

LayerTracer: Cognitive-Aligned Layered SVG Synthesis via Diffusion Transformer

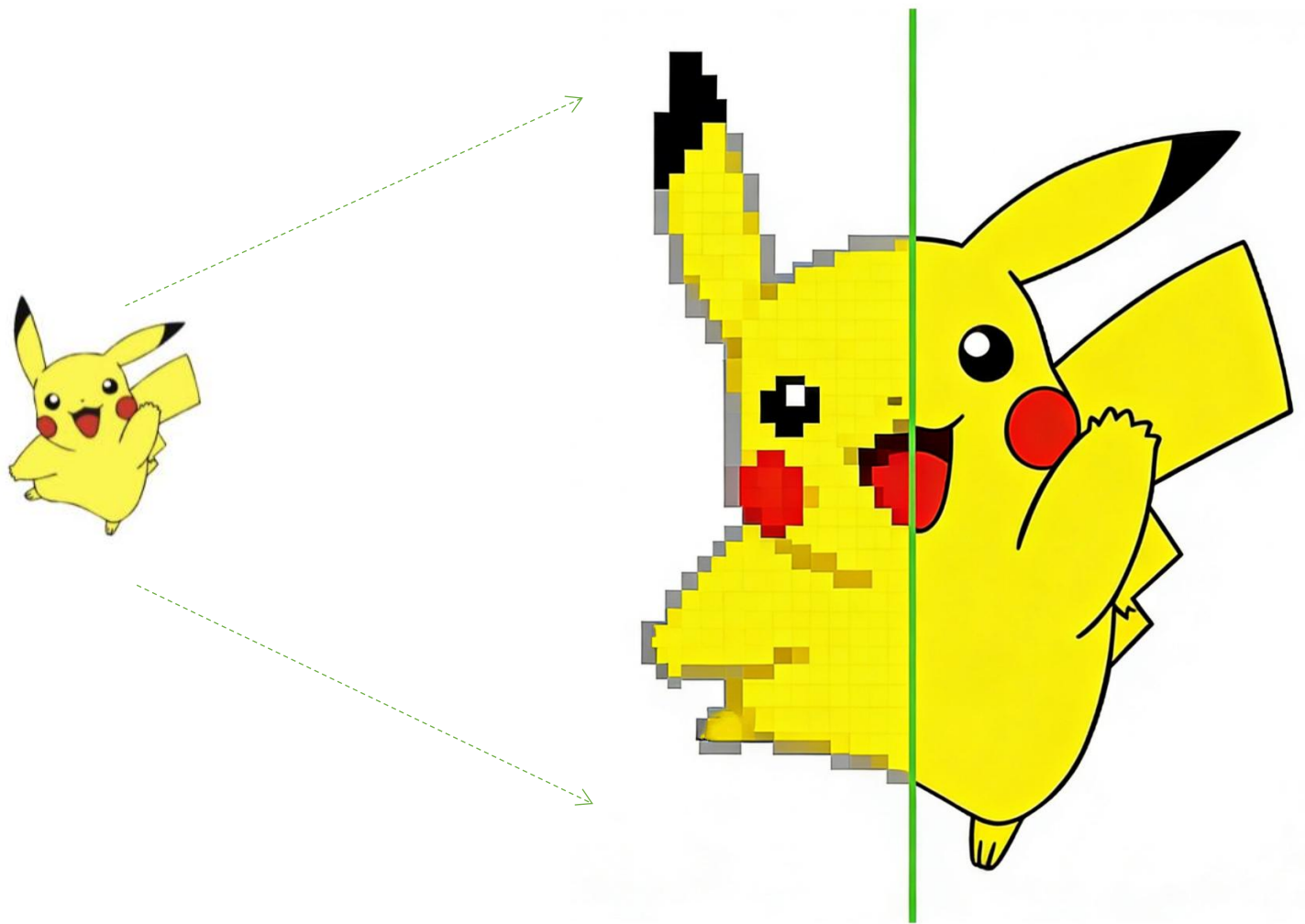
Yiren Song Danze Chen Mike Zheng Shou

Show Lab, National University of Singapore

ICCV 2025

Presenter: Jiangyue Zeng
2025.12.14

What is SVG?



What is SVG?

```
<!-- 根元素<svg>: 定义画布的显示尺寸、坐标系 -->
<svg width="200" height="200" viewBox="0 0 100 100" xmlns="http://www.w3.org/2000/svg">
  <!-- 图形元素1: 皮卡丘的脸 (圆形) -->
  <circle
    cx="50"      <!-- 圆心x坐标 (基于viewBox的坐标系) -->
    cy="50"      <!-- 圆心y坐标 -->
    r="40"       <!-- 圆的半径 -->
    fill="#FFDE00" <!-- 填充色 (皮卡丘黄色) -->
  />

  <!-- 图形元素2: 左腮红 (圆形) -->
  <circle cx="30" cy="60" r="8" fill="#FF6B6B" />

  <!-- 图形元素3: 右腮红 (圆形) -->
  <circle cx="70" cy="60" r="8" fill="#FF6B6B" />

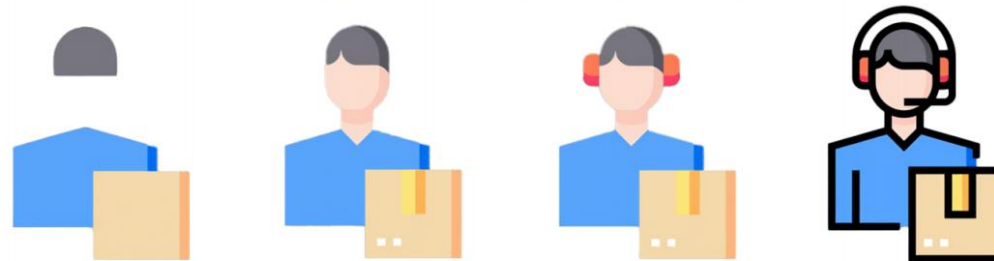
  <!-- 图形元素4: 左眼 (圆形) -->
  <circle cx="40" cy="45" r="5" fill="#000" />
</svg>
```



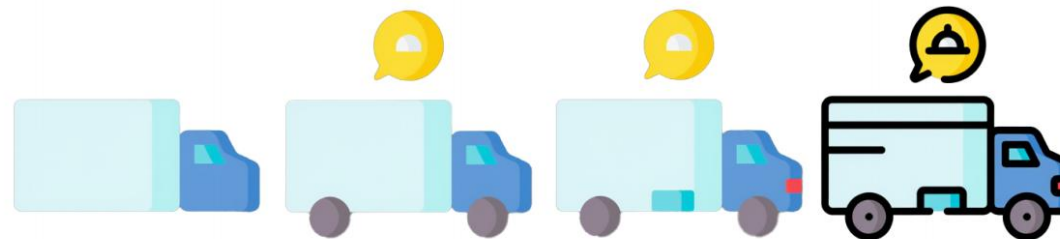
What is SVG?



an icon of iced lemon tea



an icon of a logistics customer service



An icon of a delivery truck with a notification bell

Text2SVG

An icon of man in a suit holding a gavel.



