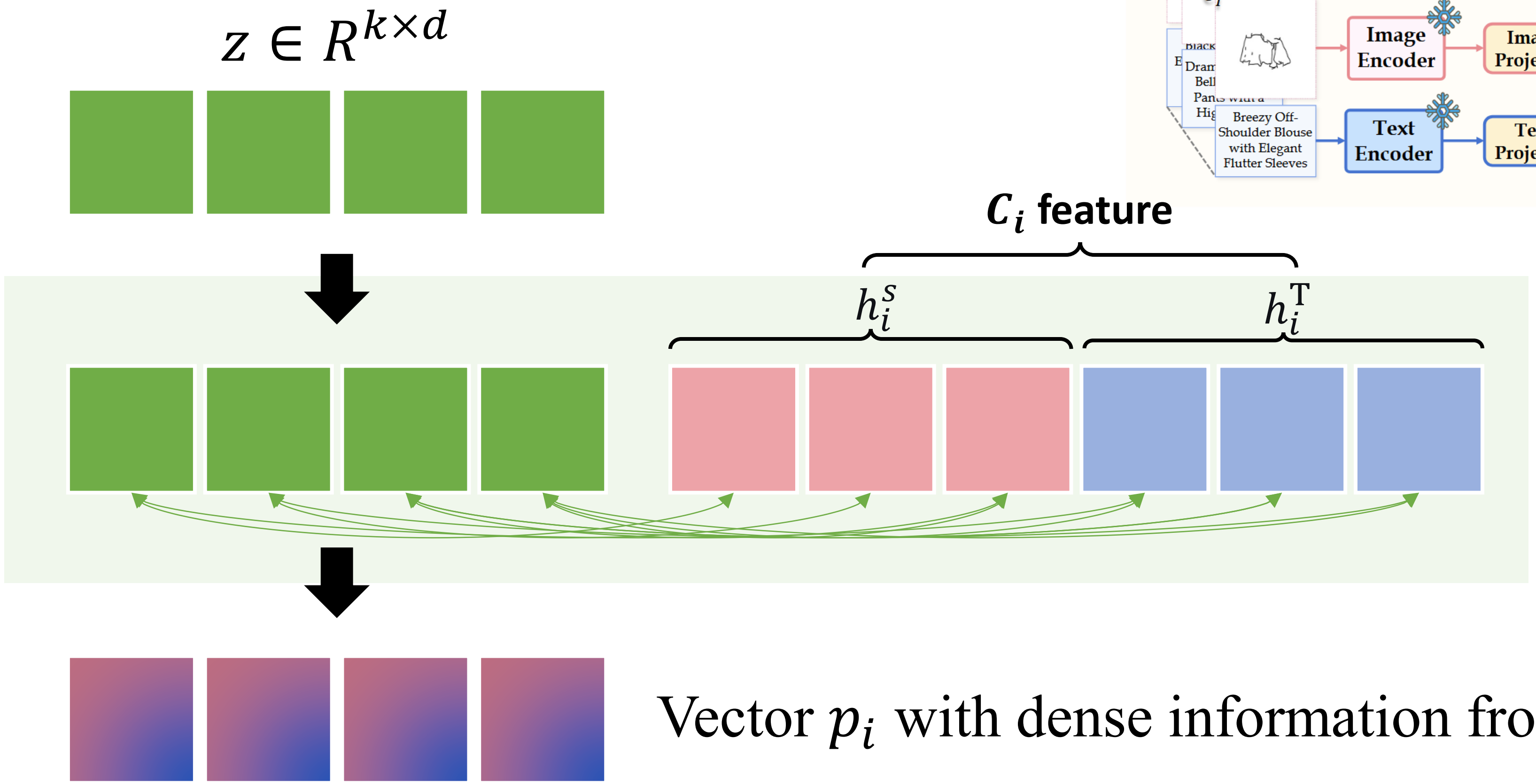
Use encoder of CLIP to process the text description T_i into h_i^T.

Encoders are followed by a **trainable FC layer** .

