北京大學 PEKING UNIVERSITY

Generative Photomontage

Sean J. Liu, Nupur Kumari, Ariel Shamir, Jun-Yan Zhu

CVPR 2025

Presenter: XuShenghan
2025.03.23

Introduction: Photomontage

The Constructor, El Lissitsky, 1924

Album cover for The Beatles, Sgt. Pepper's Lonely Hearts Club Band, 1967

Mask XXXV, John Stezaker, 2007

Introduction: Photomontage

The Constructor, El Lissitsky, 1924

Album cover for The Beatles, Sgt. Pepper's Lonely Hearts Club Band, 1967

Mask XXXV, John Stezaker, 2007

Generative Photomontage: Task Description

Generation of ControlNet:

akin to a ***dice roll***, hard to achieve a single image that captures everything a user wants.

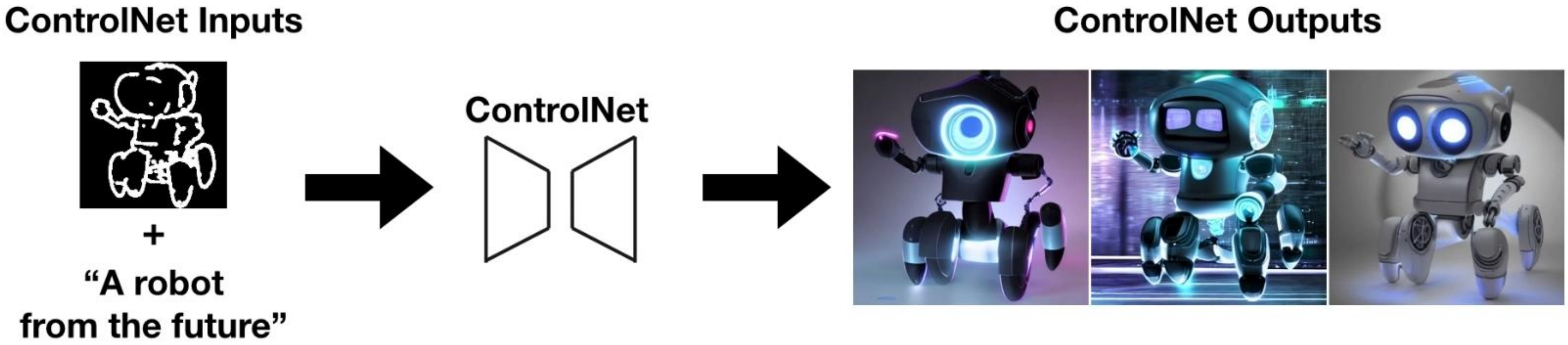
Input Canny edge | Default | "masterpiece of fairy tale, giant deer, golden antlers" | "..., quaint city Galic"

What if the user wants to keep the dear, moon and background from each result respectively?

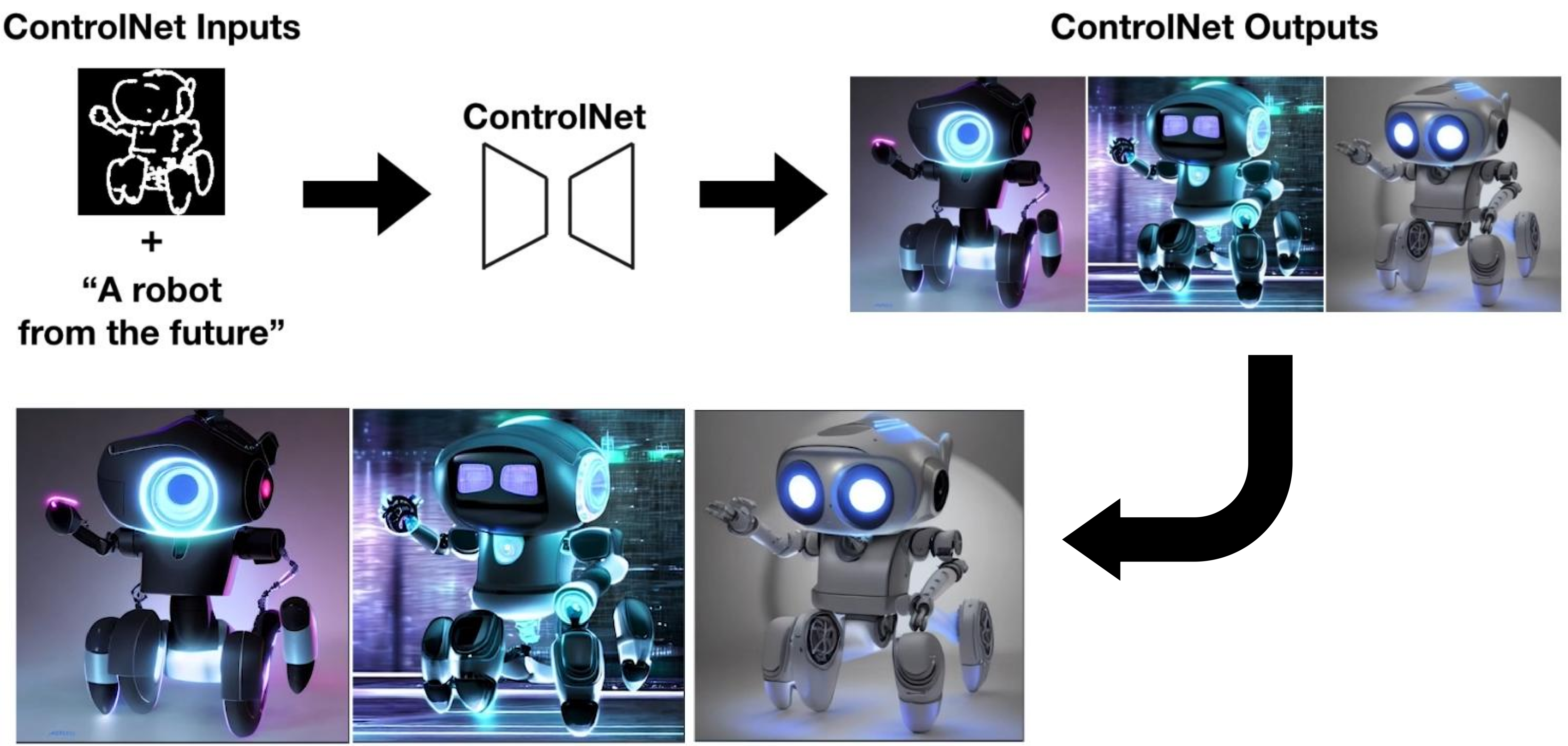
Generative Photomontage: Task Description

Existing methods that add various conditions to text-to-image models for greater user control fail to adhere closely to the input conditions.

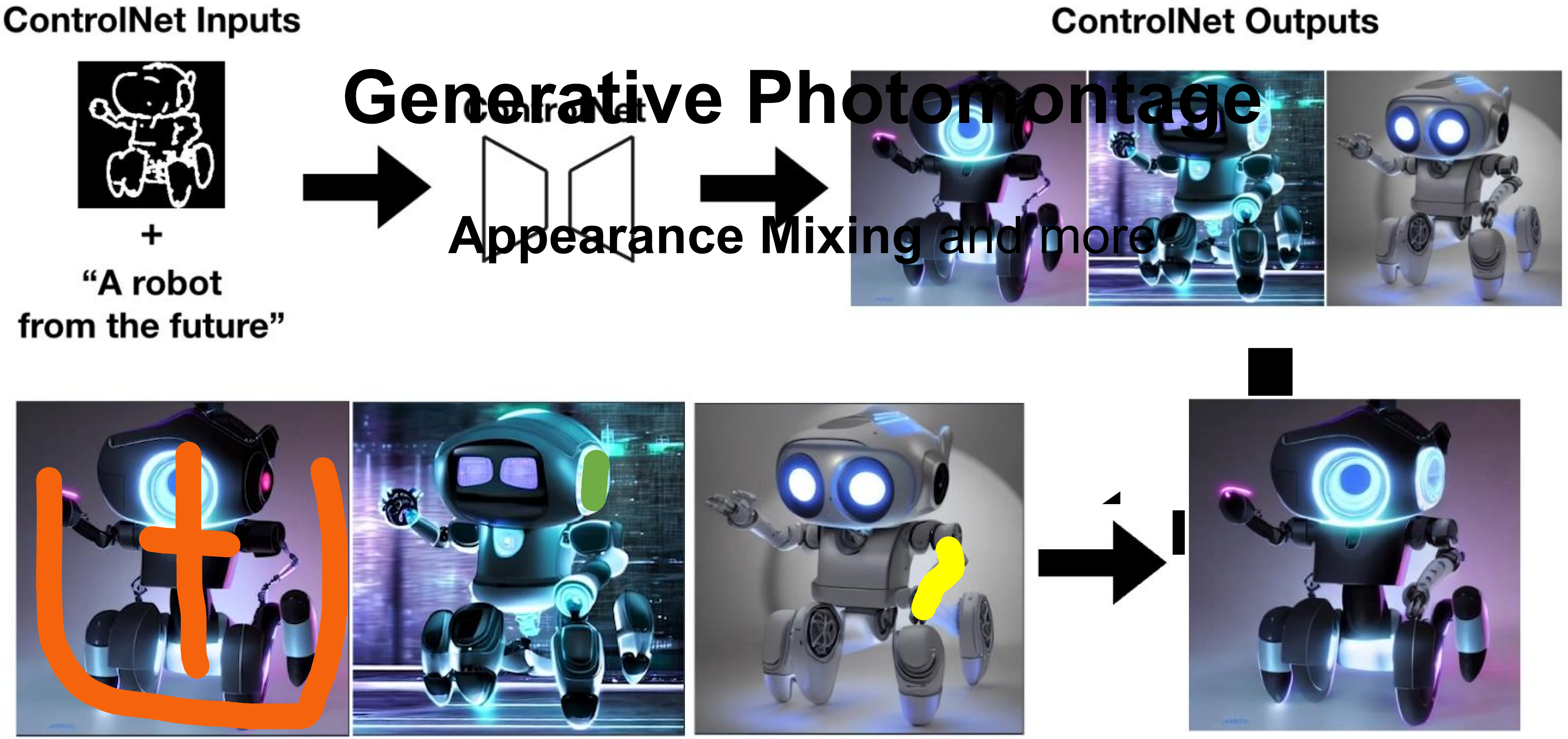
Faithfully preserve & compositing harmoniously, using a stack of ControlNet output image.



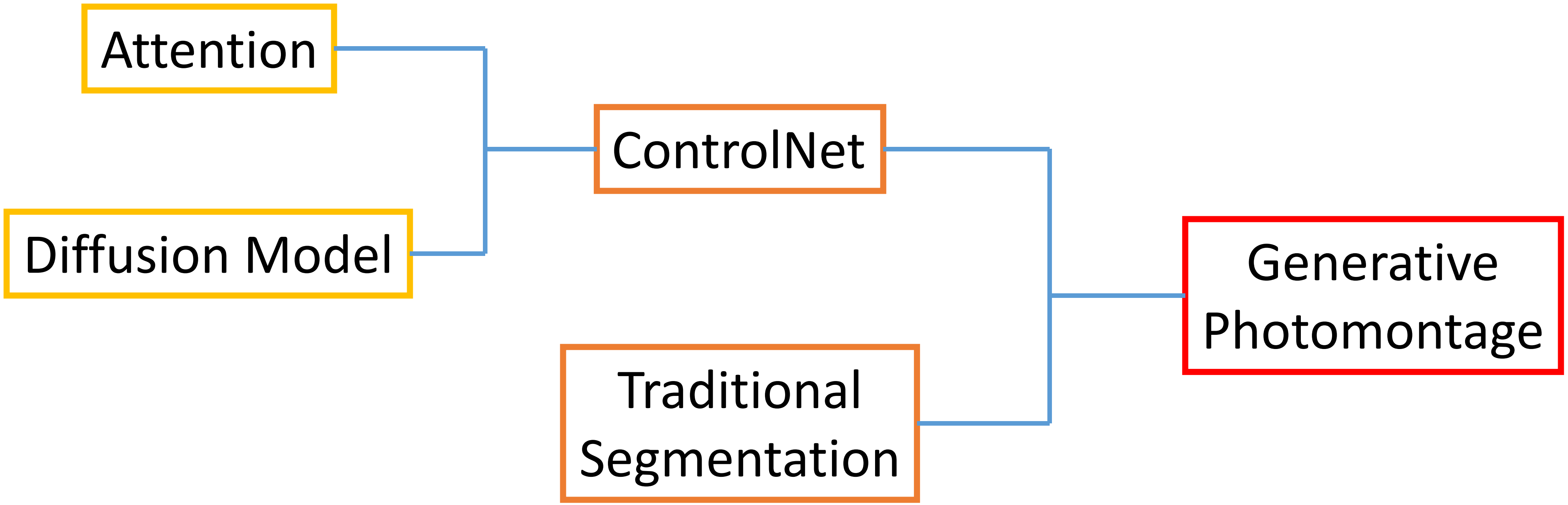
Generative Photomontage: Task Description