Image2SVG



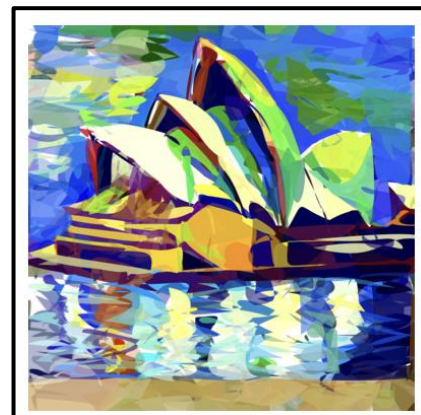
Text2SVG



CLIPDraw



VectorFusion

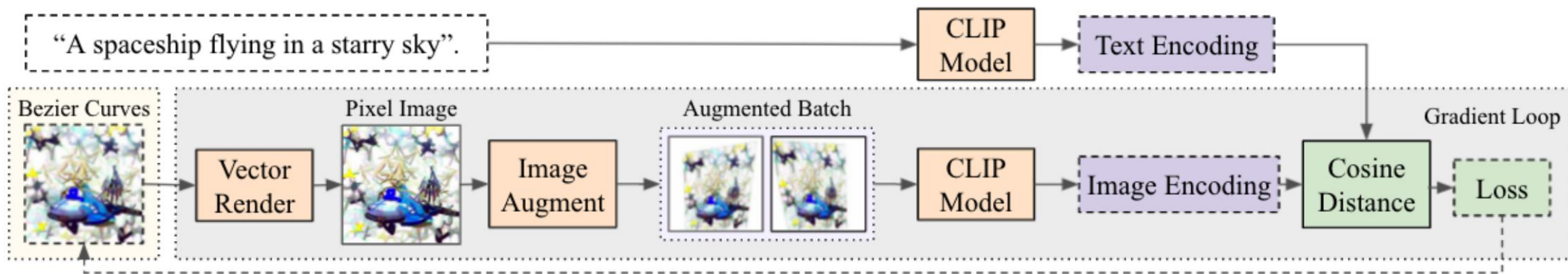


SVGDreamer



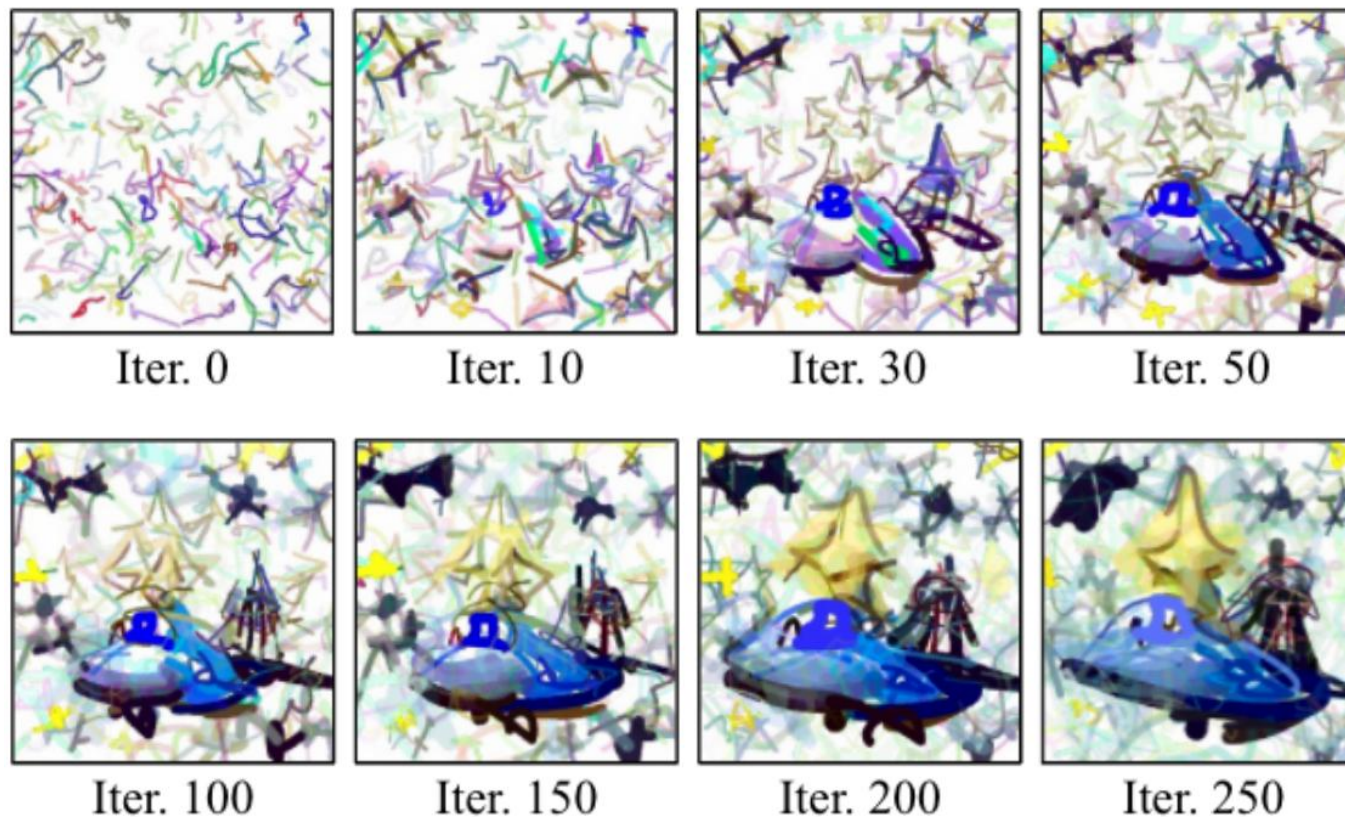
DiffSketcher

Text2SVG: CLIPDraw (NeurIPS 2022)

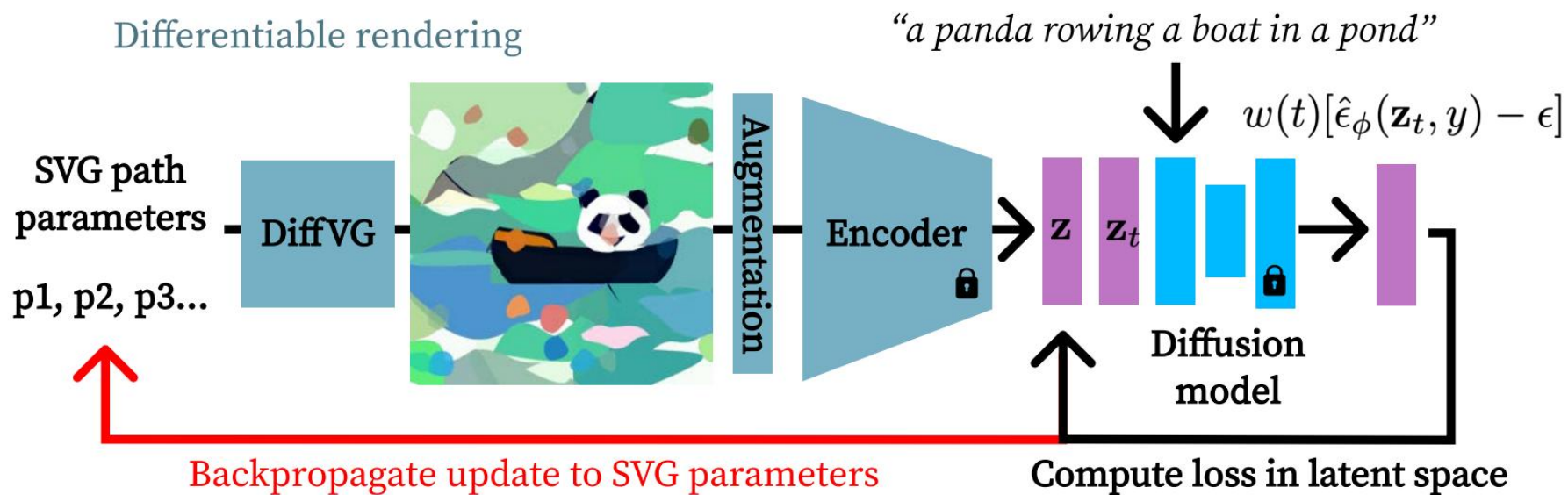


$$L_{\text{CLIP}}(\theta) = 1 - \cos(E_{\text{img}}(R(\theta)), E_{\text{text}}(y))$$

Text2SVG: CLIPDraw (NeurIPS 2022)



Text2SVG: VectorFusion (CVPR 2023)

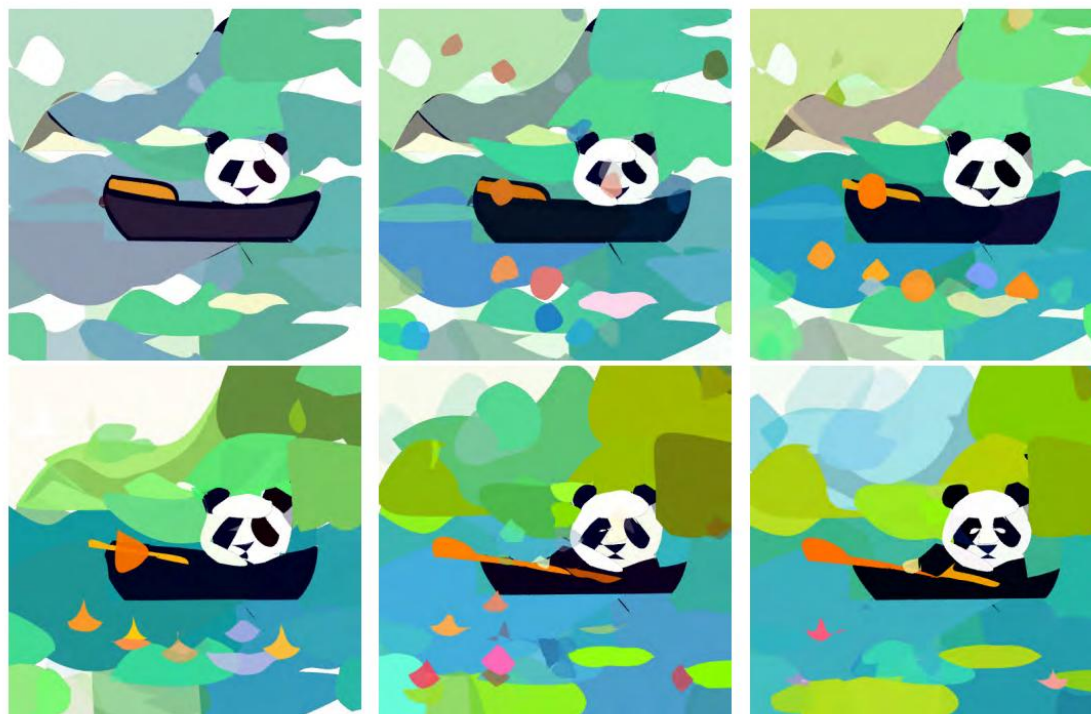


$$\nabla_{\theta} L_{\text{SDS}} = \mathbb{E}_{t, \epsilon} \left[w(t) \left(\hat{\epsilon}_{\phi}(x_t; y, t) - \epsilon \right) \frac{\partial x}{\partial \theta} \right]$$

Text2SVG: VectorFusion (CVPR 2023)



A panda rowing a boat in a pond

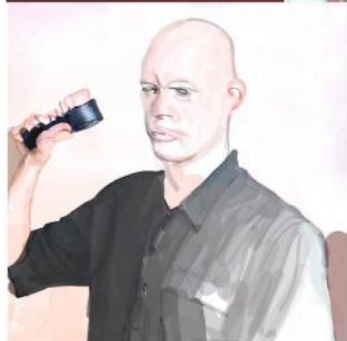


Text2SVG: DiffSketcher (NeurIPS 2023) & SVGDreamer (CVPR 2024)

Illustrator, a women
stand beside a window,
holding a cafe



An icon of man in a
suit holding a gavel.