LOTS Method: Modularized Pair-Centric Representation

Step 3: PairFormer



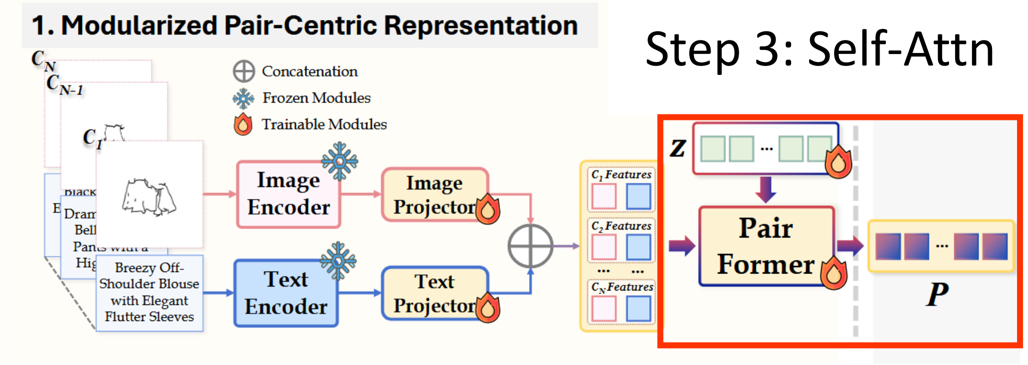


Self-attention

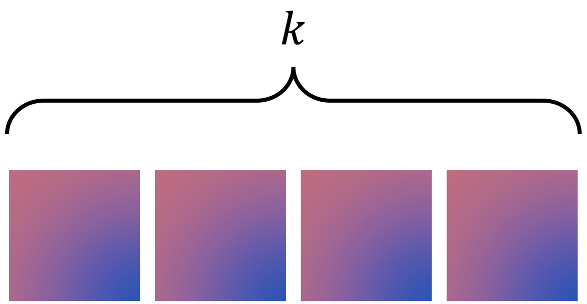
Vector p_i with dense information from $C_i = (S_i, T_i)$

Put sequence C_i into the self-attention layer of a Transformer. By doing so, the self-attention mechanism enables each token to focus on all other tokens in the sequence.

LOTS Method: Modularized Pair-Centric Representation



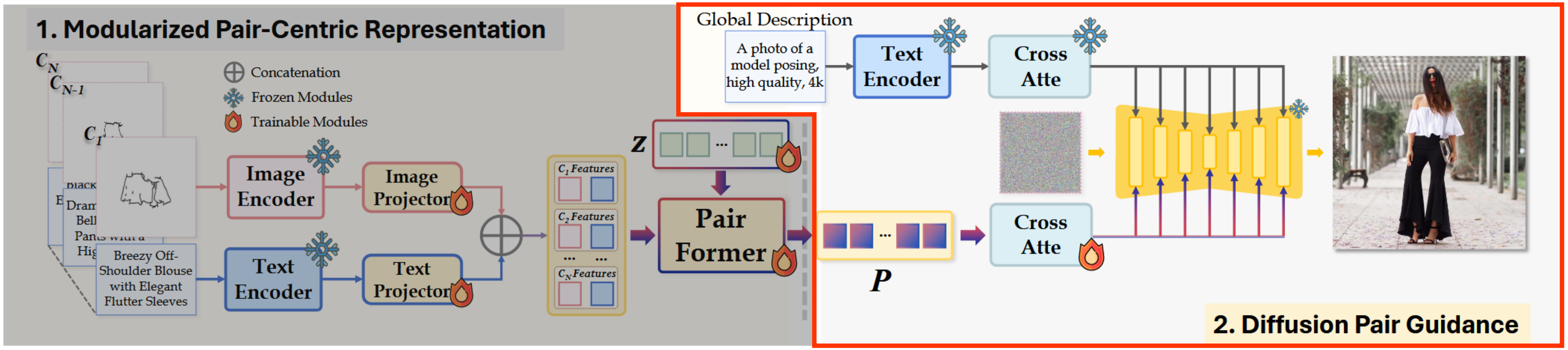
Step 3: Self-Attn



Vector p_i with dense information from $C_i = (S_i, T_i)$

The output of the self-attention layer is a sequence of the same length as the input sequence. We only take the **first k tokens** (that is, the part corresponding to the initial learnable token z) as the final fused representation of this pair p_i.