Generative Photomontage: Task Description



Background


Attention
Diffusion Model
ControlNet
Traditional Segmentation
Generative Photomontage

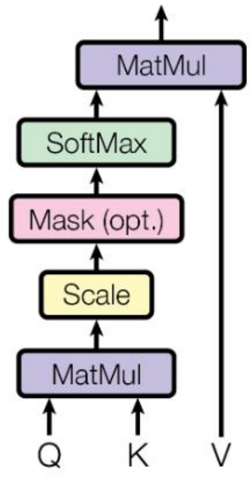
Background: Attention Mechanism

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

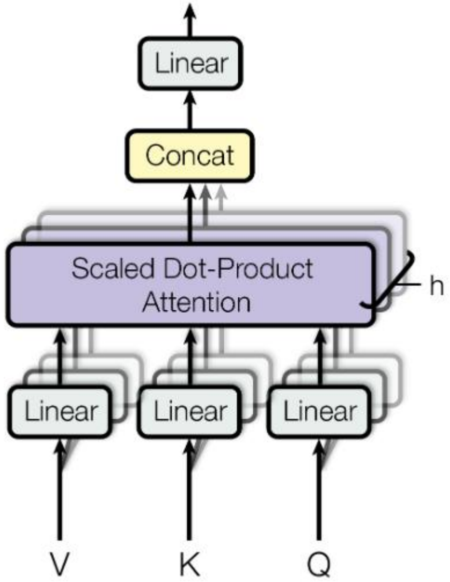
- *Long-range dependency and dynamic weight*
- *Global information capturing*

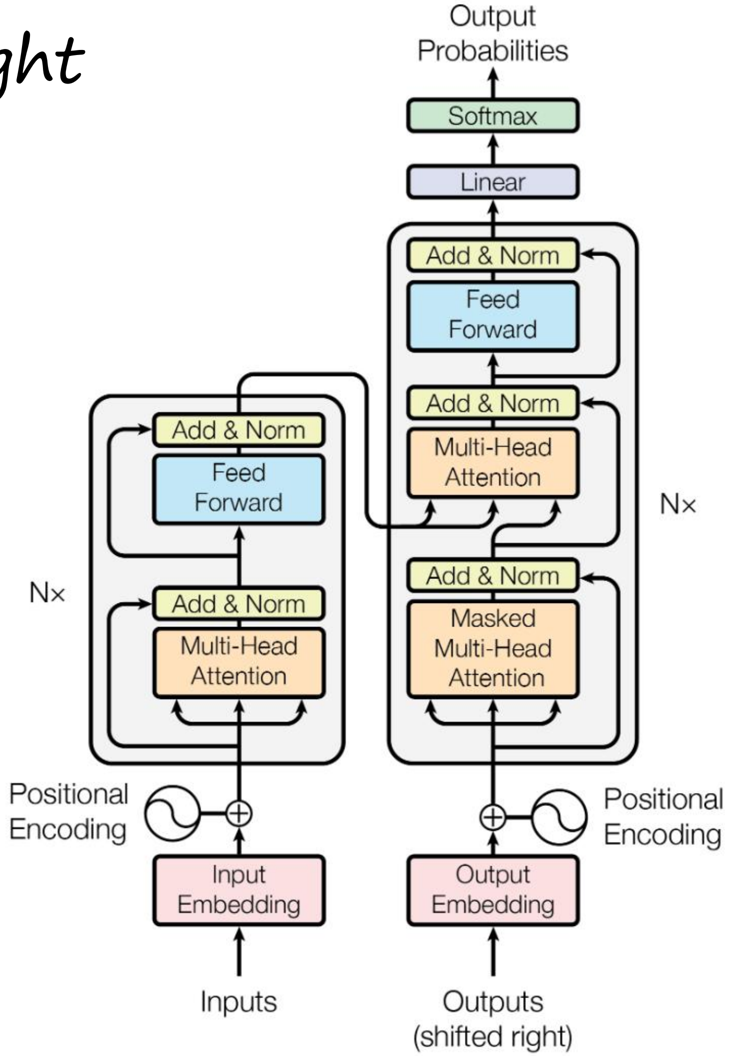
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})\,V$$

Scaled Dot-Product Attention



Multi-Head Attention



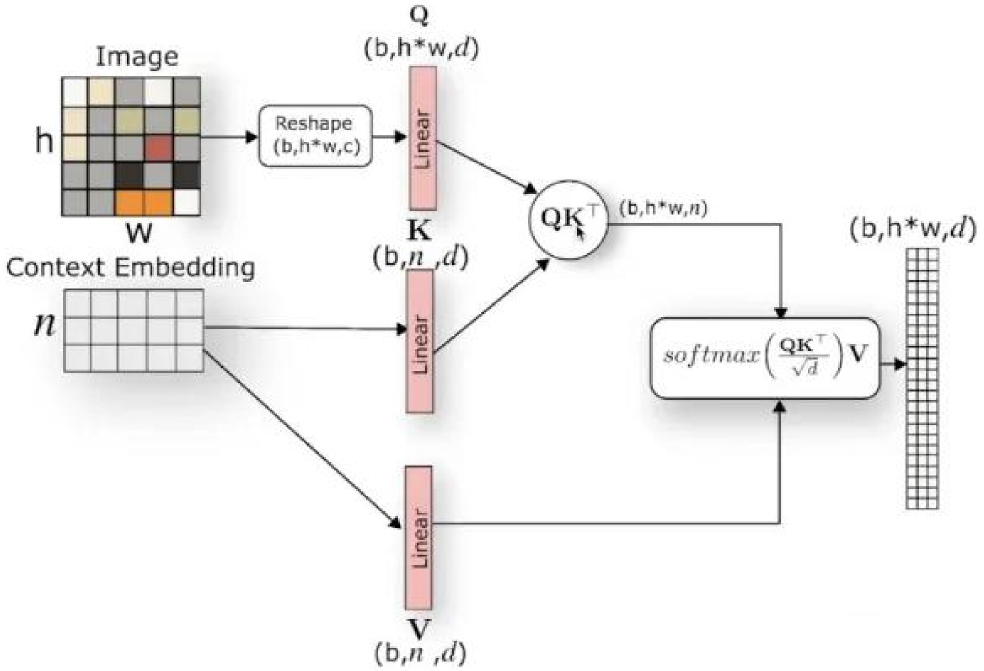


Background: Cross Attention

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

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

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

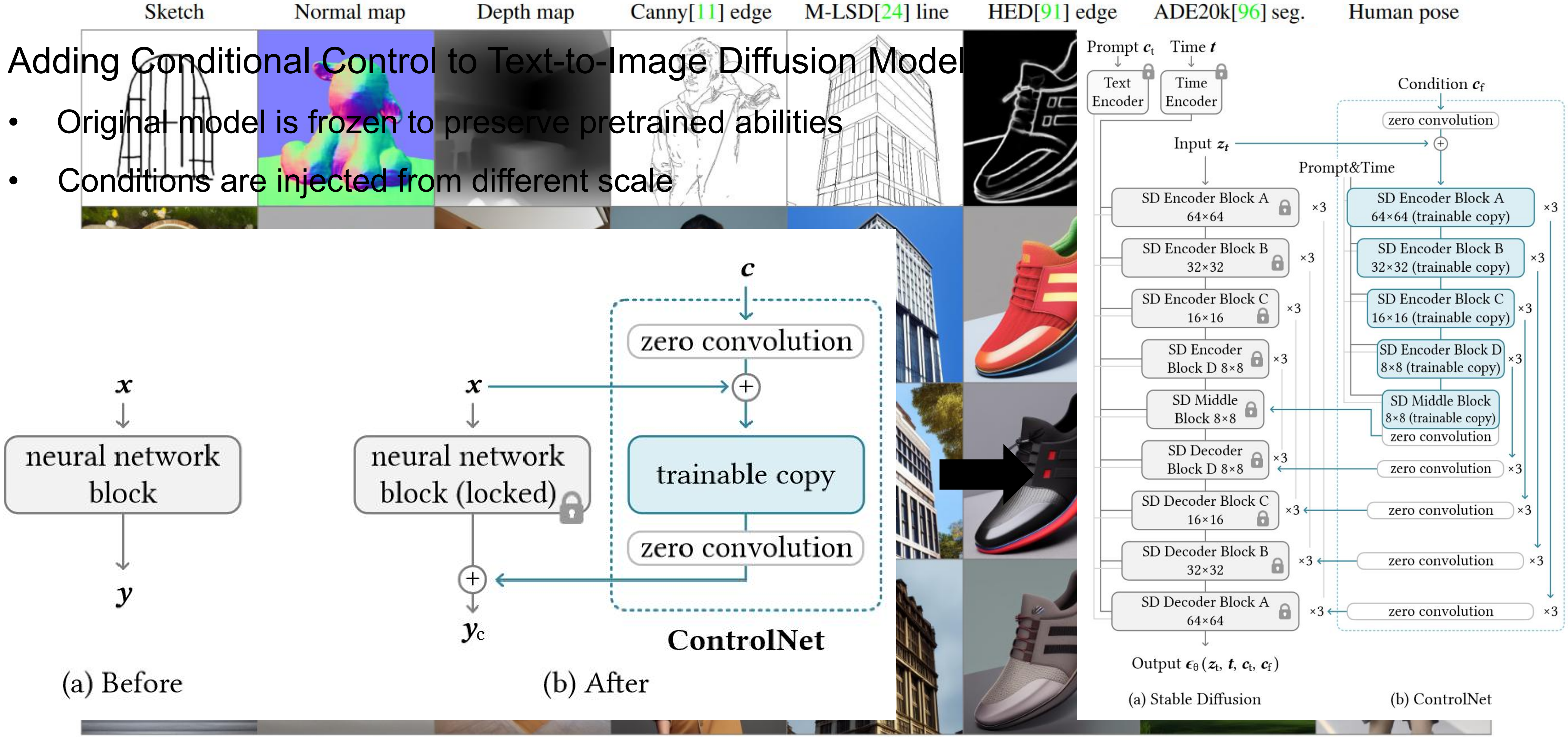
V : Injects **detailed appearance** information into the output

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

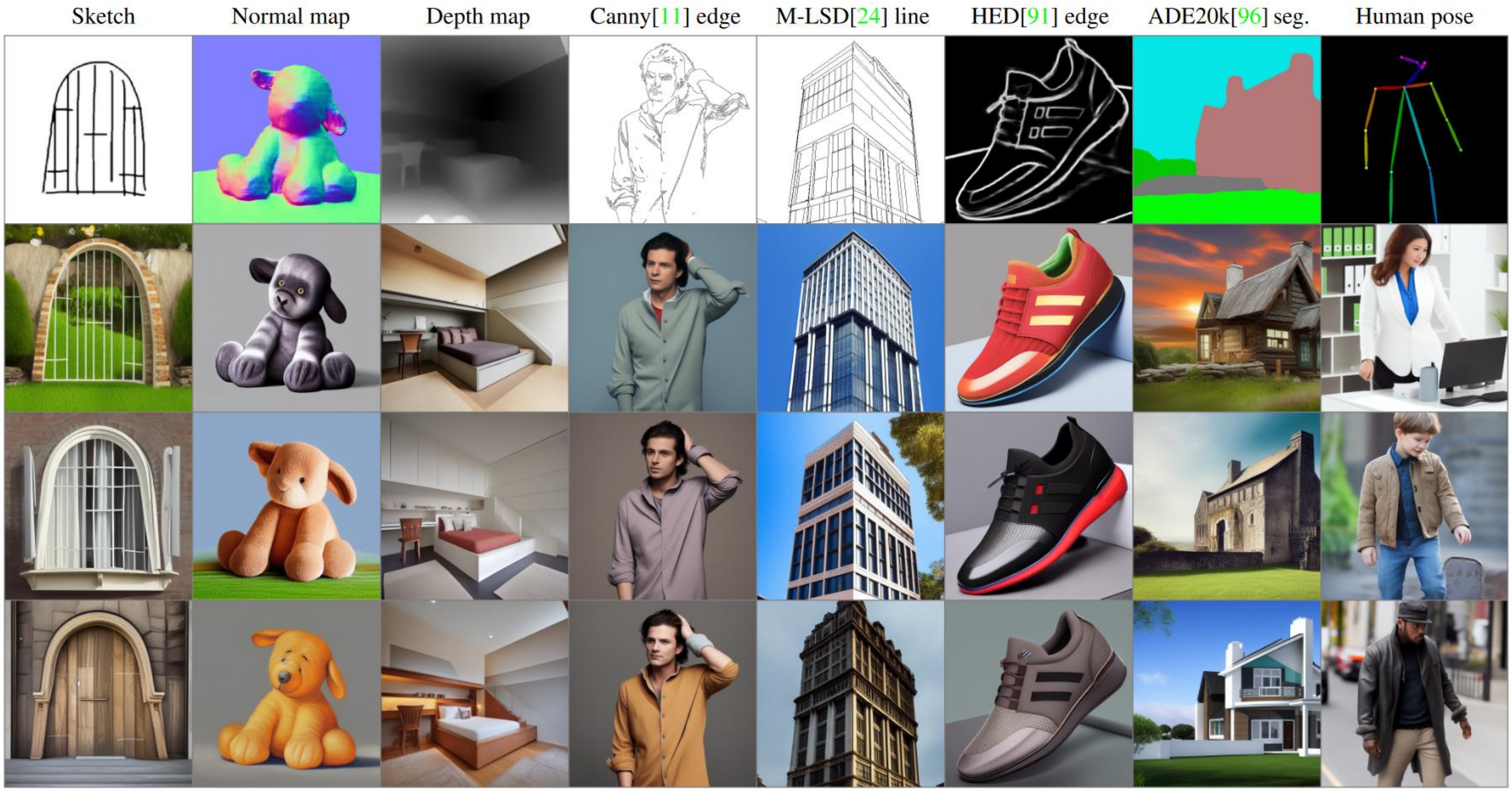
- Original model is frozen to preserve pretrained abilities
- Conditions are injected from different scale