Prompt

DiffSketcher

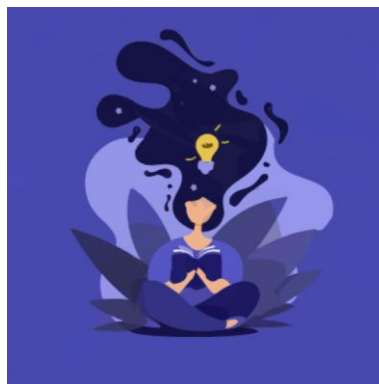


SVGDreamer

Image2SVG

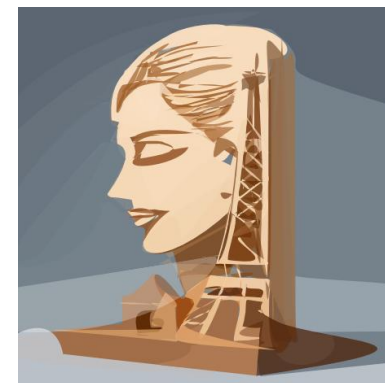


DiffVG



LIVE

“+”



O & R

“-”

Image2SVG

- **DiffVG Vectorization (SIGGRAPH ASIA 2020):**

Differentiable rasterizer

$$L(\theta) = \|R(\theta) - I\|^2$$

- **LIVE (CVPR 2022)** : More emphasis to the loss near the edges, impose constraints on the angles



Differentiable Vector Graphics Rasterization for Editing and Learning

Towards Layer-wise Image Vectorization

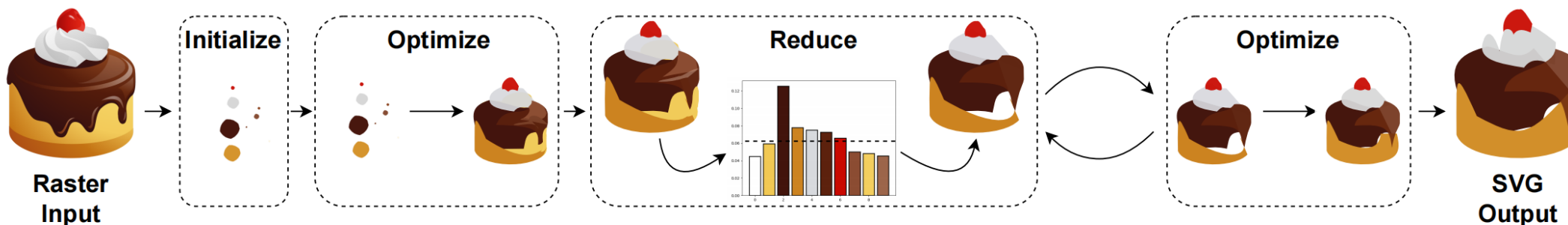
Image2SVG

- **DiffVG Vectorization (SIGGRAPH ASIA 2020):**

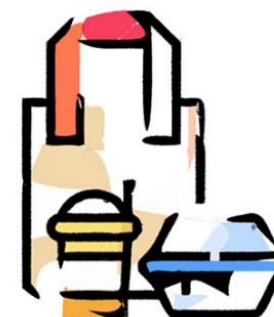
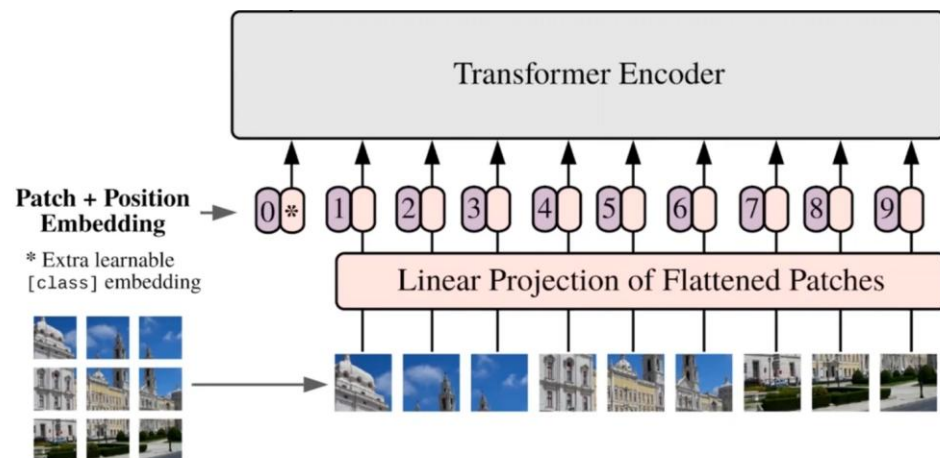
Differentiable rasterizer

$$L(\theta) = \|R(\theta) - I\|^2$$

- **O&R (AAAI 2024) :** DBSCAN Clustering Initialization,
Geometric reg: penalizes self-intersections, sharp corners...

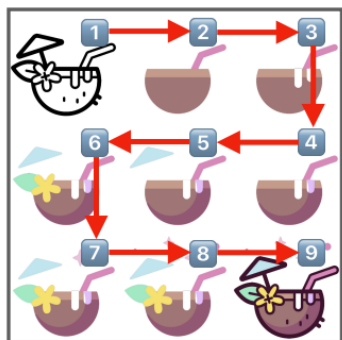


DiT

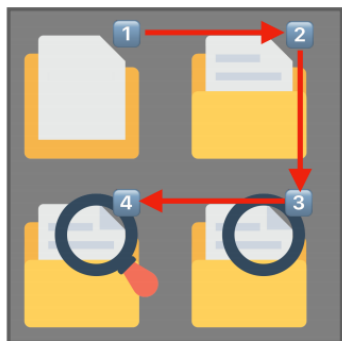


How to generate beautiful SVG?

Overall Architecture

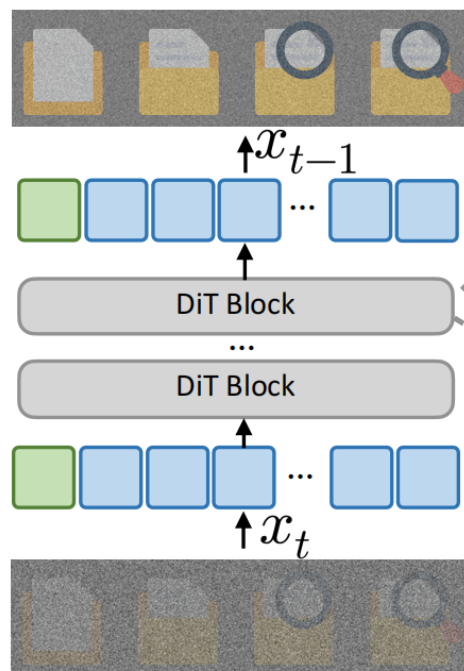


“This image illustrates the nine-step creation process of an SVG icon illustration, arranged in a 3*3 grid. The top left tile is a line art sketch of a icon, depicting its basic shape and structure. Subsequent tiles gradually add color and detail to the illustration...”



“The image is a 2x2 grid of folder icons. The top-left shows a folder with a plain document, while the top-right depicts a folder containing a document with visible text lines...”