LOTS Method

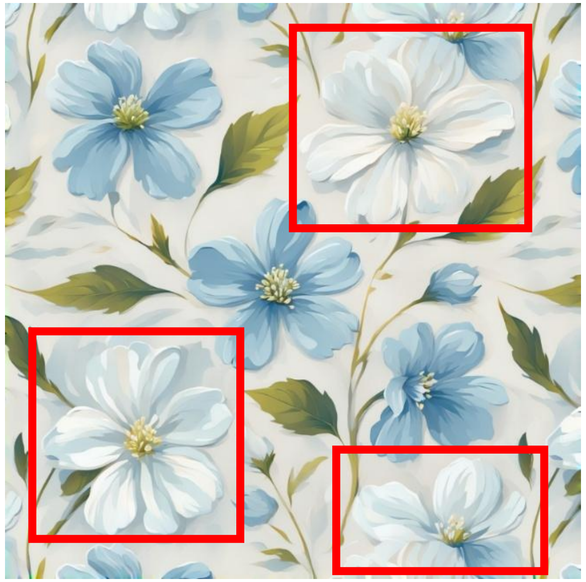


1. Modularized Pair-Centric Representation
Concatenation
Frozen Modules
Trainable Modules
Image Encoder
Image Projector
Text Encoder
Text Projector
Breezy Off-Shoulder Blouse with Elegant Flutter Sleeves
C_1 Features
C_2 Features
C_N Features
Z
Pair Former
P
Global Description
A photo of a model posing, high quality, 4k
Text Encoder
Cross Atte
Cross Atte
2. Diffusion Pair Guidance

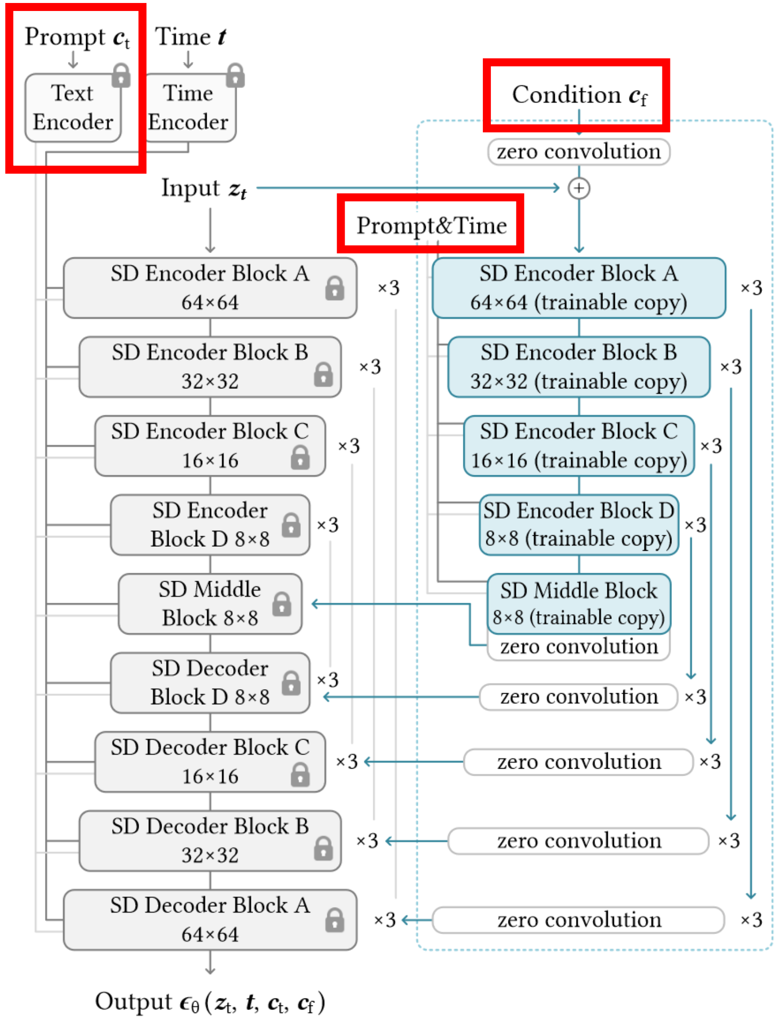
LOTS Method: Diffusion Pair Guidance

Fundamental Flaws in Traditional Methods:

- Attribute Confusion
- Premature condition merging



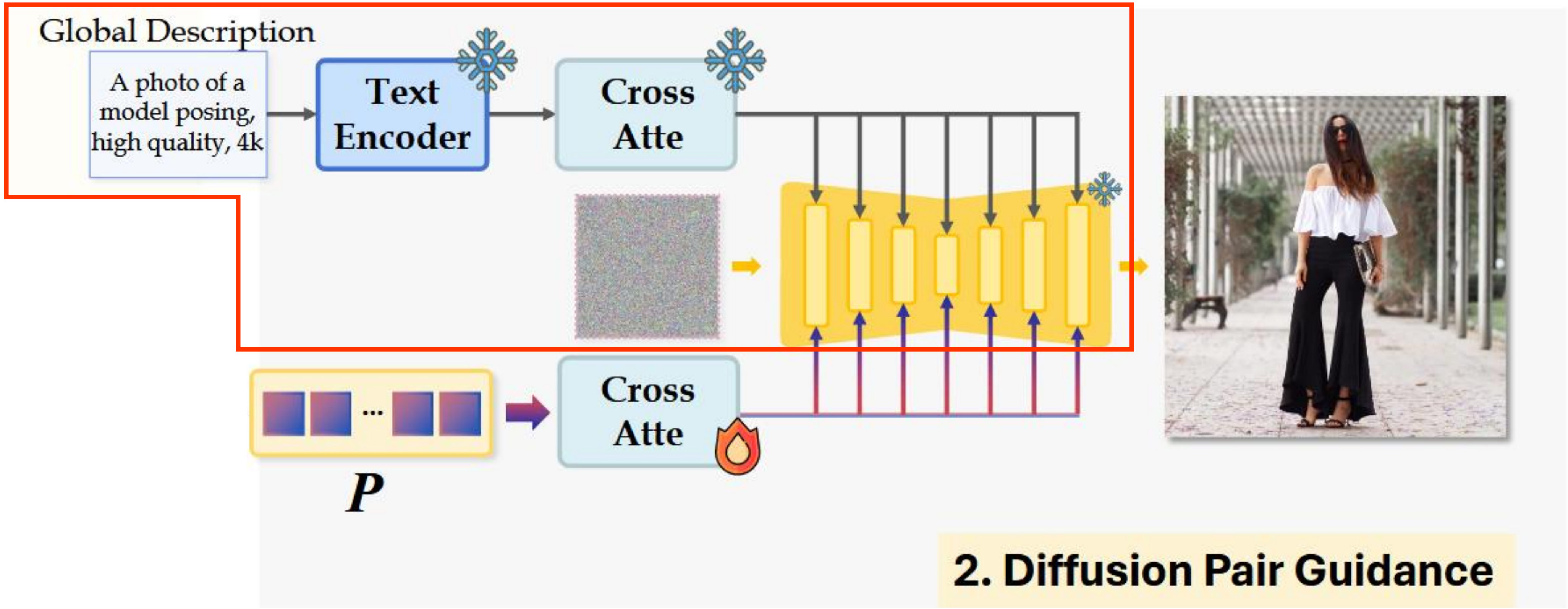
blue flower. green leaves. white background.



