Week 2

Matthew Perlman

Project Overview

Input: **Text Prompts** and **Reference Images**

Output: **SVG Code** for the desired image



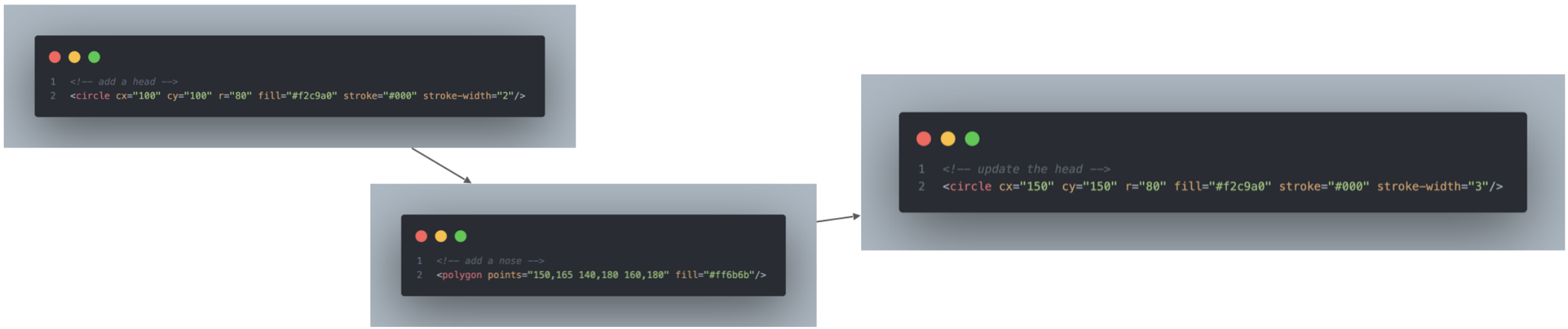
\+ “A picture of a cat” ⇒

```svg
<svg width="300" height="300" viewBox="0 0 300 300" xmlns="http://www.w3.org/2000/svg">
  <!-- head -->
  <circle cx="150" cy="150" r="80" fill="#f2c9a0" stroke="#000" stroke-width="3"/>

  <!-- ears -->
  <polygon points="90,90 120,40 140,100" fill="#f2c9a0" stroke="#000" stroke-width="3"/>
  <polygon points="210,90 180,40 160,100" fill="#f2c9a0" stroke="#000" stroke-width="3"/>

  <!-- eyes -->
  <circle cx="120" cy="150" r="8" fill="#000"/>
  <circle cx="180" cy="150" r="8" fill="#000"/>

  <!-- nose -->
  <polygon points="150,165 140,180 160,180" fill="#ff6b6b"/>

  <!-- mouth -->
  <path d="M150 180 Q140 190 130 185" stroke="#000" stroke-width="2" fill="none"/>
  <path d="M150 180 Q160 190 170 185" stroke="#000" stroke-width="2" fill="none"/>

  <!-- whiskers -->
  <line x1="70" y1="170" x2="120" y2="175" stroke="#000" stroke-width="2"/>
  <line x1="70" y1="185" x2="120" y2="185" stroke="#000" stroke-width="2"/>
  <line x1="180" y1="175" x2="230" y2="170" stroke="#000" stroke-width="2"/>
  <line x1="180" y1="185" x2="230" y2="185" stroke="#000" stroke-width="2"/>
</svg>
```

Project Overview

Mechanism: **Structured Iterative Refinement** of the SVG using a small **VLM**

Why: **Smaller Model**, **No Fine-Tuning**, **General Purpose**, **Not Auto Regressive**



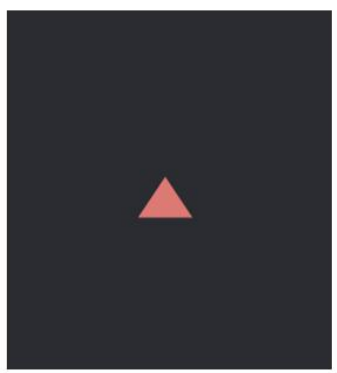
What I’ve Finished

- Setting up **Qwen7B VLM** and other utilities on the Newton Cluster
- **Initial Interface** for the model to deconstruct and visualize SVGs

```
1  <circle cx="150" cy="150" r="80" fill="#f2c9a0" stroke="#000" stroke-width="3"/>
```

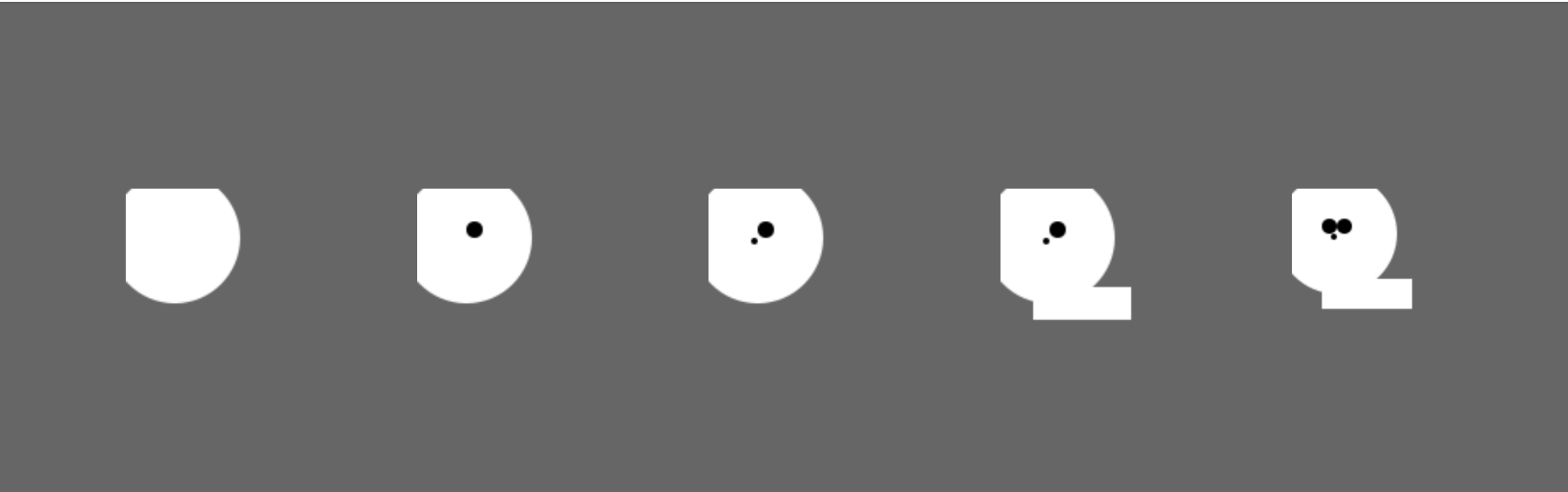


```
1  <polygon points="150,165 140,180 160,180" fill="#ff6b6b"/>
```



What I’ve Finished

- **Structured LLM-Generated SVG Actions** from text and image prompts
- Minimal **Refinement Loop**: Visualize → Reason → Action → Effect



What I’m Working On

- Optimizing the **Visualization Interface** and **System Prompting**
- Testing **User Prompt Coverage** and **Multiple Reference Images**
- Adding **New Actions** for the VLM to use during the refinement loop

What I’m Planning

- Comparison across **Different Models** (Qwen, LLaMA, Intern, LLaVA)
- Formalizing **Evaluation Metrics**
- **Ablating Inference-Time Hyperparameters** (prompts, interface, action set)
- Investigating input **Attention** and **Saliency** maps