Serpentin Dataset Construction



Layer-wise Model Training

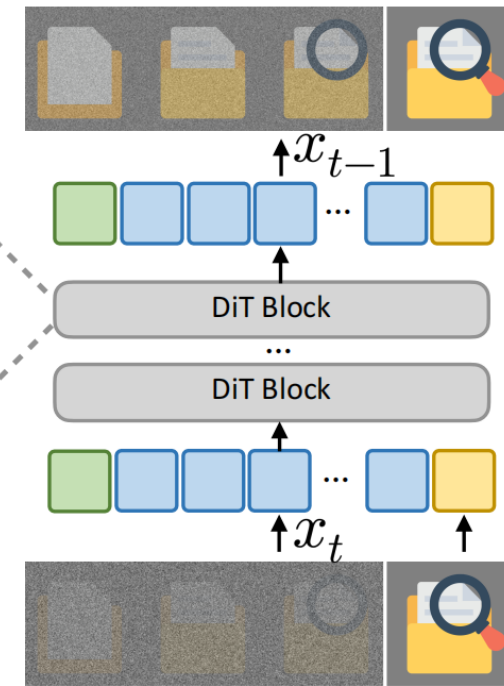
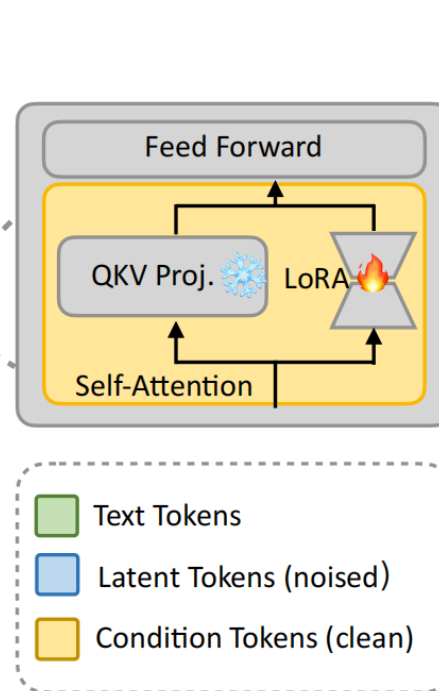
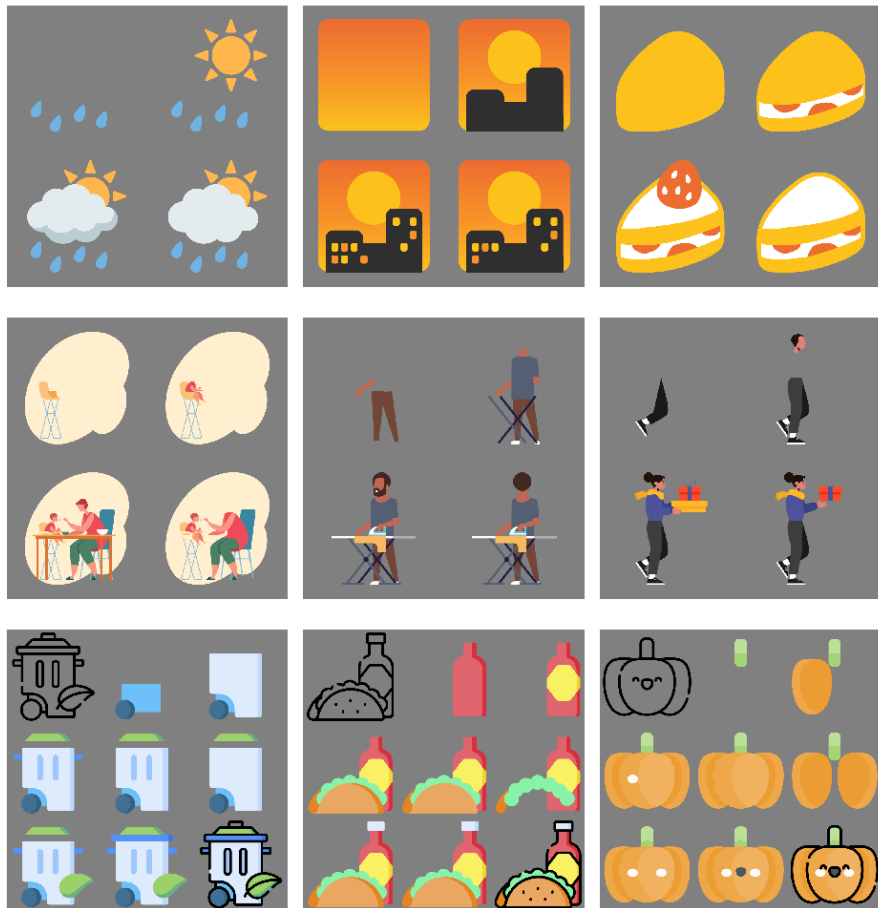


Image2Layers Model Training

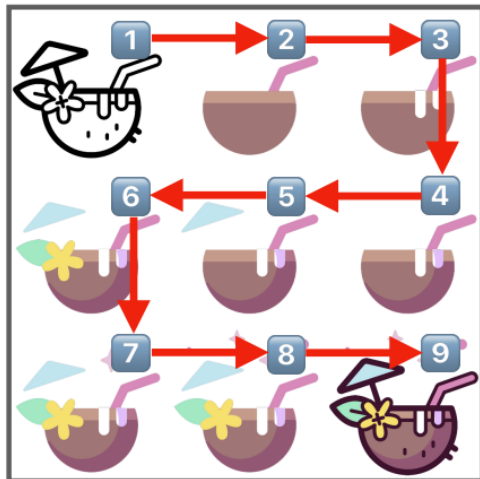
- Text Tokens
- Latent Tokens (noised)
- Condition Tokens (clean)

Serpentin Dataset Construction

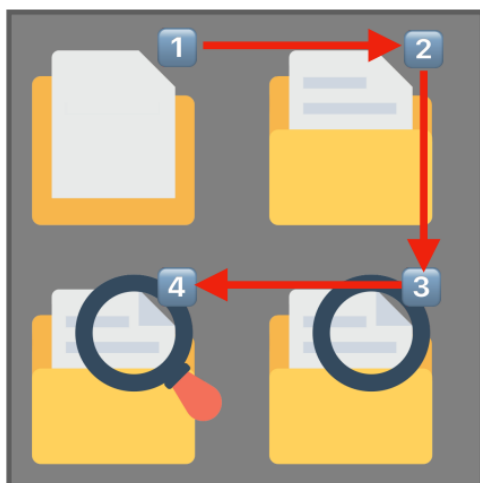


- **Sources:**
internet / vendors / designer-created assets
- **Human-in-the-loop**
- **20K samples across 3 styles:**
Black outline icons (3K)
Illustrations (2K)
Emojis (15K)

Serpentin Dataset Construction



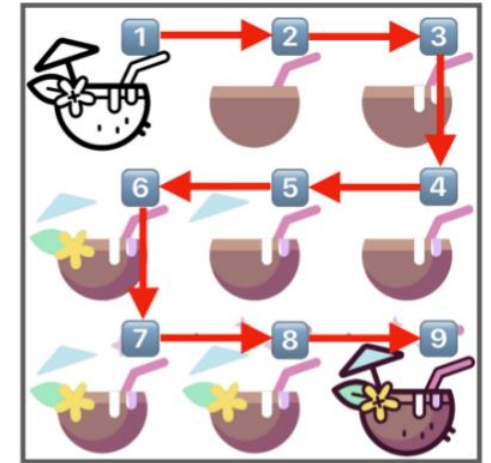
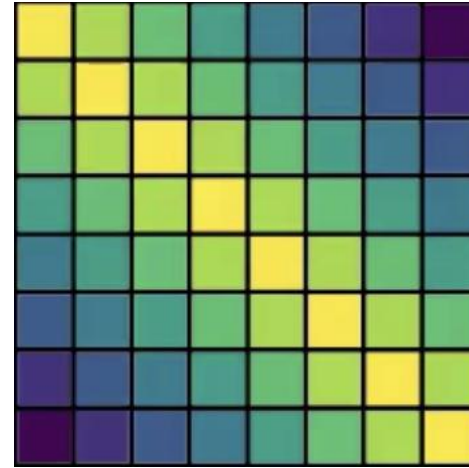
“This image illustrates the nine-step creation process of an SVG icon illustration, arranged in a 3*3 grid. The top left tile is a line art sketch of a icon, depicting its basic shape and structure. Subsequent tiles gradually add color and detail to the illustration...”



“The image is a 2x2 grid of folder icons. The top-left shows a folder with a plain document, while the top-right depicts a folder containing a document with visible text lines...”

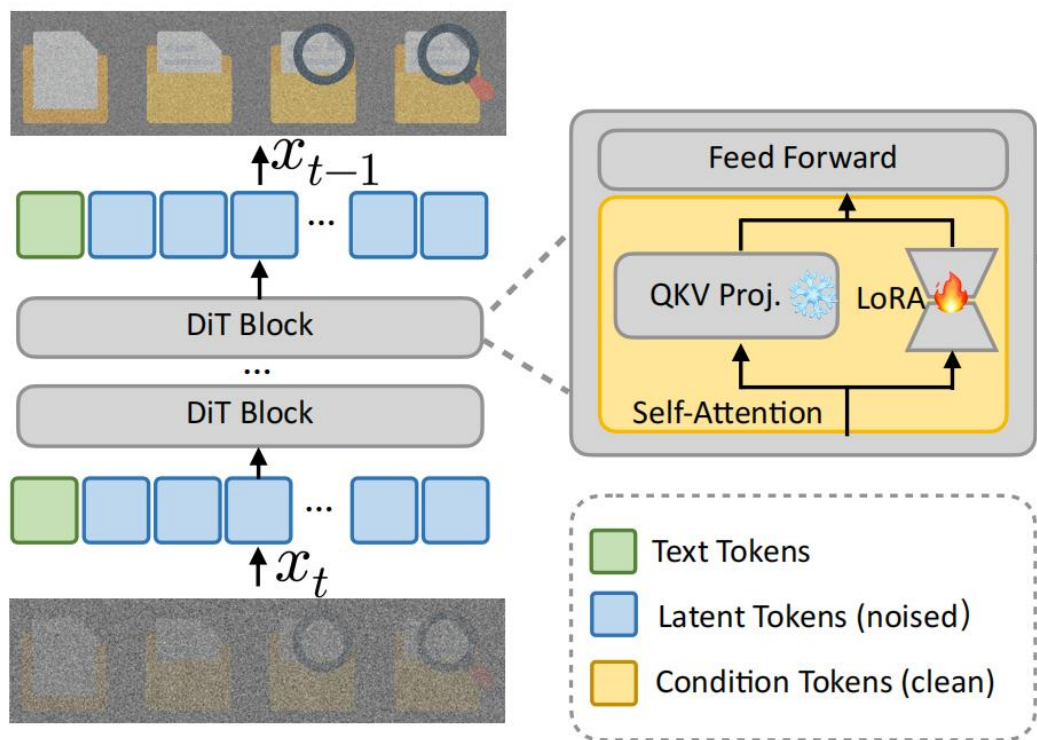
- 9-frame → 3×3 grid
4-frame → 2×2 grid
- **Serpentine order:**
to keeps temporally adjacent
steps spatially adjacent
- Outline to color blocks to details

Serpentin Dataset Construction



- Tends to focus on spatial neighbors
- Serpentine layout: **time adjacency to space adjacency**
- Crucial for **coherent sequence generation** (validated in ablation)

Layer-wise Model Training



Layer-wise Model Training

- **LoRA Fine-tuning**

$$W = W_0 + \Delta W$$

- Keep W_0 fixed; learn lightweight ΔW

- **Input:** text prompt $\rightarrow$

Output: multi-step raster grid

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, p_t(X|\epsilon), p(\epsilon)} \|v_{\Theta}(X, t) - u_t(X|\epsilon)\|^2$$

Layer-wise Model



... illustration, a woman with silver hair sitting in a cozy corner of a library ...