Background: ControlNet

Addi

- C
- C



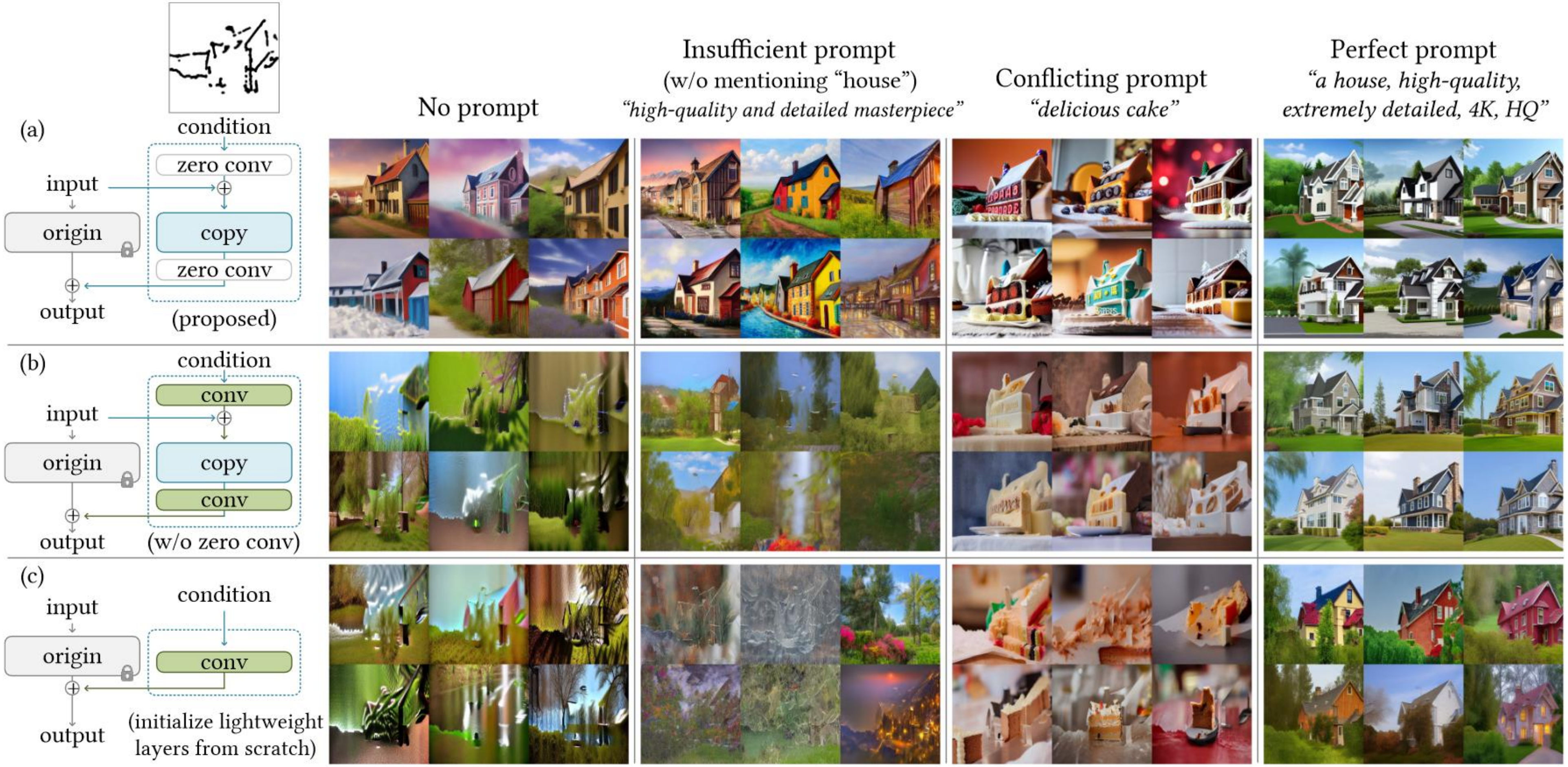
neu

(a

Background: ControlNet

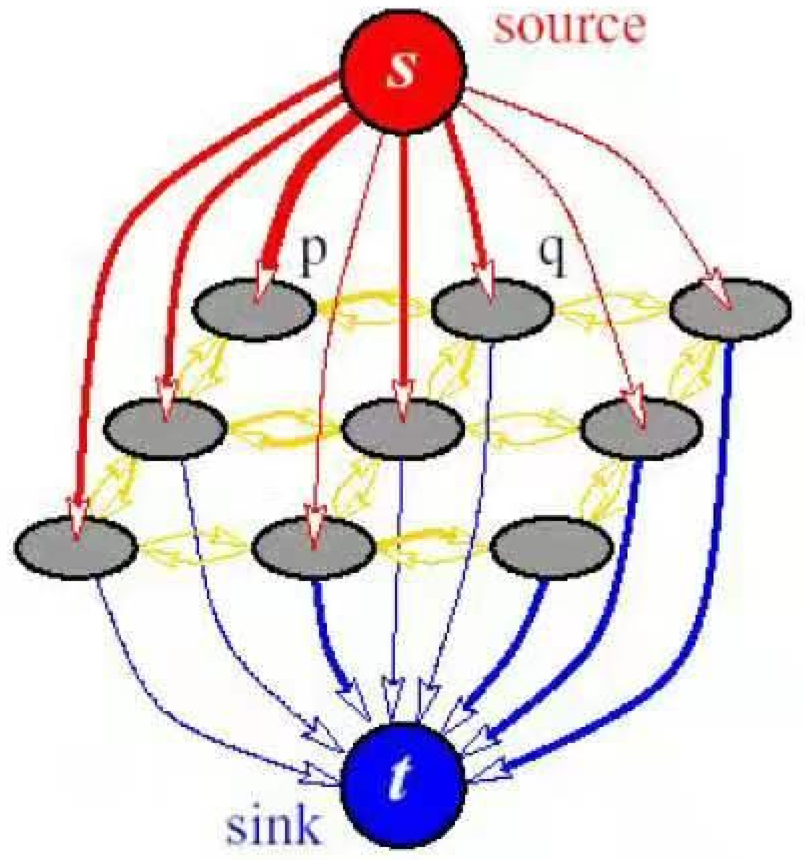
Due to the **zero convolutions**, ControlNet always predicts high-quality images during the entire training.

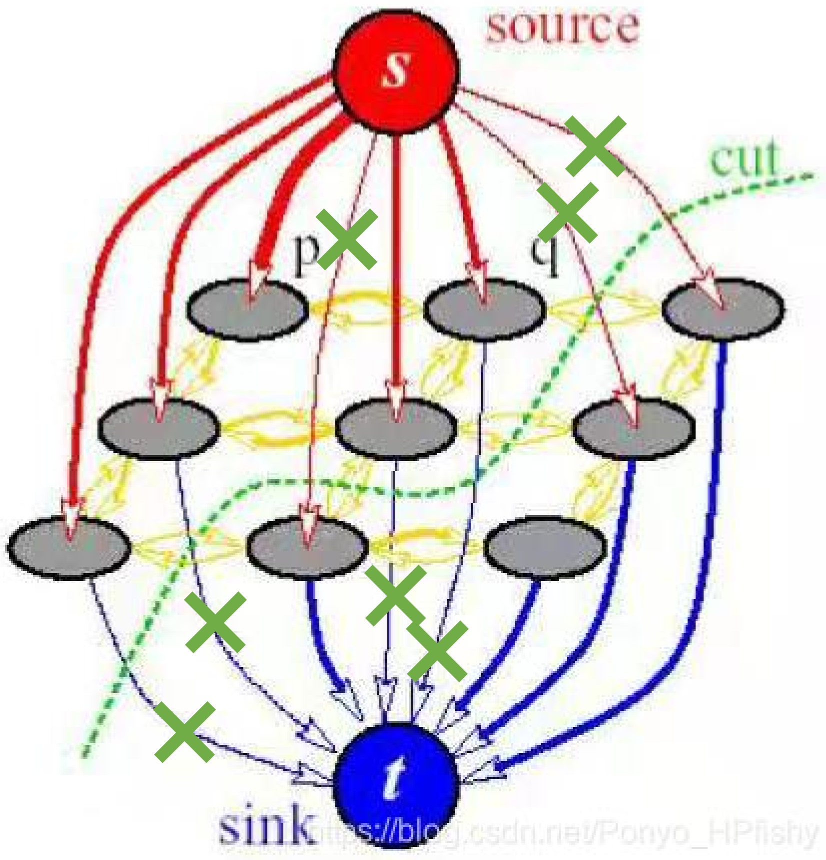
At a certain step in the training process, the model suddenly learns to follow the input condition.