(a) Stable Diffusion (b) ControlNet

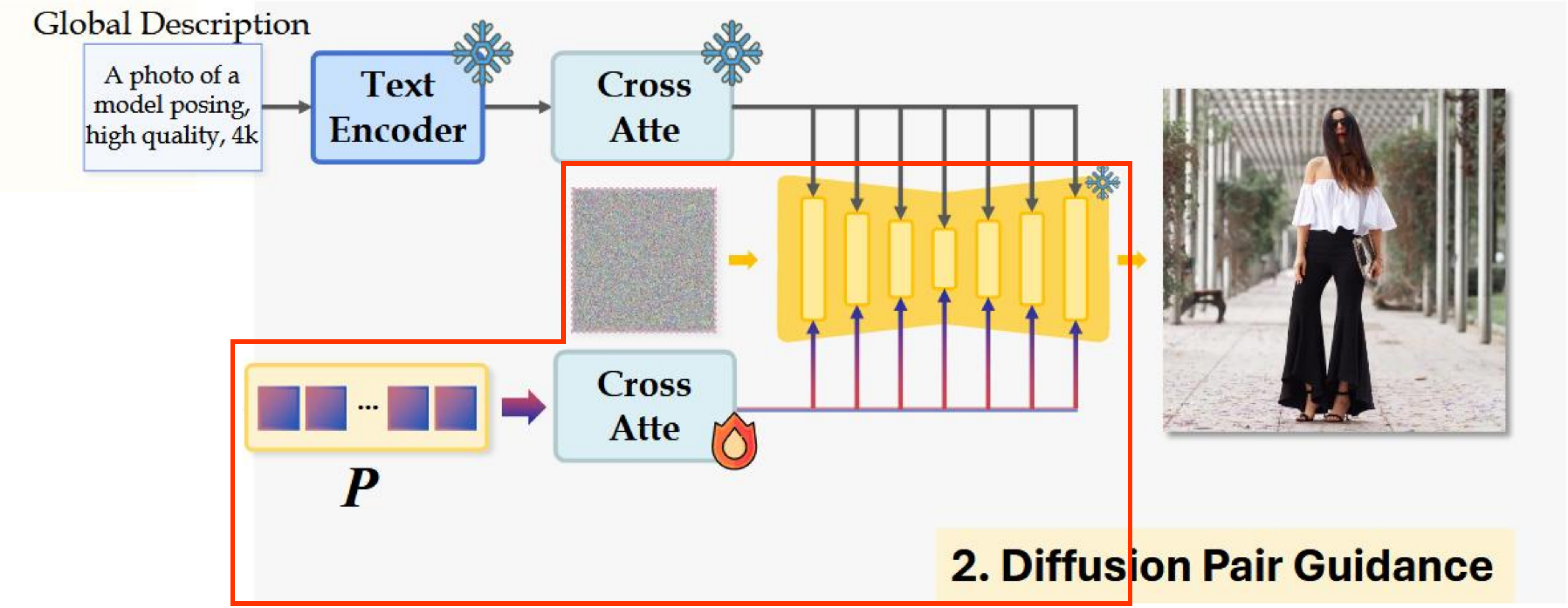
LOTS Method: Diffusion Pair Guidance

Traditional global text control on UNet via CrossAttn $w(x, h_g^T)$



LOTS Method: Diffusion Pair Guidance

U-Net Modification in LOTS:



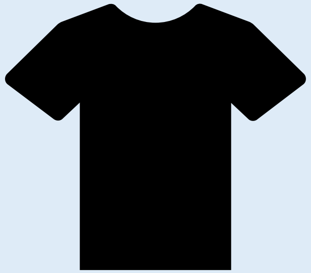
Local pair conditioning with a separate CrossAttn $\hat{w}(x, P)$