... a four-step illustration, a man standing under a road sign checking his phone ...



... a nine-step SVG creation of a modern office building icon in blue shades... The icons progress from a simple outline to more detailed versions with glass windows ...

Image2Layers Model Training

- To reverse **engineering layers**
- Given a **final raster image**
- Predict earlier **construction steps**
- Reframe hierarchical vectorization as **process prediction**

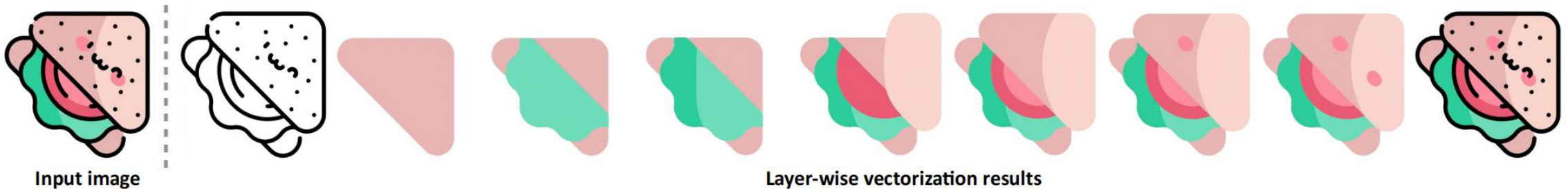
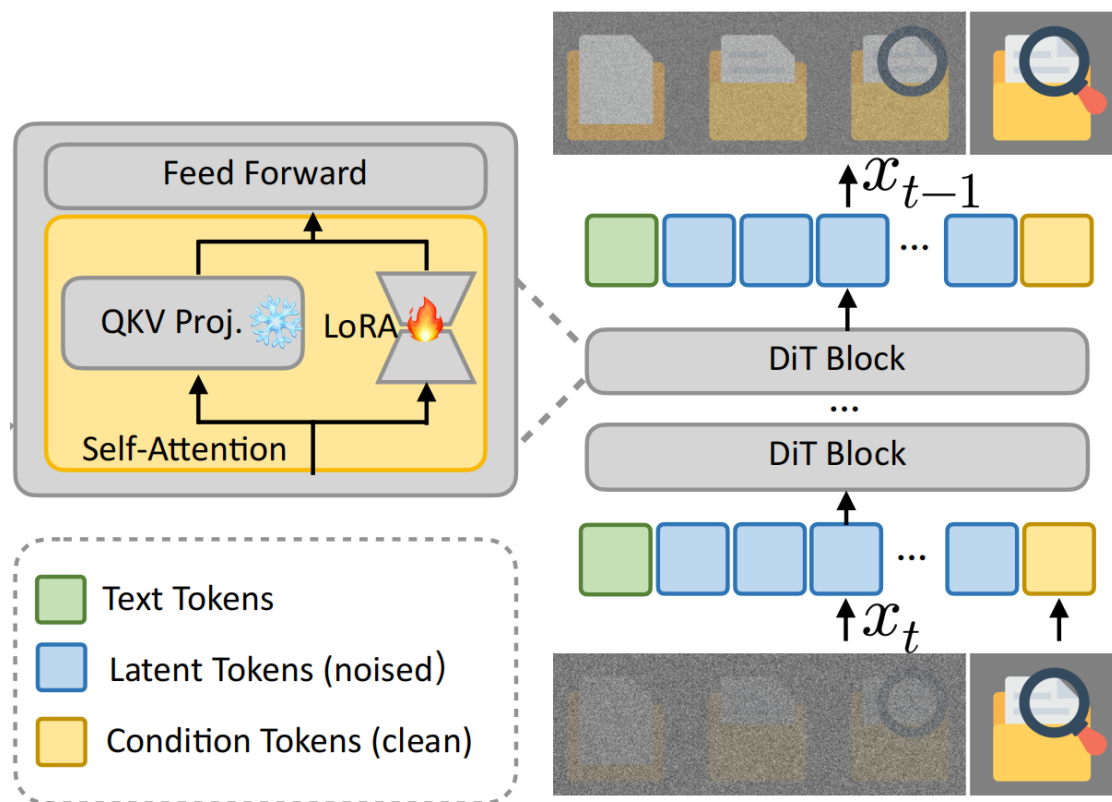


Image2Layers Model Training

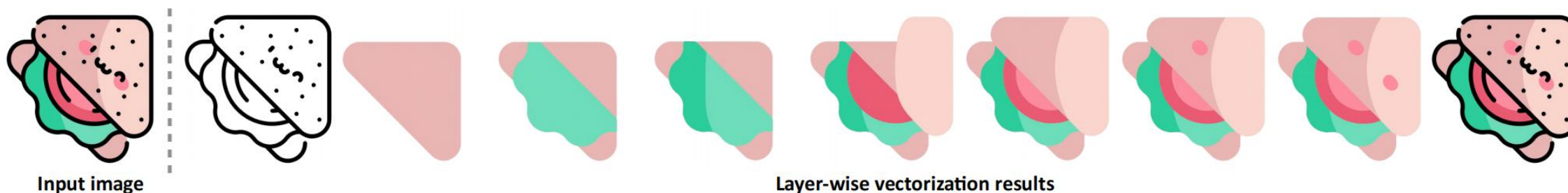


- **Image - Conditioned Generation**
- Build 2×2 / 3×3 grids
take final frame as context
- VAE encode context $\rightarrow C_I$
append to denoising latents

$$\text{MMA}([X; C_I; C_T]) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V$$

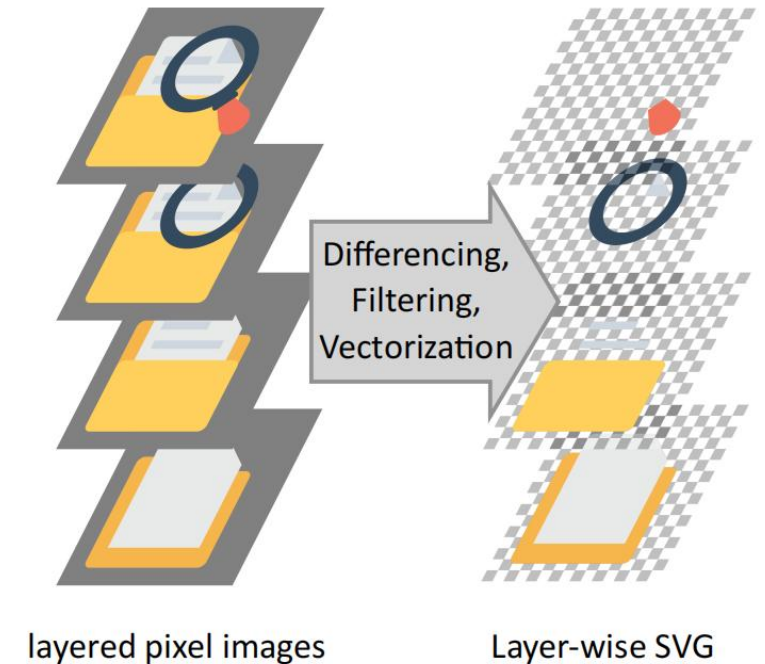
Image2Layers Model

- **Inference: Predict Earlier Layers**
- Input: reference image
- Output: **a plausible creation sequence**
- Enables **cognitive-aligned layered vectorization**