Background: Traditional Segmentation

Max-Flow/Min-Cut

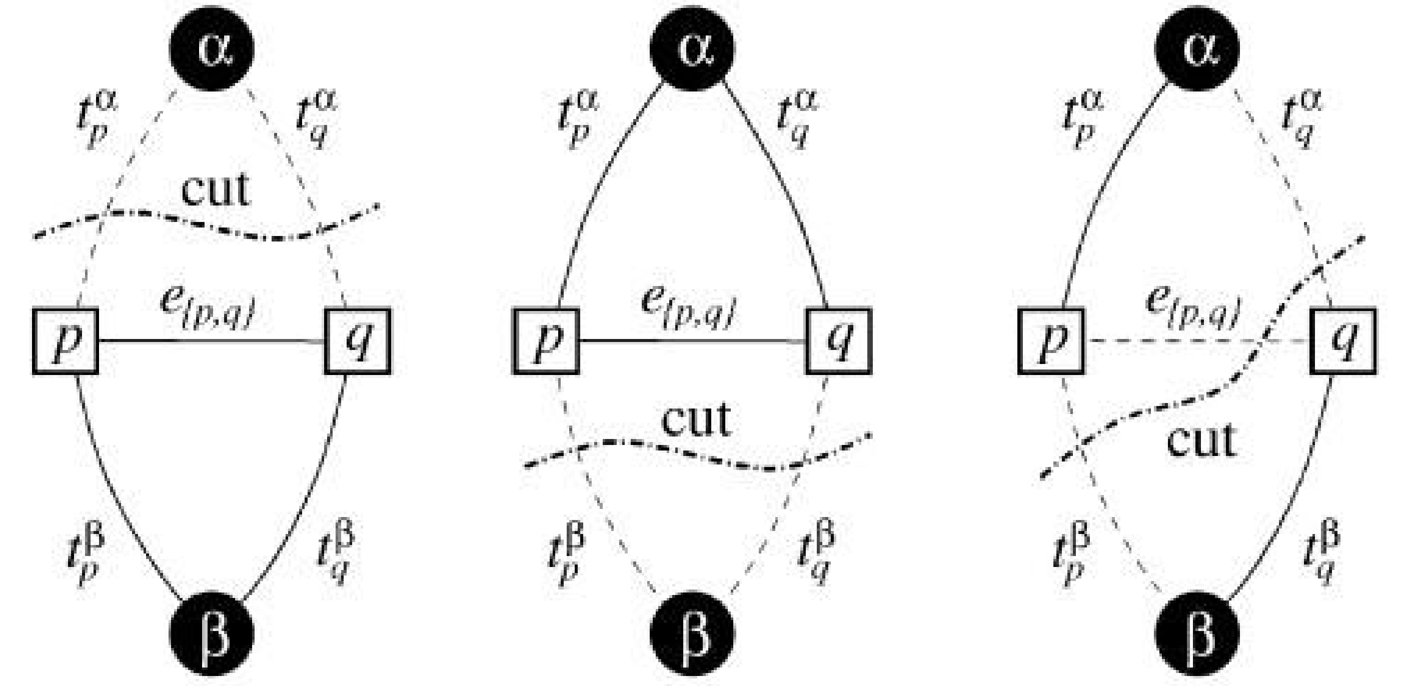




https://blog.csdn.net/Ponyo_HPfishy/article/details/103127138

Background: Traditional Segmentation

Segmentation with 2 labels α and β

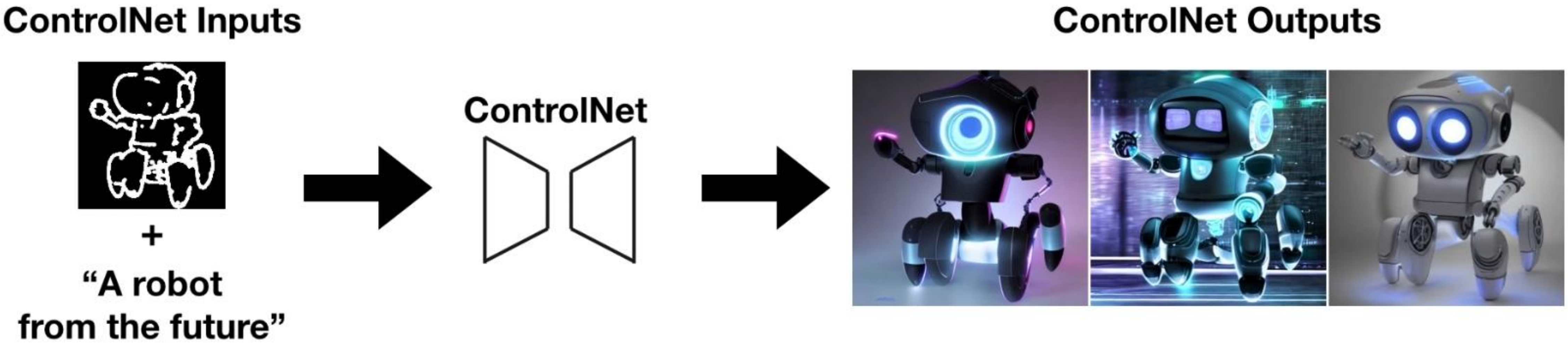


Property 4.2(a) Property 4.2(b) Property 4.2(c,d)

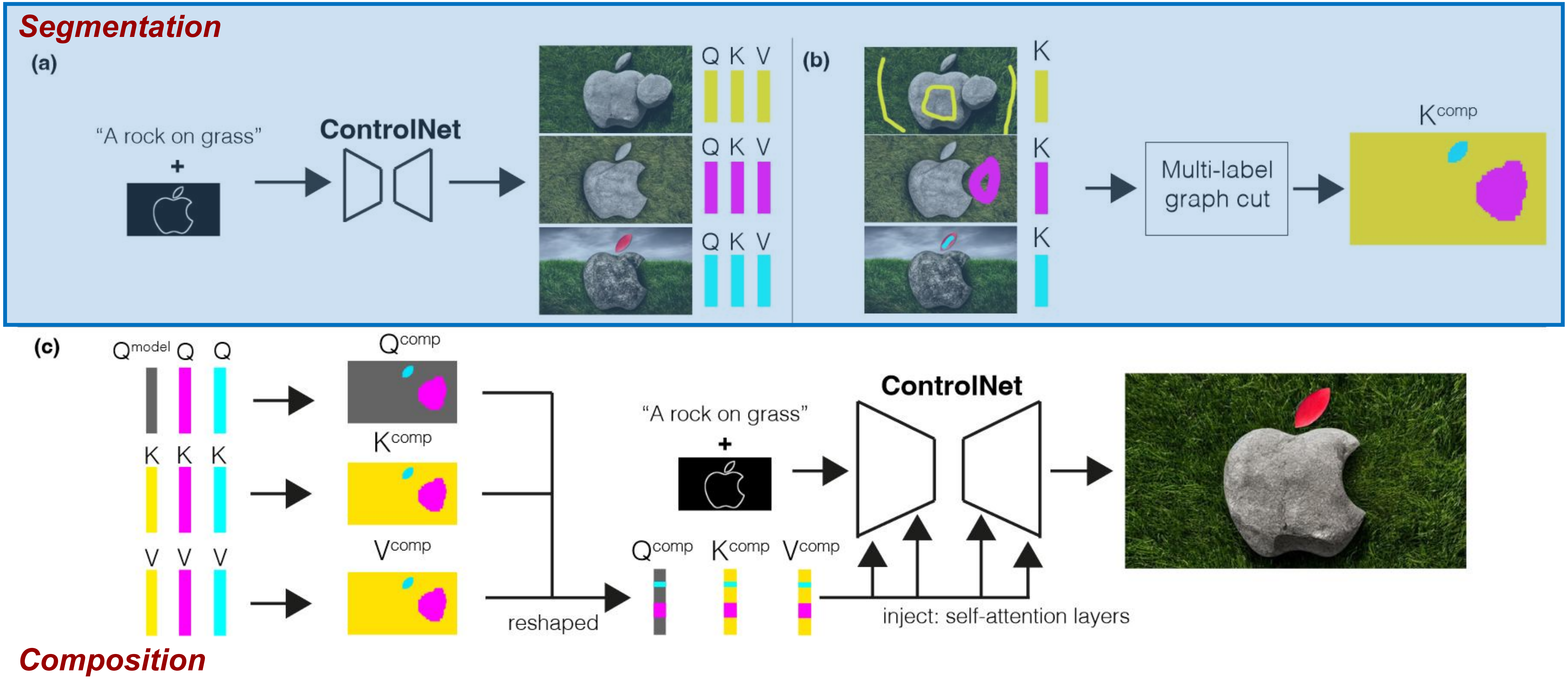
Generative Photomontage: Task Description

Existing methods that add various conditions to text-to-image models for greater user control fail to adhere closely to the input conditions.

Faithfully preserve & compositing harmoniously, using a stack of ControlNet output image.



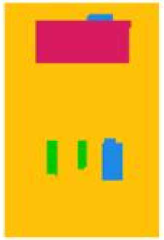
Generative Photomontage: Task Description

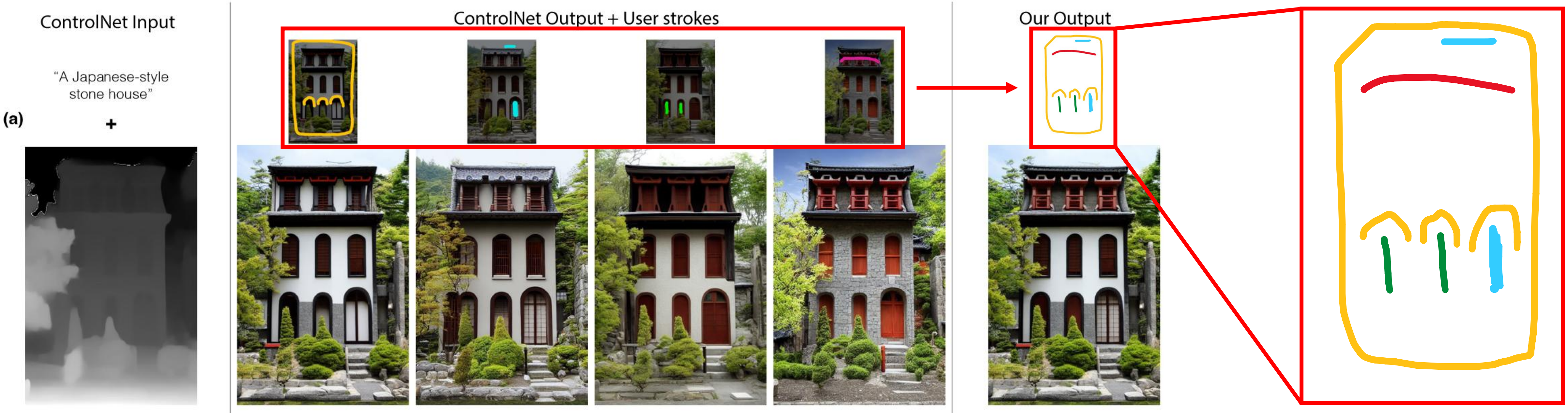


Step1: Segmentation with Graph Cut

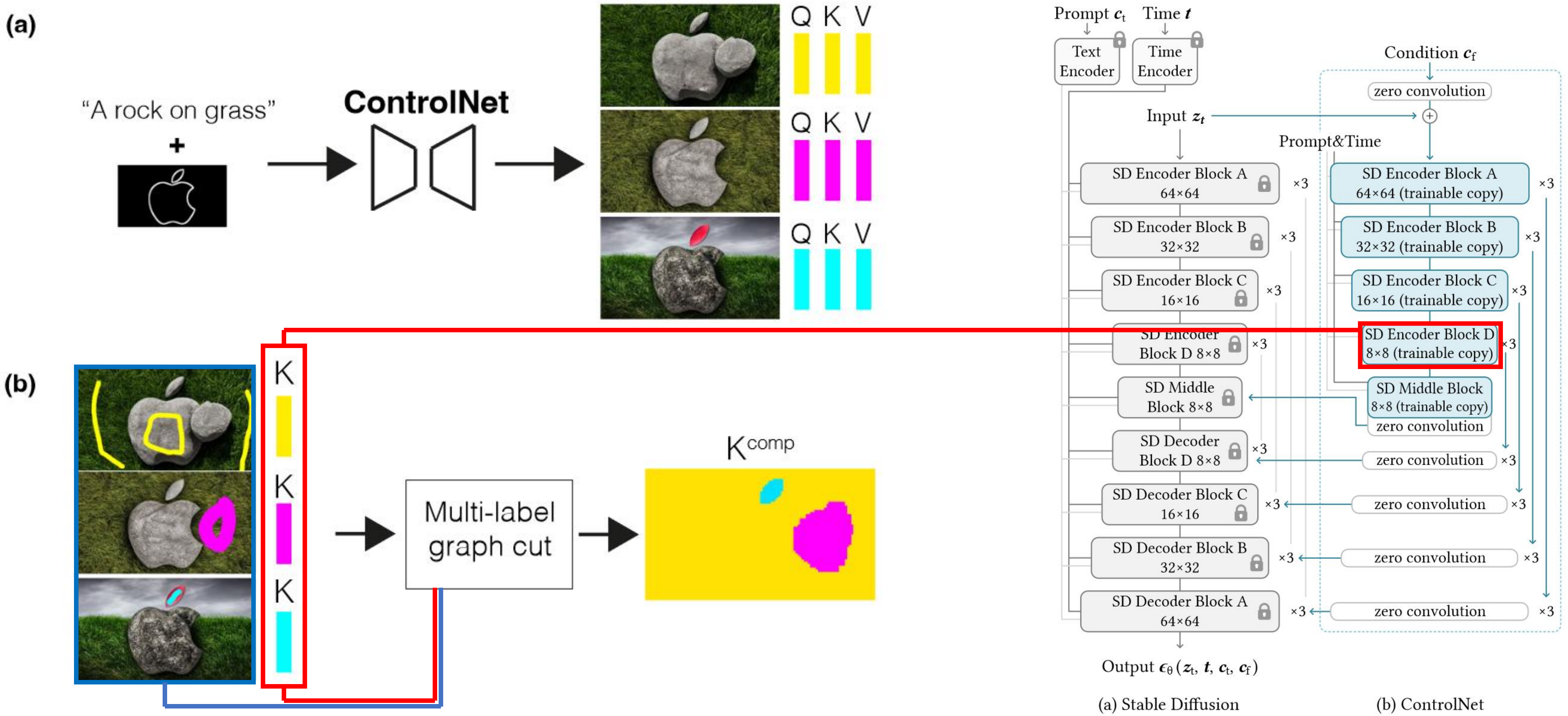
An image stack of N images, labeled 1 to N.
Assign the image label i to pixel $I_O(p)$, $I_O(p) = I_i(p)$.