LOTS Method: Diffusion Pair Guidance

Total Feature P

p_{jacket}

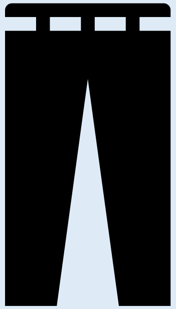
Striped jacket



K_{jacket}

p_{pants}

Plain pants

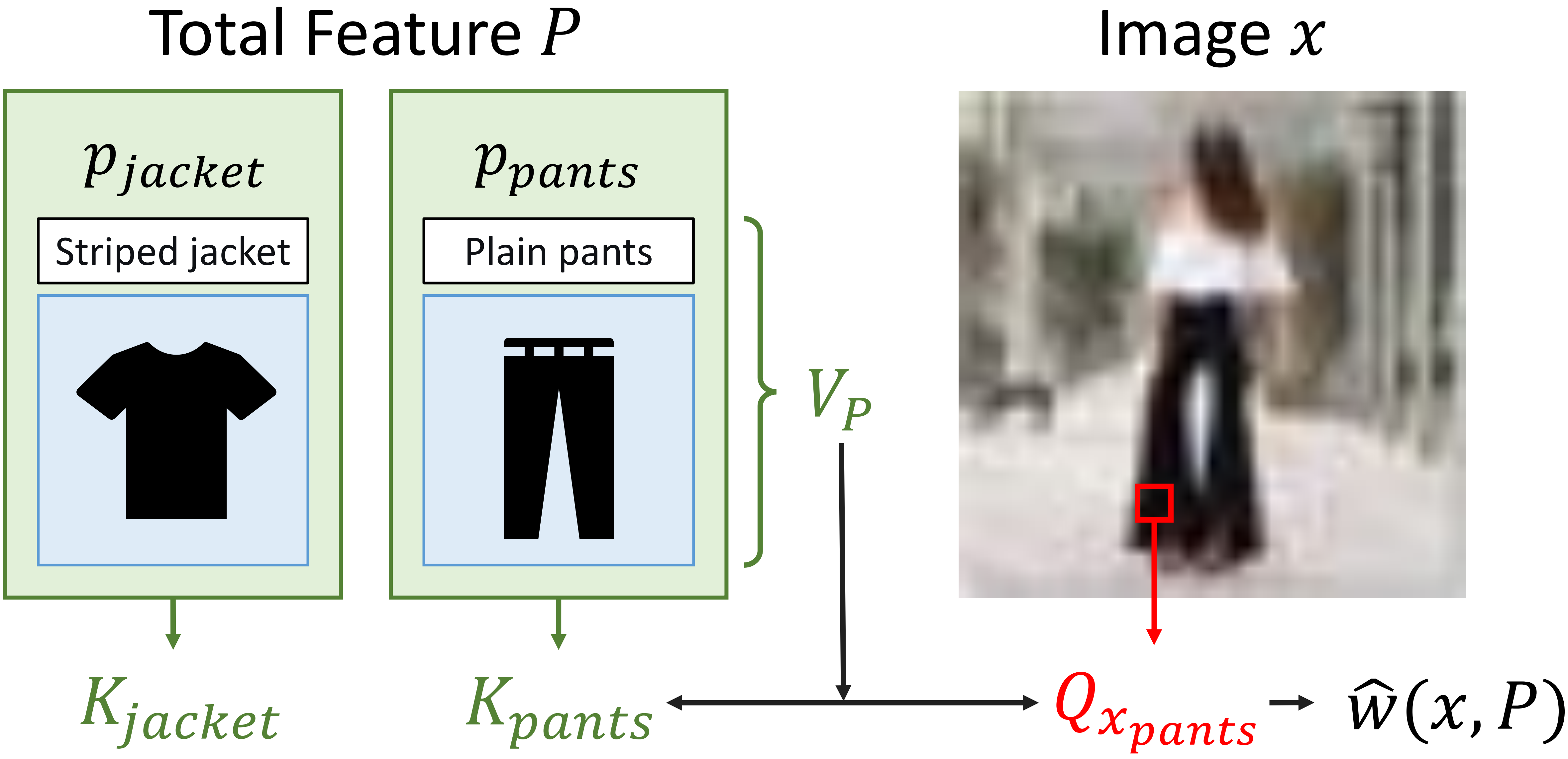


K_{pants}

Image x

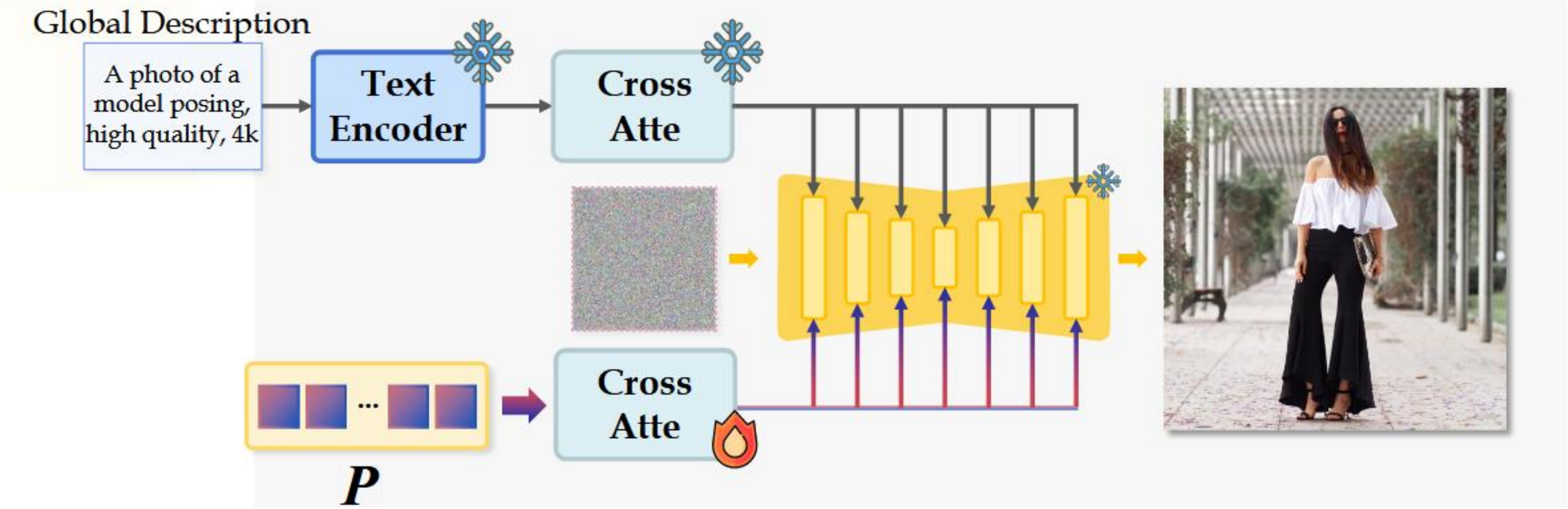
$Q_{x_{pants}}$

LOTS Method: Diffusion Pair Guidance



LOTS Method: Diffusion Pair Guidance

$w(x, P)$ Global information (text)