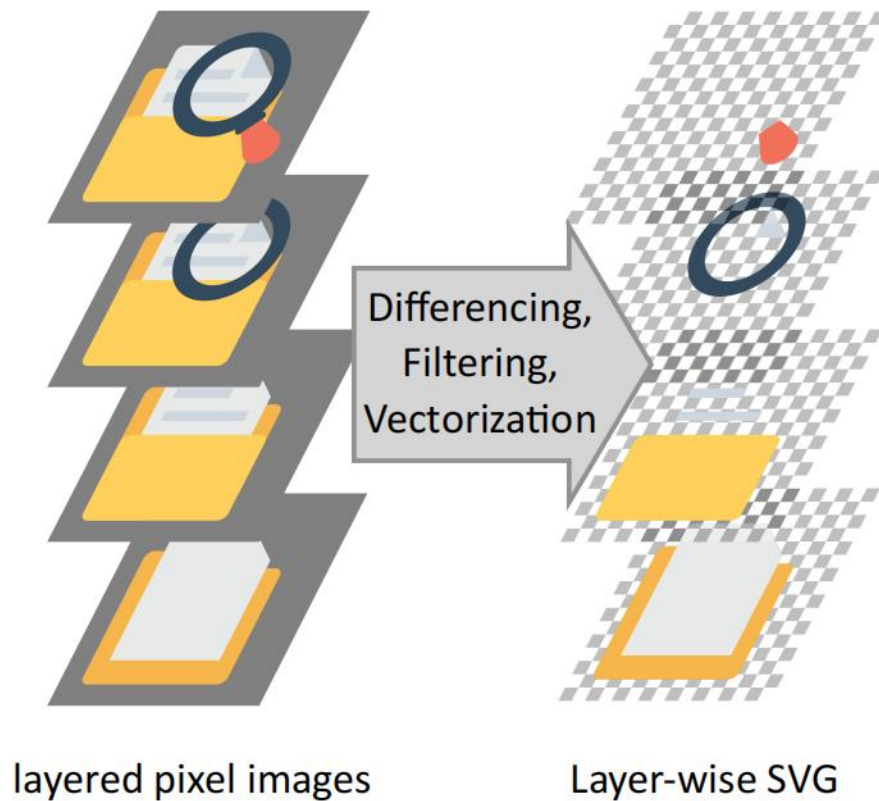


Layer-wise Vectorization

- Split grid into **individual frames**
- **Outline icons:**
grayscale, Gaussian blur, Otsu threshold
- **Adjacent-frame differencing:**
abs diff, binarize, morphology
- Save each change mask as **transparent PNG**



Layer-wise Vectorization



- Vectorize each mask with **vtracer**
- Tune for detail vs. paths
- Merge all vector layers → layered, editable SVG
- **VTracer**: Raster-to-SVG
 - (1) color clustering → (2) contour tracing → (3) curve fitting & simplification

Setup

- Backbone: FLUX 1.0 dev (DiT)
- Resolutions: 1056×1056 (3×3), 1024×1024 (2×2)
- LoRA: rank 256, batch 16, lr 1e-3, 20k steps
- Image2Layers: merge weight 1.0, then +20k steps
- Compute: 1×A100 80GB
- Baselines: SVGDreamer / VecFusion / DiffSketcher; DiffVG / LIVE / O&R

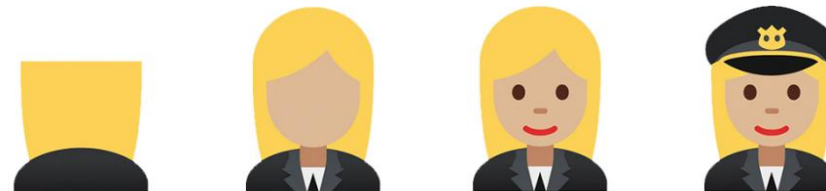
Text2SVG



... a yellow face with a frown, a blue tear rolling down its cheek, sadness ...



... illustration, a woman with silver hair sitting in a cozy corner of a library ...

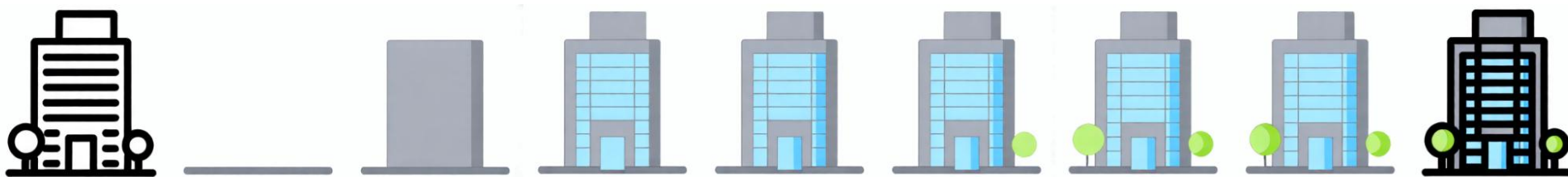


... an emoji, a blonde woman with short hair and a black blazer, pilot's hat...



... a four-step illustration, a man standing under a road sign checking his phone ...

Text2SVG



... a nine-step SVG creation of a modern office building icon in blue shades... The icons progress from a simple outline to more detailed versions with glass windows ...



... a set of nine minimalistic icons of a delivery truck with a white arrow to complex arrangements of orange boxes with directional arrows...

Text2SVG: Gradient

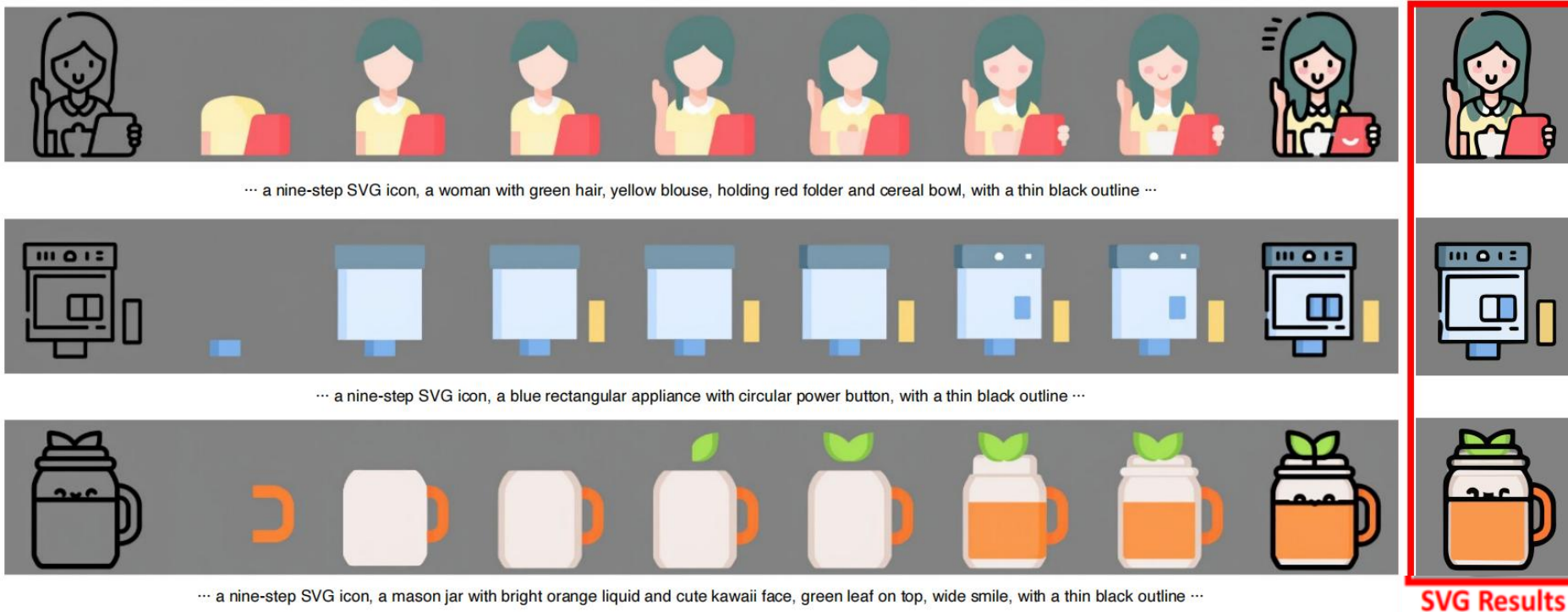


... an emoji, an anxious emoji with a sweat drop on its forehead, set against a gradient green-to-yellow background ...



... an illustration, a mother in a white top and purple skirt tenderly holds her baby ...

Text2SVG



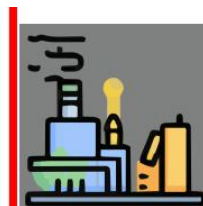
Text2SVG



... a nine-step SVG icon, a cityscape line drawing with tall modern building and power plant, with a thin black outline ...



... a nine-step SVG icon, a soft gold wristwatch with a thin black outline ...