From sparse to dense: optimization





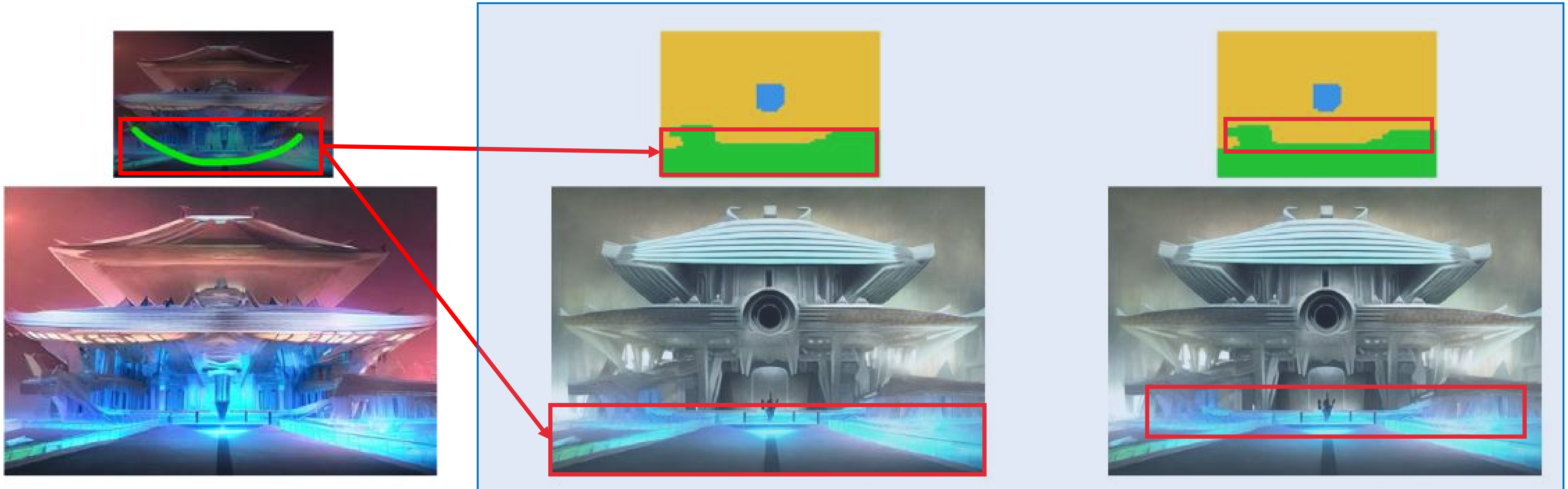
Step1: Segmentation with Feature-Space Graph Cut



"K represents features of the generated image"

Step1: Segmentation with Feature-Space Graph Cut

Optimization: to minimize the **energy cost function**.



Obeying the conditional control is highly appreciated.

A natural seam is highly appreciated.

Segmentation with Feature-Space Graph Cut (Skip)

Energy Cost for Optimization: pixel-scale layer label assignment

$$E_{\text{total}}(L) = \sum_{p} \underbrace{E(p, L_p)}_{\text{Unary costs}} + \sum_{p,q} \underbrace{E(p, q, L_p, L_q)}_{\text{Pairwise costs}},$$

S: Stroke L: Label p, q: pixel from I_O

$\boxed{C \approx 10^6}$

$$\underline{E(p, L_p)} = \begin{cases} C & \text{if } S(p,i) = 1 \text{ and } L_p \neq i \\ 0 & \text{otherwise,} \end{cases}$$

$\boxed{\lambda \approx 10^2}$ $\boxed{\sigma \approx 10^1}$

$$\underline{E(p, q, L_p, L_q)} = \begin{cases} \sum_{i=1}^{N} \lambda e^{-\frac{|f_i(p) - f_i(q)|}{2\sigma}} & \text{if } L_p \neq L_q \\ 0 & \text{otherwise,} \end{cases}$$

Encourage the segmentation boundaries to align with the common semantic edges across all images(Ks).

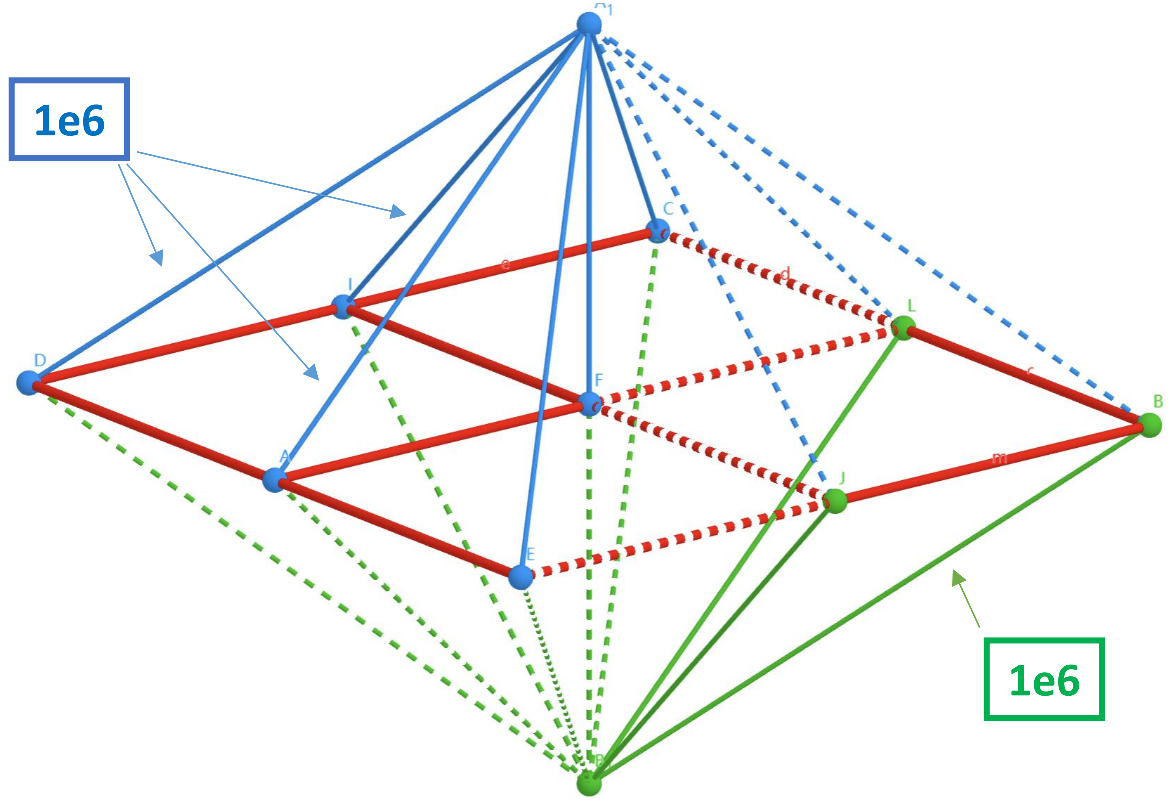
$f_i(p)$ is a feature vector derived from the key features K of image i at location p. To capture the most important features, $f_i(p)$ consists of the top-10 PCA components of K at location p.

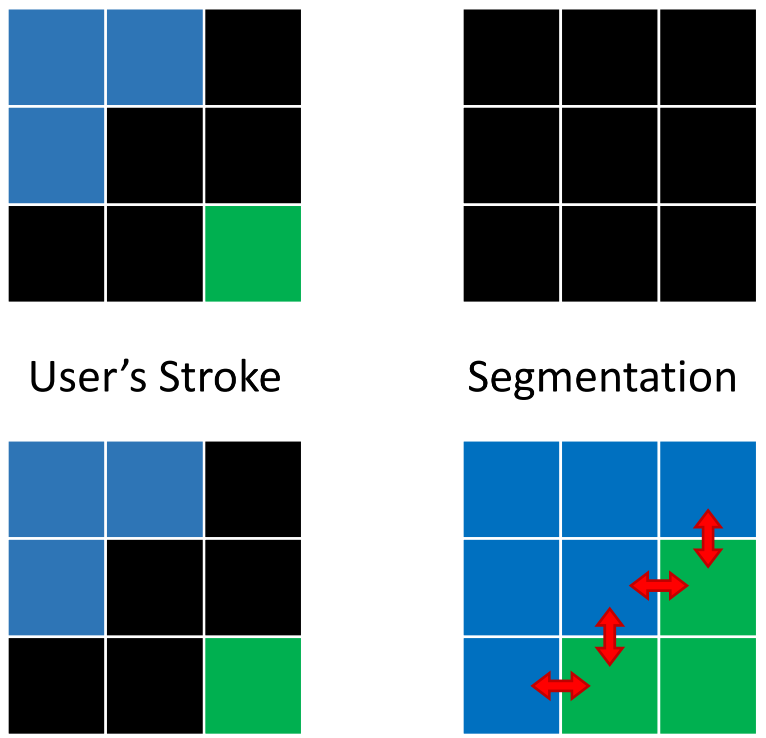
Segmentation with Feature-Space Graph Cut: Max-Flow/Min-Cut

Energy Cost for Optimization: pixel-scale layer label assignment

$$E_{\text{total}}(L) = \sum_{p} \underbrace{E(p, L_p)}_{\text{Unary costs}} + \sum_{p,q} \underbrace{E(p, q, L_p, L_q)}_{\text{Pairwise costs}},$$

Suppose that there are only two labels.



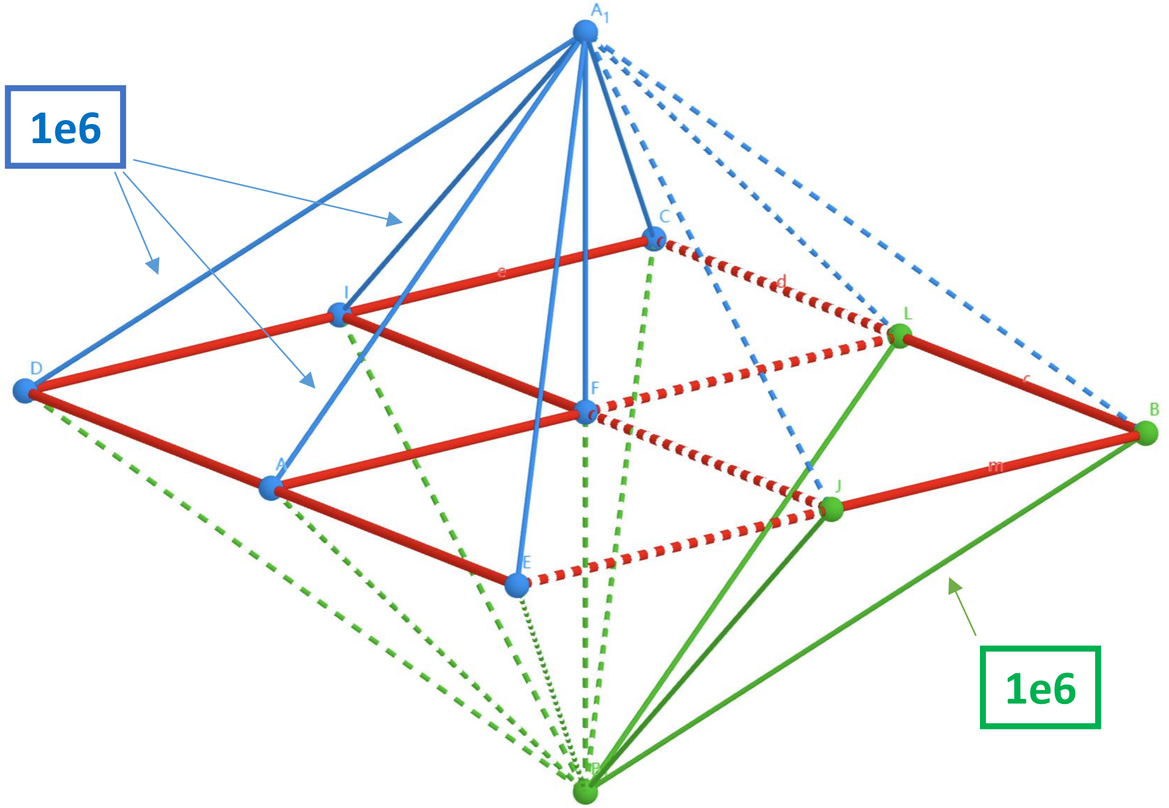


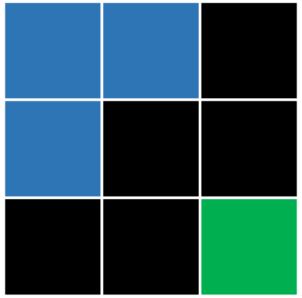
Segmentation with Feature-Space Graph Cut: Max-Flow/Min-Cut

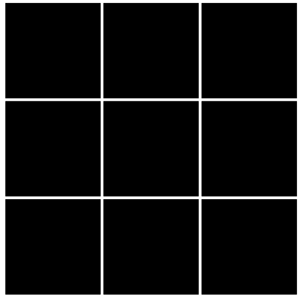
Energy Cost for Optimization: pixel-scale layer label assignment

$$E_{\text{total}}(L) = \sum_{p} \underbrace{E(p, L_p)}_{\text{Unary costs}} + \sum_{p,q} \underbrace{E(p, q, L_p, L_q)}_{\text{Pairwise costs}},$$

Suppose that there are only two labels.