$\hat{w}(x, P)$ Localized details (text & sketch)

LOTS Method: Diffusion Pair Guidance

U-Net Modification in LOTS:

Base Formula:

$$x' = w(x, h_g^T) + \alpha \cdot \widehat{w}(x, P)$$

Detailed Expansion:

$$w(x, h_g^T) = Softmax(\frac{Q_x \cdot K_{h_g^T}^T}{\sqrt{d}}) \cdot V_{h_g^T}$$

$$\widehat{w}(x, P) = Softmax(\frac{Q_x \cdot K_P^T}{\sqrt{d}}) \cdot V_P$$

Where:

$Q_x = x \cdot W_Q$ (Image feature query)

$K_P = P \cdot W_K$ (Pair feature key)

$V_P = P \cdot W_V$ (Pair feature value)

Sketchy Dataset

- Based on Fasionpedia, with 47000+ images and 79000 annotions.
- 14 higher level categories(shirt, skirt, pants, etc.) & 21 lower level categories(sleeve, pocket, etc.)
- Generate clothing sketch from image via Photo-Sketching
- Apply LLaMMA-3.1-8B to generate descriptions on sketches with a average length of 16 words

LOTS of Fashion: Results

Conditions		LOTS	Multi-T2I-Adapter (zero-shot)	IP-Adapter	T2I-Adapter
	A hip-length, single-breasted blazer with plain design, notched lapels, wrist-length set-in sleeves, and two kangaroo pockets with a welt pocket. A classic floral, above-the-hip, regular fit, symmetrical top with no waistline, featuring a v-neck. A striped, maxi-length, straight pair of pants with a normal waist and regular fit, featuring a symmetrical design and a fly opening, and a simple buckle.				
	A floral, single-breasted blazer jacket with a regular fit, normal waist, above-the-hip length, notched lapel, welt pockets, and set-in sleeves with wrist-length cuffs. Low-waisted, striped, symmetrical, fly-front, straight pants with curved pockets. A plain, tight-fitting blouse with a normal waist, single-breasted front, symmetrical design, a flap pocket, a shirt collar and wrist-length sleeves				

LOTS of Fashion: Results

Conditions		LOTS	Conditions		LOTS
	Double-breasted, floral jacket with peak lapels, and wrist-length set-in sleeves. Maxi length, symmetrical and straight sailor pants with a check pattern.			A regular-fit shirt with a striped pattern [...], featuring set-in sleeves and a regular collar. A plain, classic printed t-shirt with a round neckline. Loose, maxi-length, straight, check pants.	
	A tight-fitting, floral shirt with a hip length, short sleeves and a traditional shirt collar. Striped bermuda shorts.			A dotted, regular-fit, hip-length shirt, featuring short set-in sleeves and a traditional shirt collar. Check, loose-fitting, above-the-knee bermuda shorts.	

LOTS of Fashion: Results

Model	Conditioning Visual/Textual	Global Quality		Compositional Alignment		
		FID (↓)	GlobalCLIP (↑)	LocalCLIP (↑)	VQAScore (↑)	SSIM (↑)
SD [34]	-/G	1.11	.603	.745	.719	.663
SDXL [30]	-/G	1.77	.529	.701	.660	.544
GLIGEN [20]	-/L	0.93	.568	.704	.395	.614
ControlNet [46]	G/G	1.08	.622	.789	.733	.674
Multi-ControlNet [46]	L/G	1.10	.615	.780	.730	.672
IP-Adapter [45]	G/G	2.80	.537	.682	.611	<u>.715</u>
T2I-Adapter [25]	G/G	2.16	.534	.705	.635	.482
Multi-T2I-Adapter [25]	L/G	1.14	.583	.766	.697	**.723**
AnyControl [37]	L/G	0.99	.602	.777	.712	.544
GLIGEN [20]	-/L	1.70	.564	.713	.419	.514
ControlNet [46]	G/G	0.80	<u>.645</u>	<u>.801</u>	.717	.574
Multi-ControlNet [46]	L/G	0.84	.638	.799	.720	.572
IP-Adapter [45]	G/G	**0.69**	.621	.787	.714	.631
T2I-Adapter [25]	G/G	1.03	.570	.753	**.749**	.612
Multi-T2I-Adapter [25]	L/G	1.11	.559	.744	<u>.734</u>	.605
LOTS (Ours)	$\boxed{\text{L/L}}$	<u>0.79</u>	**.679**	**.813**	**.749**	.678