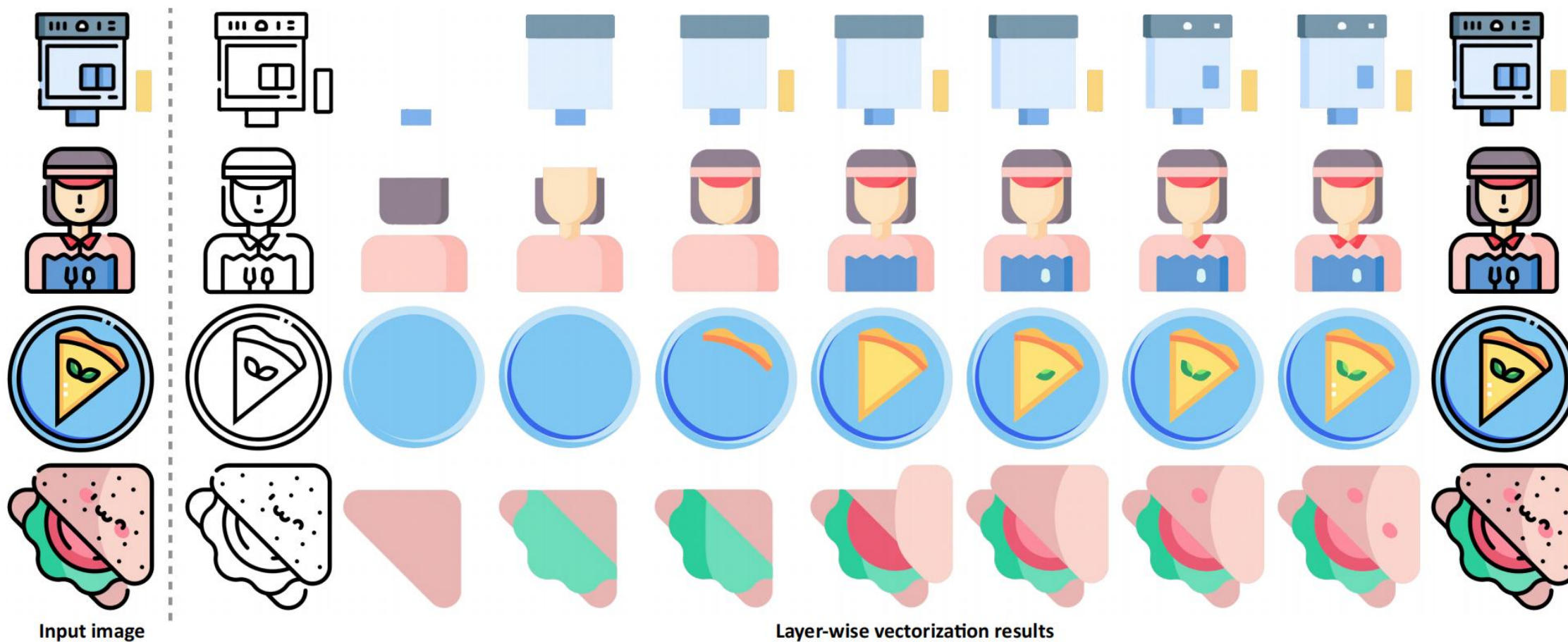


SVG Results

Quantitative Comparison

Methods	Semantic consistency	SVG complexity		
	CLIP- Score↑	Time Cost(s)↓	No. Paths↓	layer- wise
Vecfusion [17]	31.10	4668	128.00	False
SVGDreamer [57]	32.68	5715	512.00	False
DiffSketcher [56]	31.47	3374	512.00	False
Ours	33.76	27	35.39	True

Image2SVG



Comparison

An icon of a pink jerry can with a small green leaf symbol.

An icon of sashimi sushi.

Illustrator, a women stand beside a window, holding a cafe

An icon of man in a suit holding a gavel.

Prompt



SVGDreamer



Vecfusion

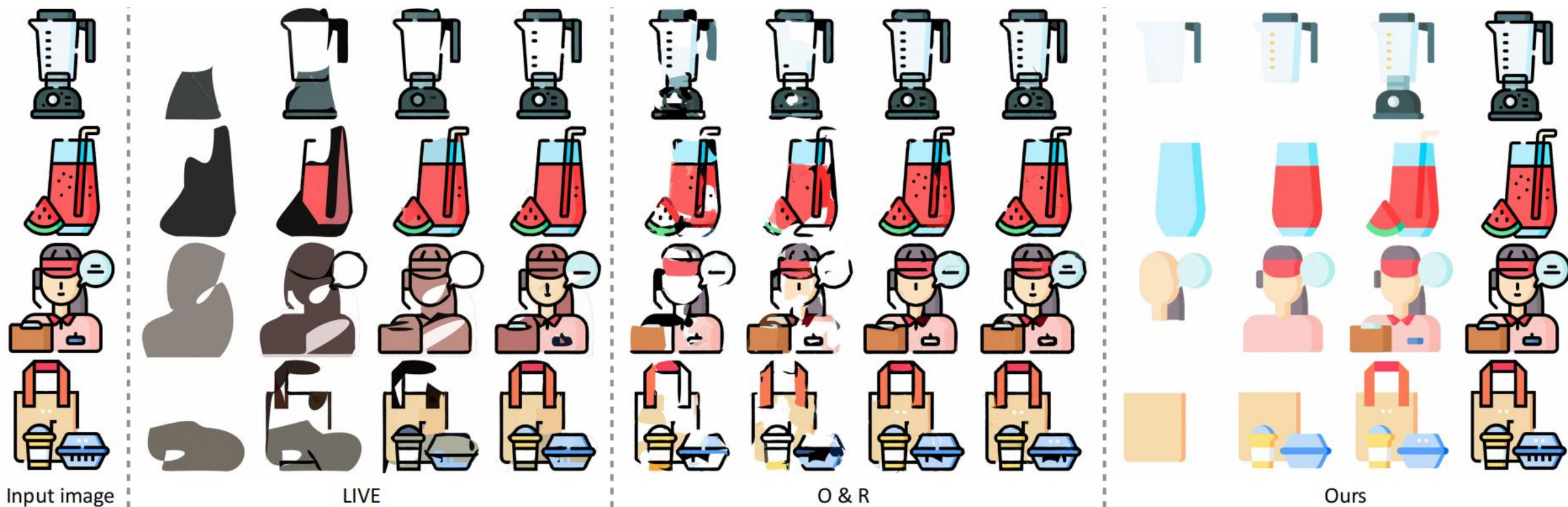


DiffSketcher



Ours

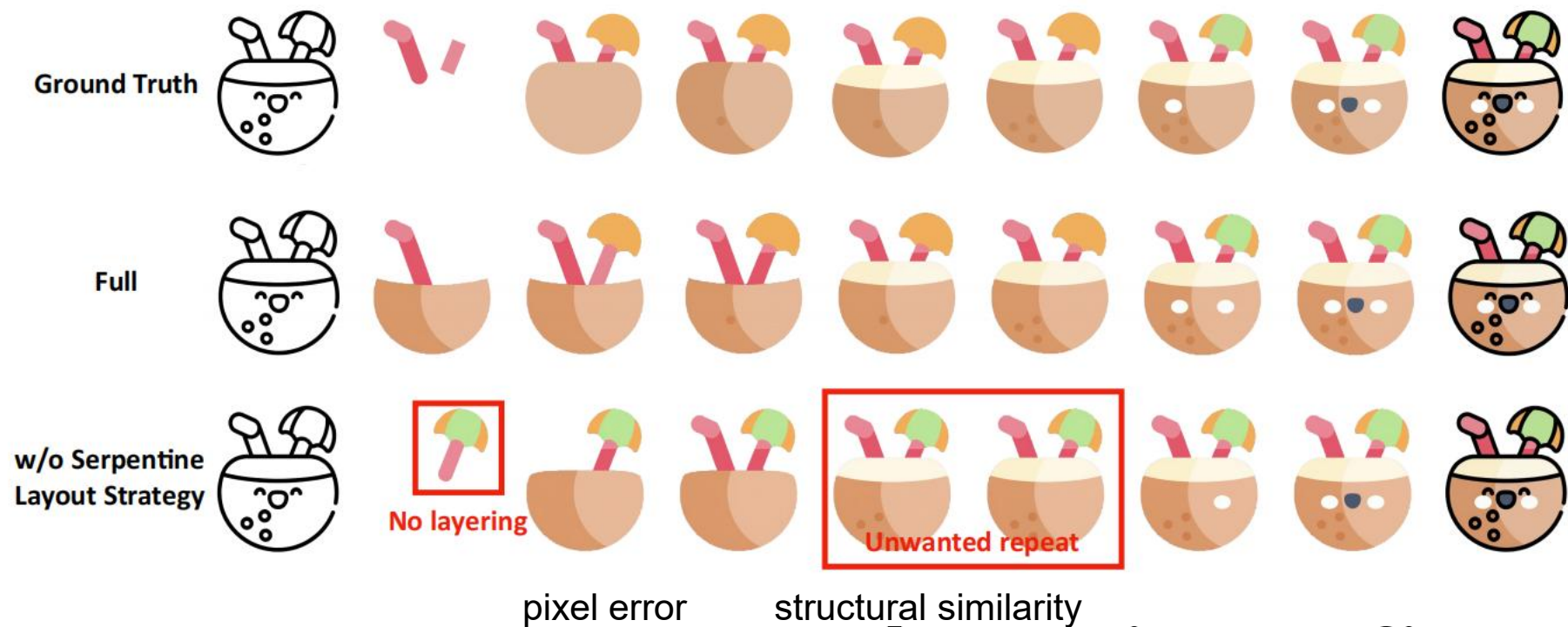
Comparison



Quantitative Comparison

Methods	pixel error MSE↓	SVG complexity		
		Time Cost(s)↓	No. Paths↓	Layer- wise
Diffvg [19]	2.02×10^{-4}	393	256.00	False
LIVE [23]	5.21×10^{-4}	3147	46.00	True
O&R [13]	2.01×10^{-4}	612	64.00	True
Ours	1.96×10^{-4}	34	29.98	True

Ablation



Methods	MSE_train↓	SSIM_train↑	MSE_test↓	SSIM_test↑
w/o Serpentine Layout	2.03×10^{-4}	0.964	2.41×10^{-4}	0.959
Full	1.65×10^{-4}	0.971	1.96×10^{-4}	0.963

User Study

Part 1: Text-to-Vector Graphics (Generation)