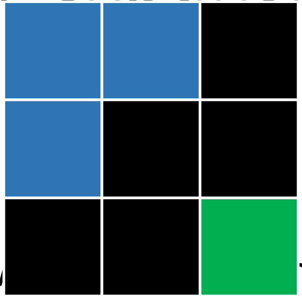


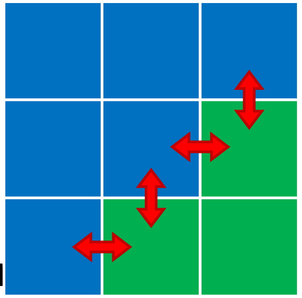


User's Stroke

Segmentation

$+\ \alpha$-expansion

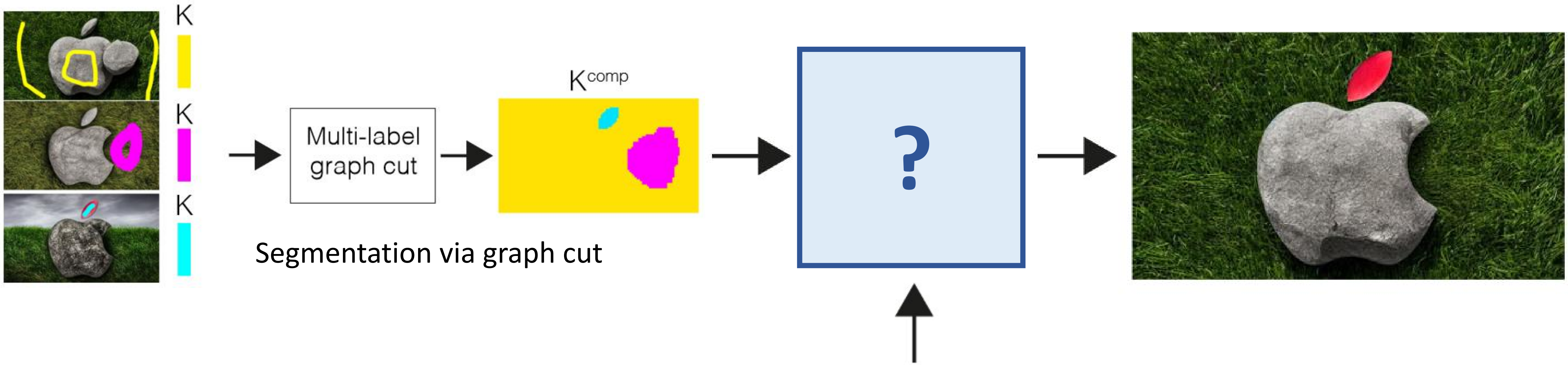


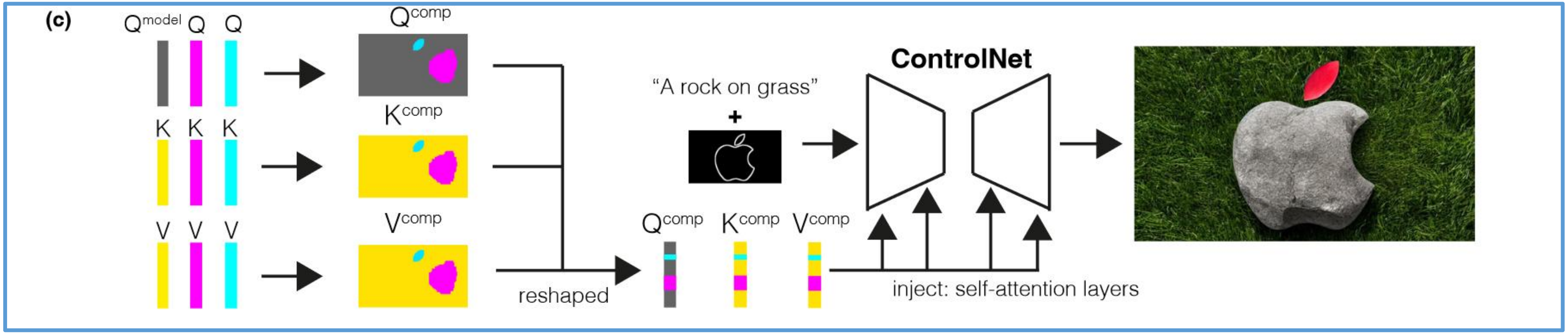


Thus v

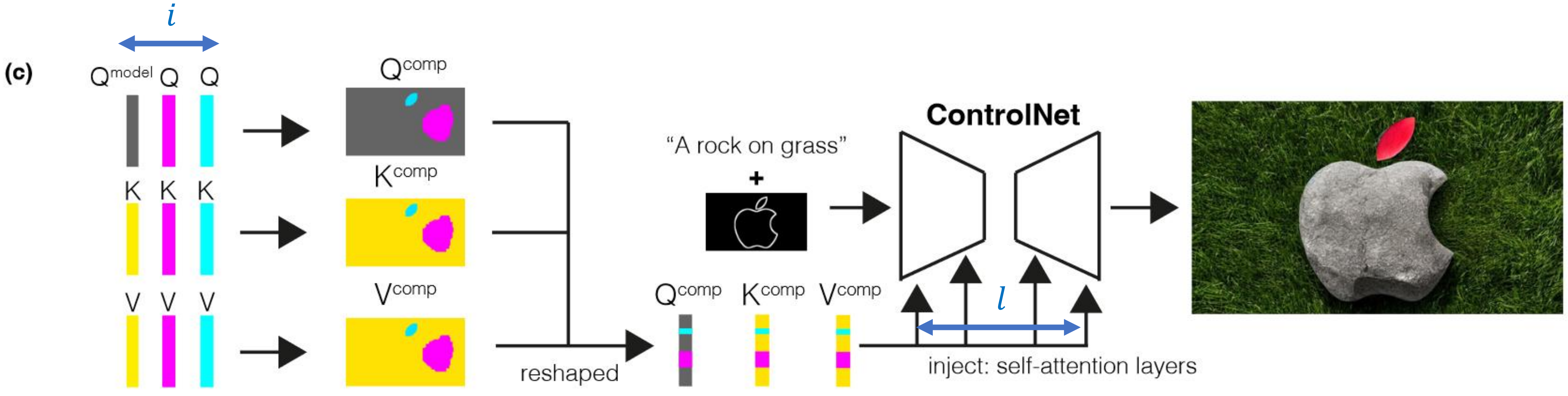
entatio

Generative Photomontage Method: Composition





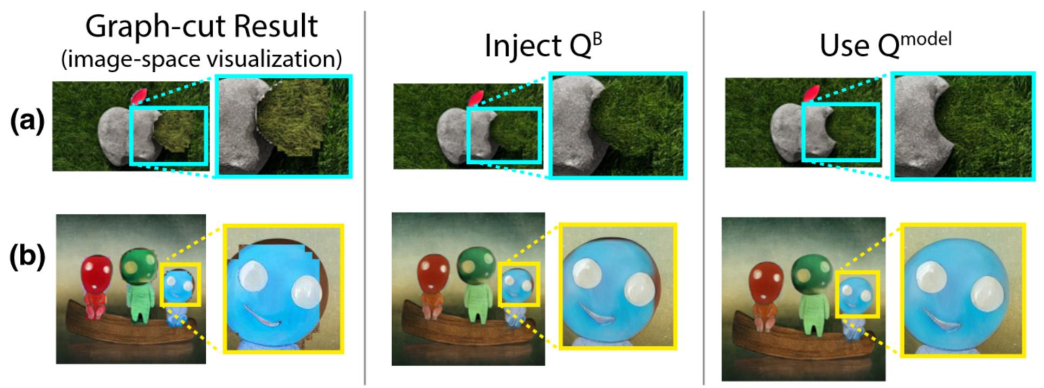
Generative Photomontage Method: Composition



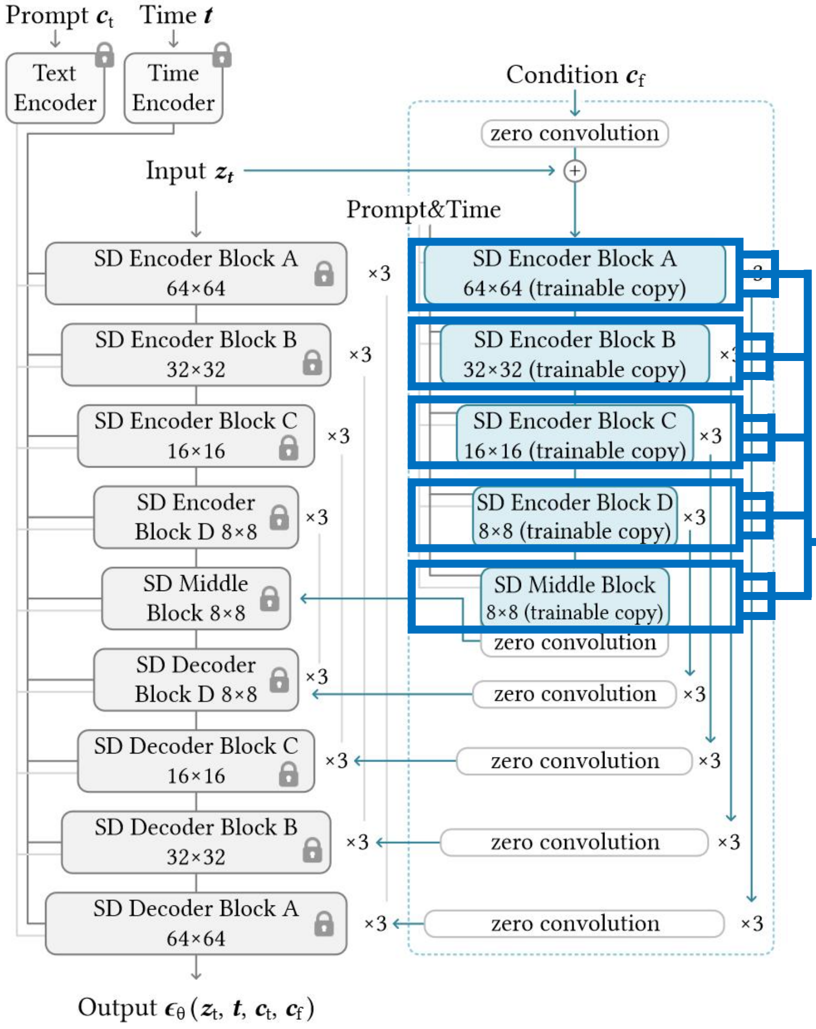
$$Q_l^{\text{comp}} = M_l^B \odot Q_l^{\text{model}} + \sum_{i \neq B} M_l^i \odot Q_l^i,$$

$$K_l^{\text{comp}} = \sum_i M_l^i \odot K_l^i,$$

$$V_l^{\text{comp}} = \sum_i M_l^i \odot V_l^i,$$



Generative Photomontage Method: Composition



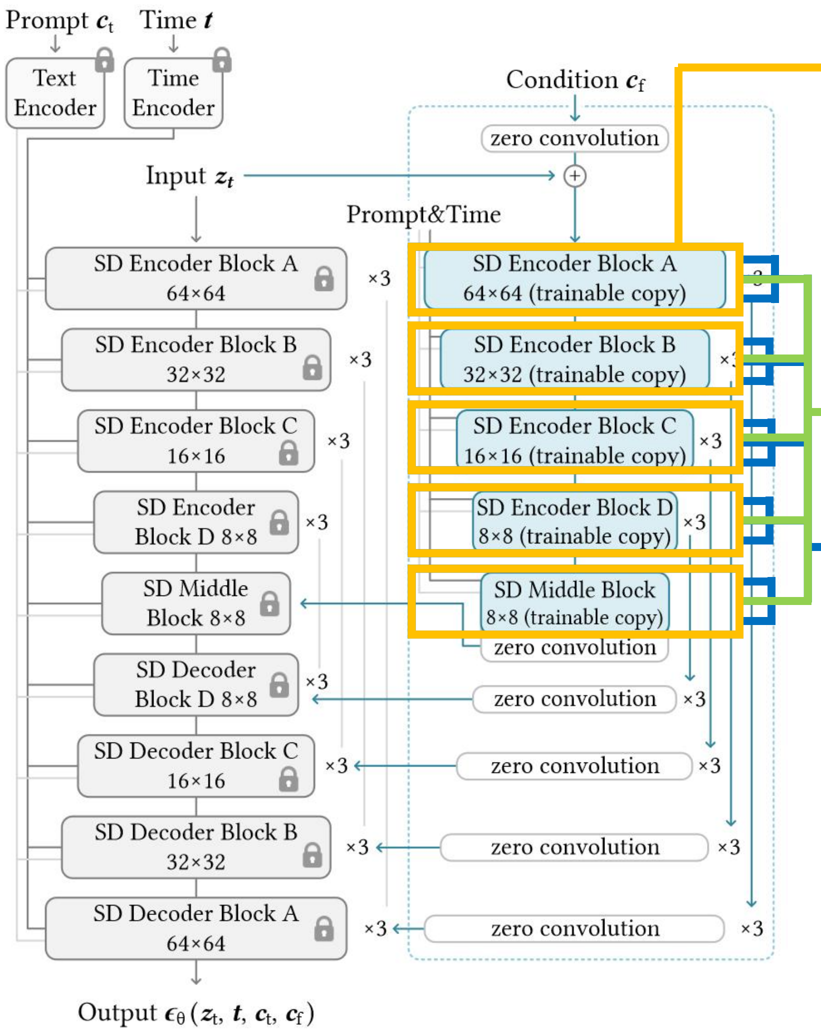
(a) Stable Diffusion (b) ControlNet

$$Q_l^{\text{comp}} = M_l^B \odot Q_l^{\text{model}} + \boxed{\sum_{i \neq B} M_l^i \odot Q_l^i}$$

$$\boxed{K_l^{\text{comp}}} = \sum_i M_l^i \odot K_l^i,$$

$$\boxed{V_l^{\text{comp}}} = \sum_i M_l^i \odot V_l^i,$$

Generative Photomontage Method: Composition



(a) Stable Diffusion (b) ControlNet

$$Q_l^{\text{comp}} = \boxed{\boxed{M_l^B} \odot \boxed{Q_l^{\text{model}}}} + \boxed{\sum_{i \neq B} M_l^i \odot Q_l^i}$$