(a) Comparisons between LOTS and state-of-the-art sketch-to-image approaches. In the Conditioning column, L and G indicate whether the model accepts Local or Global inputs as Visual or Textual conditioning. We divide the table into three sections: zero-shot approaches , fine-tuned approaches on Sketchy, and our approach LOTS . We highlight the best performance in bold and underline the second best

LOTS of Fashion: Results

北京大学 PEKING UNIVERSITY

Model	Attribute Localization		
	Precision (↑)	Recall (↑)	F1 (↑)
SDXL [30]	.636	**.754**	<u>.690</u>
ControlNet [46]	.596	.449	.512
Multi-ControlNet [46]	.487	.365	.418
IP-Adapter [45]	.625	.139	.227
T2I-Adapter [25]	.409	.170	.240
Multi-T2I-Adapter [25]	.370	.270	.312
AnyControl [37]	.281	.134	.182
ControlNet [46]	<u>.667</u>	.516	.582
Multi-ControlNet [46]	.541	.417	.471
IP-Adapter [45]	.559	.384	.455
T2I-Adapter [25]	.463	.397	.427
Multi-T2I-Adapter [25]	.551	.692	.614
LOTS (Ours)	**.813**	<u>.650</u>	**.722**

(b) Results of qualitative user study of attribute localization and confusion conducted between LOTS and other models. We highlight the best results for each metric in bold and underline the second best.

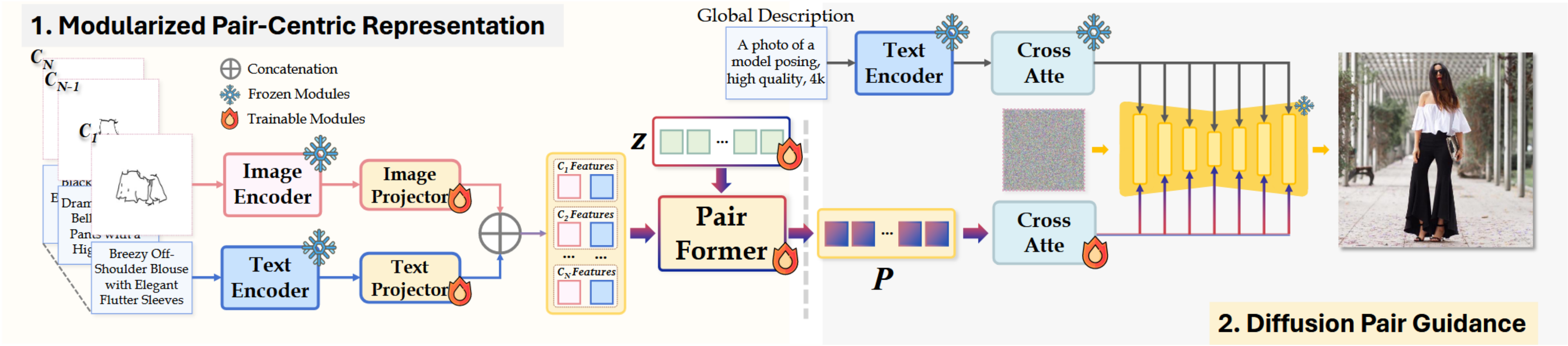
LOTS of Fashion: Conclusion

- **Novel Method:** Proposes LOTS, a new approach for fine-grained image generation using localized sketch-text pairs.

- **Technical Innovation:** Introduces a "delayed fusion" mechanism that processes conditions during diffusion, solving attribute confusion.

- **New Dataset:** Creates Sketchy, the first fashion dataset with localized sketch-text annotations.

- **SOTA Results:** Demonstrates superior performance in both image quality and attribute localization.



Reference

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/ diffusers, 2022.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In ACM SIGGRAPH, 2023.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. arXiv preprint, 2022.

Xi Wang, Hongzhen Li, Heng Fang, Yichen Peng, Haoran Xie, Xi Yang, and Chuntao Li. Lineart: A knowledgeguided training-free high-quality appearance transfer for design drawing with diffusion model. In CVPR, 2025. 3, 6

Zhifeng Xie, Hao Li, Huiming Ding, Mengtian Li, Xinhan Di, and Ying Cao. Hierafashdiff: Hierarchical fashion design with multi-stage diffusion models. In AAAI, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-toimage diffusion models. arXiv preprint, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In NeurIPS, 2024.

Thanks for listening!

Presenter: XuShenghan
2025.10.19