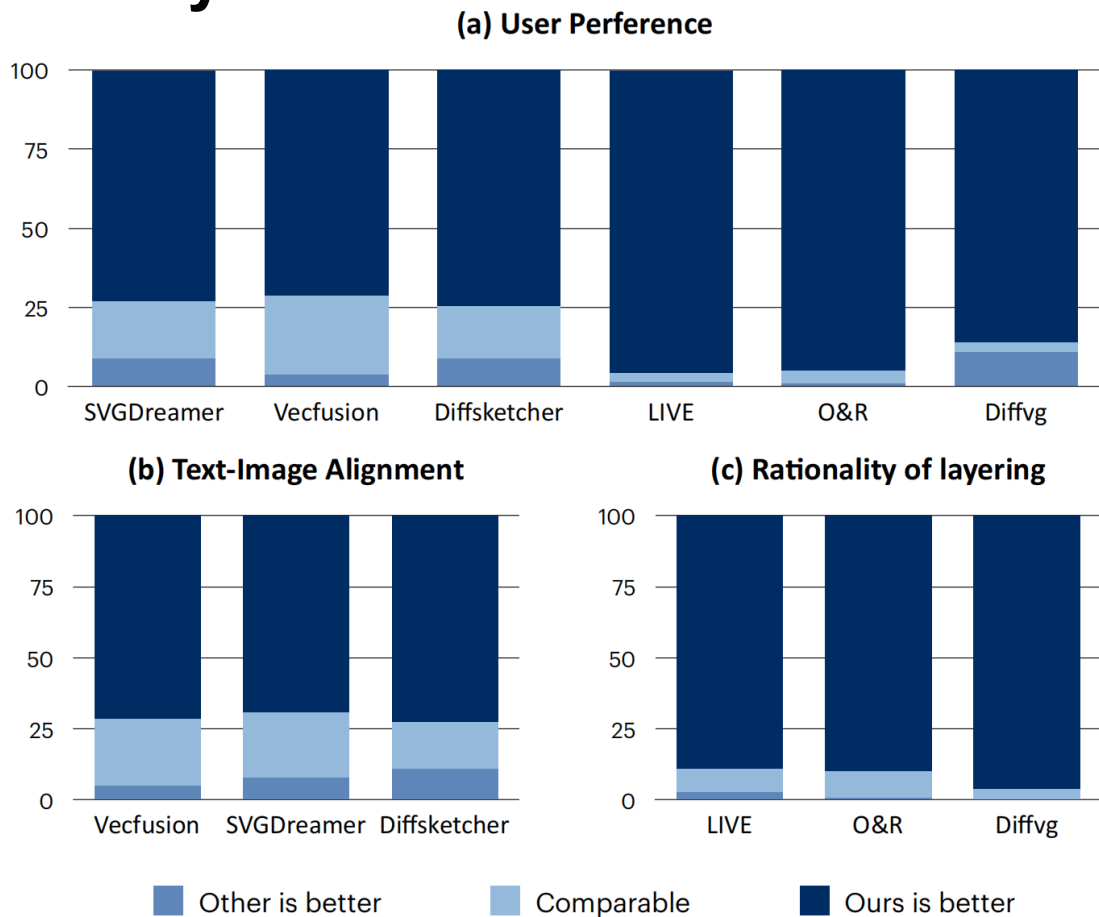
Text: an icon describes a soft gold wristwatch with a thin black outline

Part 2: Image-to-Vector Graphics (Vectorization)



- Which vector graphics result **looks better**?
- Which result **better reflects the meaning** of the original text?
- Which vectorization result **looks better**?
- Which result **better preserves the characteristics** of the original raster image

User Study



- Which vector graphics result **looks better**?
- Which result **better reflects the meaning** of the original text?
- Which vectorization result **looks better**?
- Which result **better preserves the characteristics** of the original raster image

Failure Cases



Unwanted repetition

... A freshly cut dragon fruit with vibrant pink skin, green leaves ...



Unwanted repetition

... A cheerful, cartoon-style toaster in orange and white, with golden toast popping out ...

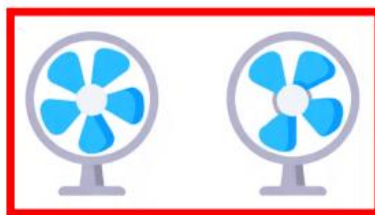


Incorrect layering

... A refreshing glass of orange drink with ice cubes, an orange slice garnish ...

SVG Results

Failure Cases



Inconsistencies

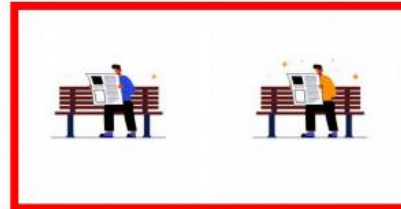
... An electric fan with blue blades ...



Incorrect layering

SVG Results

... A person in a green sweater is intently using a smartphone while seated inside a vehicle ...



Inconsistencies

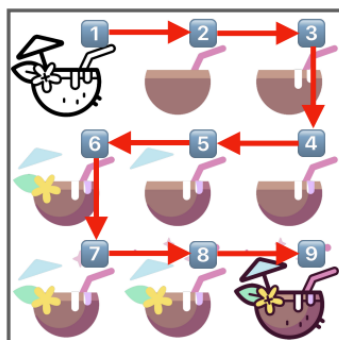


SVG Results

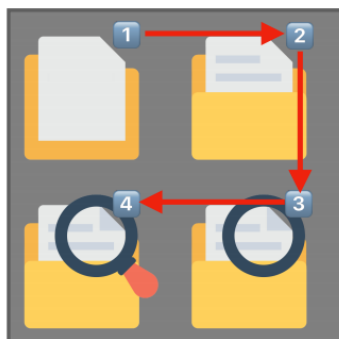
... A person sits on a bench, engrossed in reading a newspaper ...

- Shift from final SVG to creation **process supervision**
- Built 20K creation sequences + Serpentine construction
- Unifies text-to-SVG and image-conditioned layer prediction **in one DiT pipeline**
- Produces clean, editable SVGs, **better match professional design workflows**

Conclusion

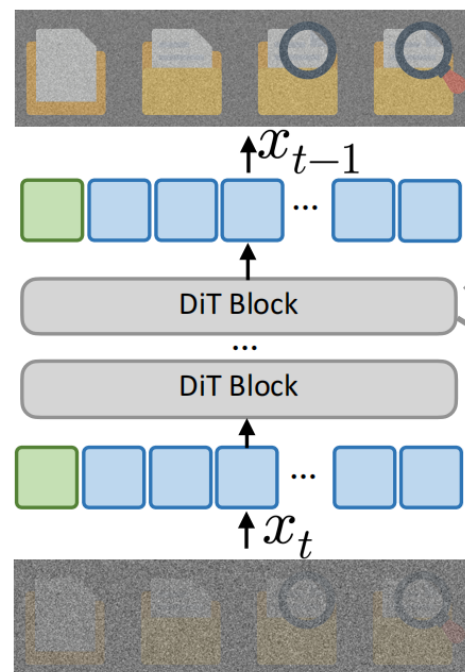


“This image illustrates the nine-step creation process of an SVG icon illustration, arranged in a 3*3 grid. The top left tile is a line art sketch of a icon, depicting its basic shape and structure. Subsequent tiles gradually add color and detail to the illustration...”



“The image is a 2x2 grid of folder icons. The top-left shows a folder with a plain document, while the top-right depicts a folder containing a document with visible text lines...”

Serpentin Dataset Construction



Layer-wise Model Training

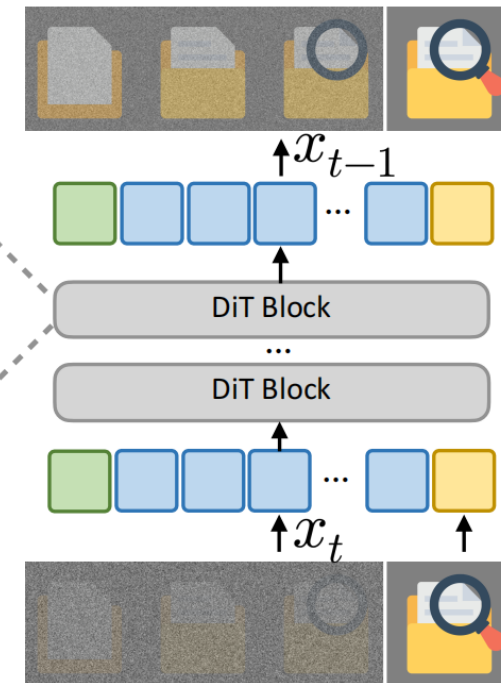
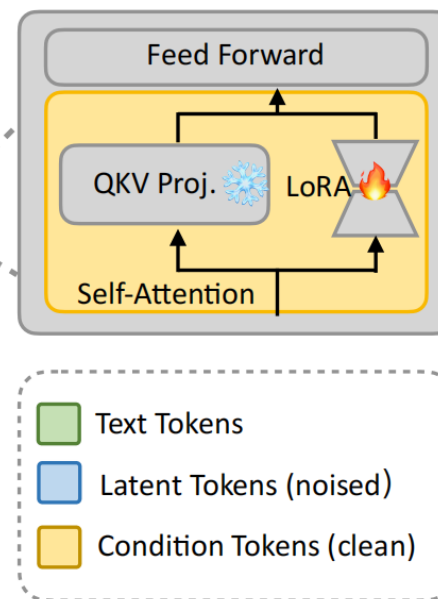


Image2Layers Model Training

Thanks for listening!

Presenter: Jiangyue Zeng
2025.12.14