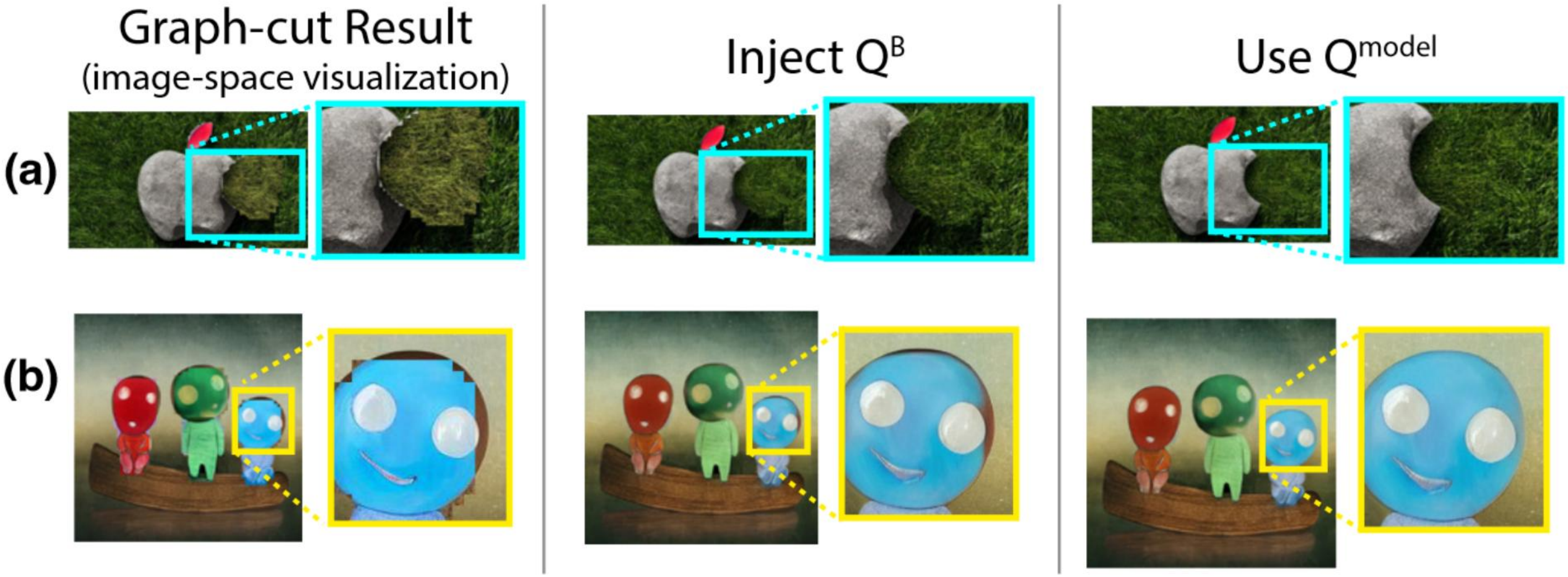
$$\boxed{K_l^{\text{comp}}} = \sum_i M_l^i \odot K_l^i,$$

$$\boxed{V_l^{\text{comp}}} = \sum_i M_l^i \odot V_l^i,$$

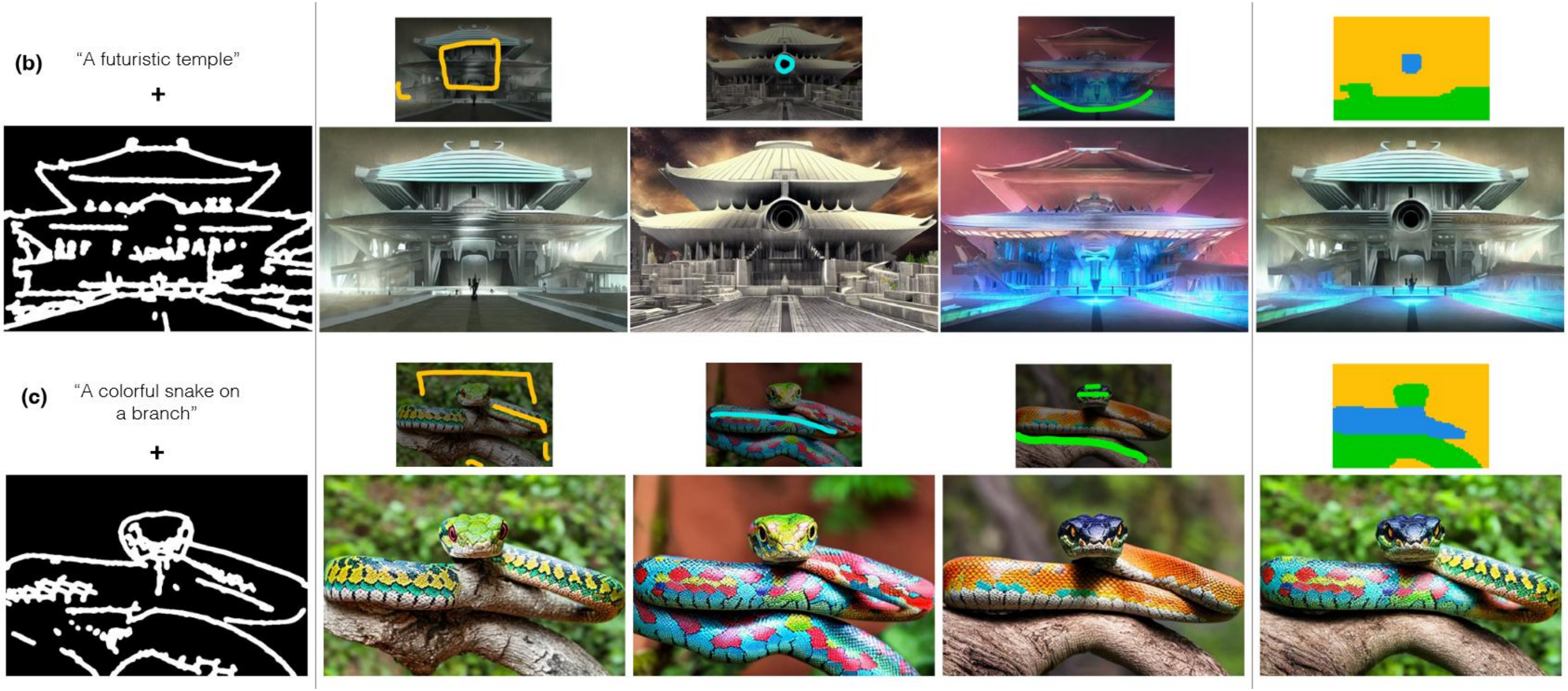
Generative Photomontage Method: Composition

K^{comp}

"Q influences the image structure, while K and V influence the appearance."

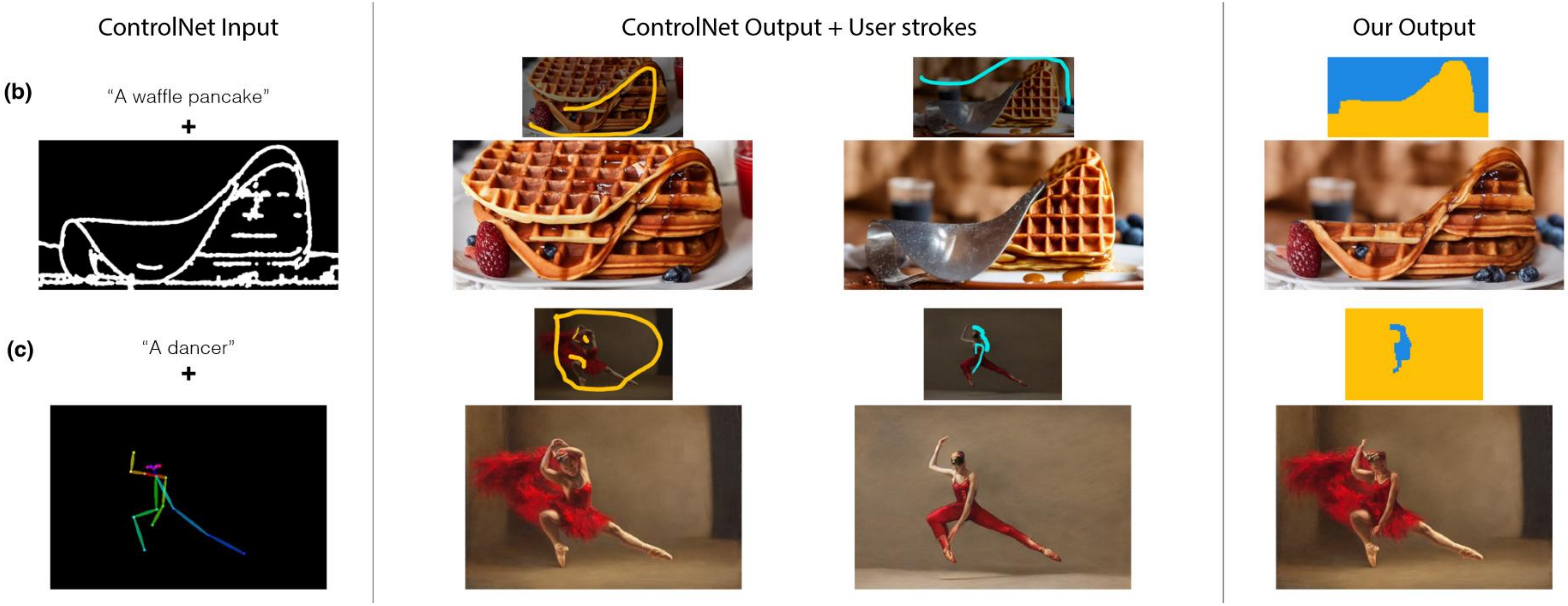


Generative Photomontage Method: Results



Appearance Mixing

Generative Photomontage Method: Results

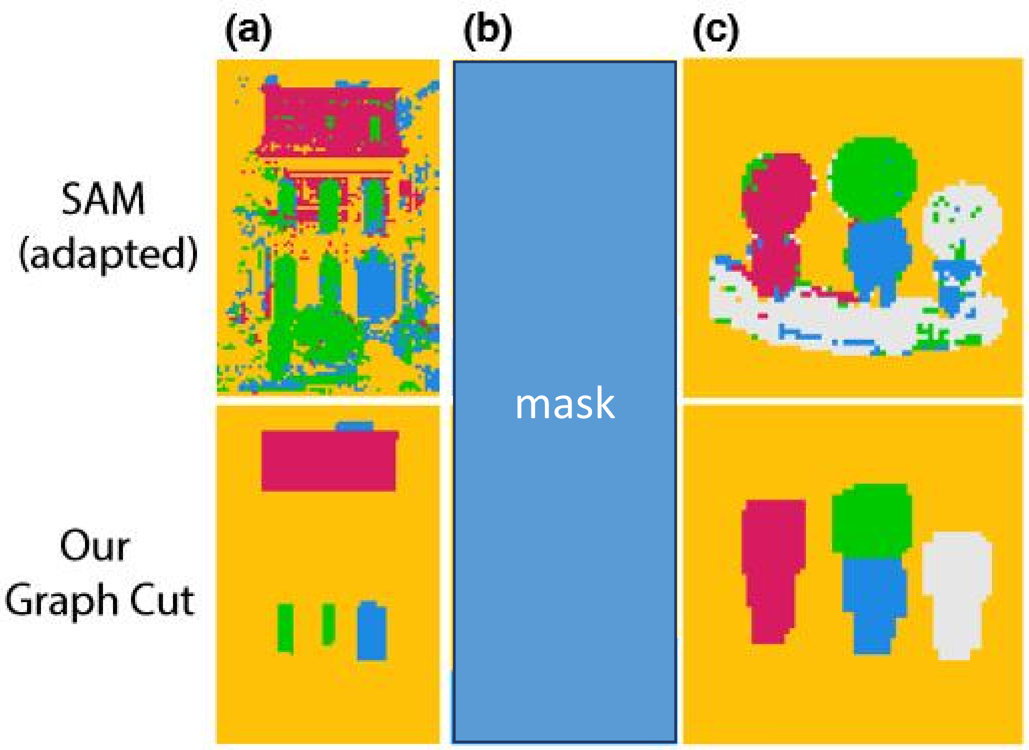


Shape and Artifacts Correction

Generative Photomontage Method: Results

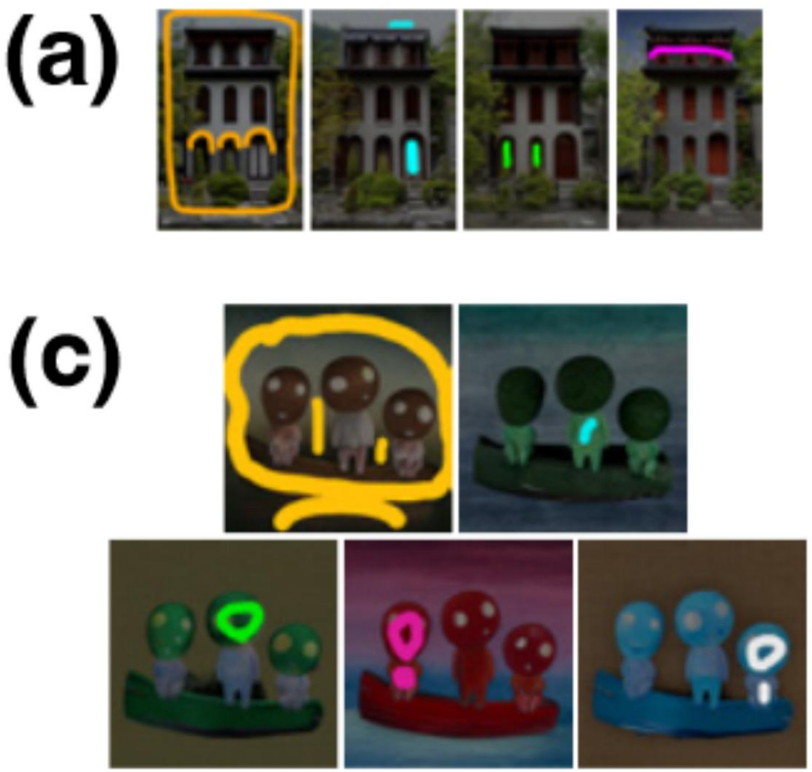
Prompt Alignment

Generative Photomontage Method: Evaluation

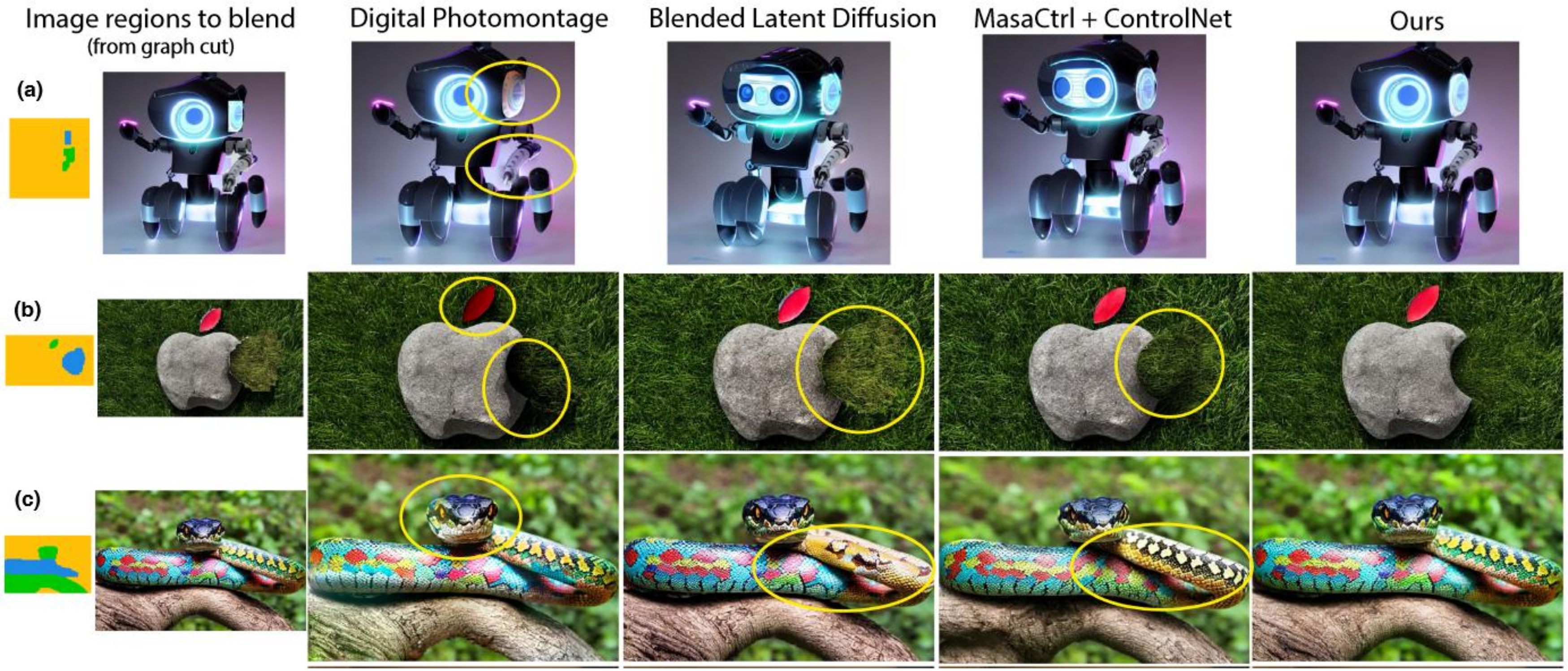


Graph Cut vs SAM

Input Strokes



Generative Photomontage Method: Evaluation



Global appearance

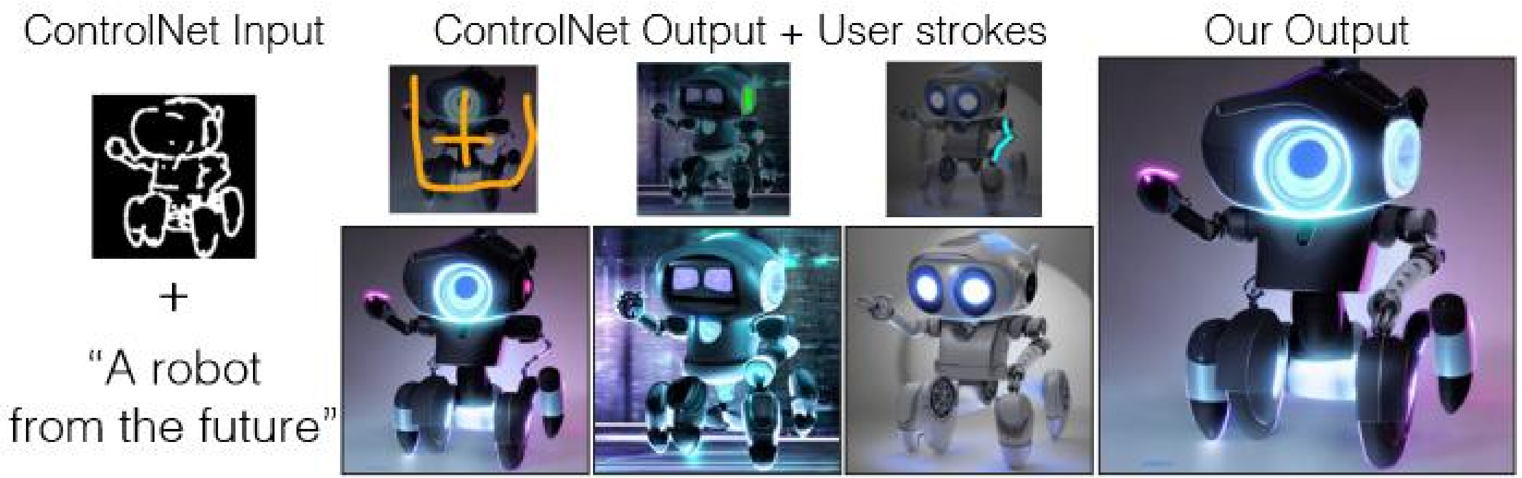
	Ours	Interactive Digital Photomontage	Blended Latent Diffusion	BLD + MultiDiffusion	MasaCtrl + ControlNet	Cross-Domain Compositing	Deep Image Blending	GP-GAN	Collage Diffusion
Masked LPIPS ↓	0.104	**0.085**	0.187	0.188	0.198	0.380	0.252	0.220	0.244
PSNR ↑	**23.44**	21.12	21.17	20.66	19.50	20.31	18.35	18.11	20.95
Seam Gradient Score min: 0.256, avg: 0.337, max: 0.427	0.335	0.312	0.386	0.326	0.341	0.394	0.301	*0.207*	*0.487*

Local appearance fidelity

Generative Photomontage Method: Conclusion

Pros:

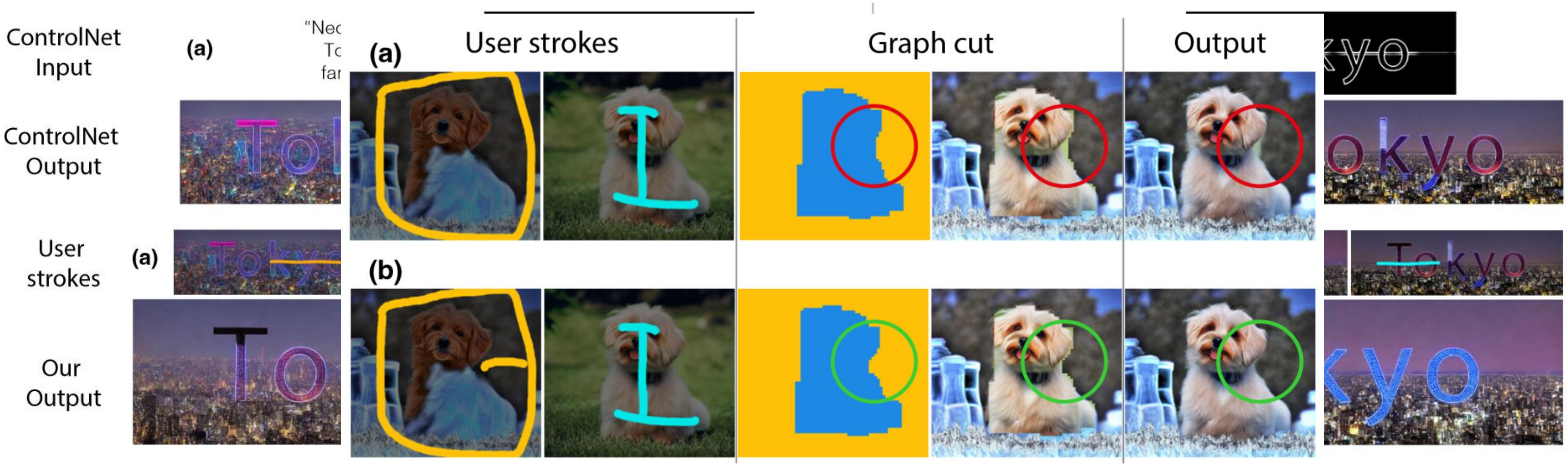
- **Treat ControlNet output as intermediate outputs**, avoid complex algorithms.
- **Training-free** method!
- Gives users more **fine-grained control** over the final output.



Generative Photomontage Method: Conclusion

Cons:

- If the target object has a curvy outline, it may require additional user strokes to obtain a finer boundary.
- If the images differ significantly in scene structure, it will rely more on the user to select proper regions to form a valid scene.
- Needs explanation to the selection of $K,\ Q_l^{model}$ and the calculation of evaluation standards.



Reference

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, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In IEEE International Conference on Computer Vision (ICCV), 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In IEEE International Conference on Computer Vision (ICCV), 2023.

Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In ACM Multimedia (MM), 2019.

Thanks for listening!

Presenter: XuShenghan
2025.03.23