

Evaluation Agent: Efficient and Promptable Evaluation Framework for Visual Generative Models (Oral)



Fan Zhang^{1*}



Shulin Tian^{2*}



Ziqi Huang^{2*‡}



Yu Qiao^{1†}



Ziwei Liu^{2†}

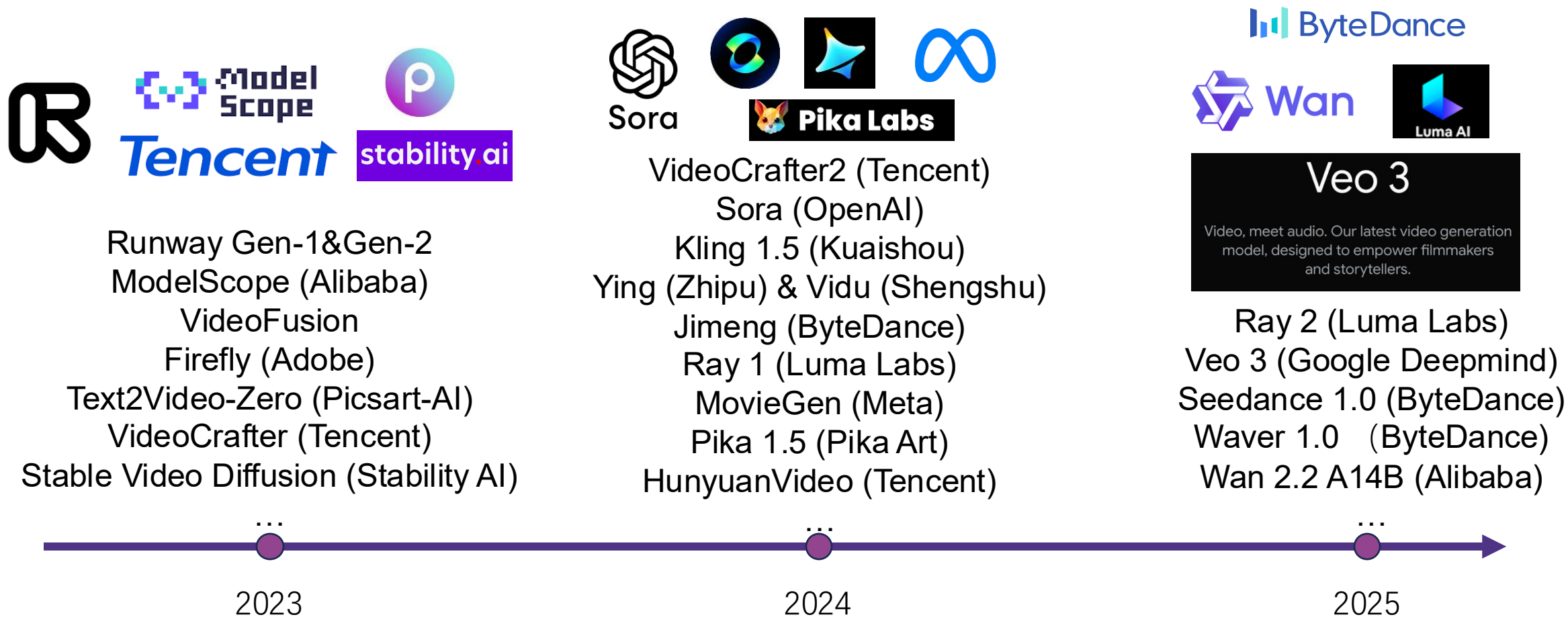
(* equal contributions, † corresponding authors, ‡ project lead)

¹ Shanghai Artificial Intelligence Laboratory ² S-Lab, Nanyang Technological University



| Background – GenAI

As an example of how video generative models have progressed...



| Problems

- Fixed (public) benchmarks are easy to hack
- Dynamic evaluation can better probe the capability of the model

Vibe checking -> Trustworthy evaluation

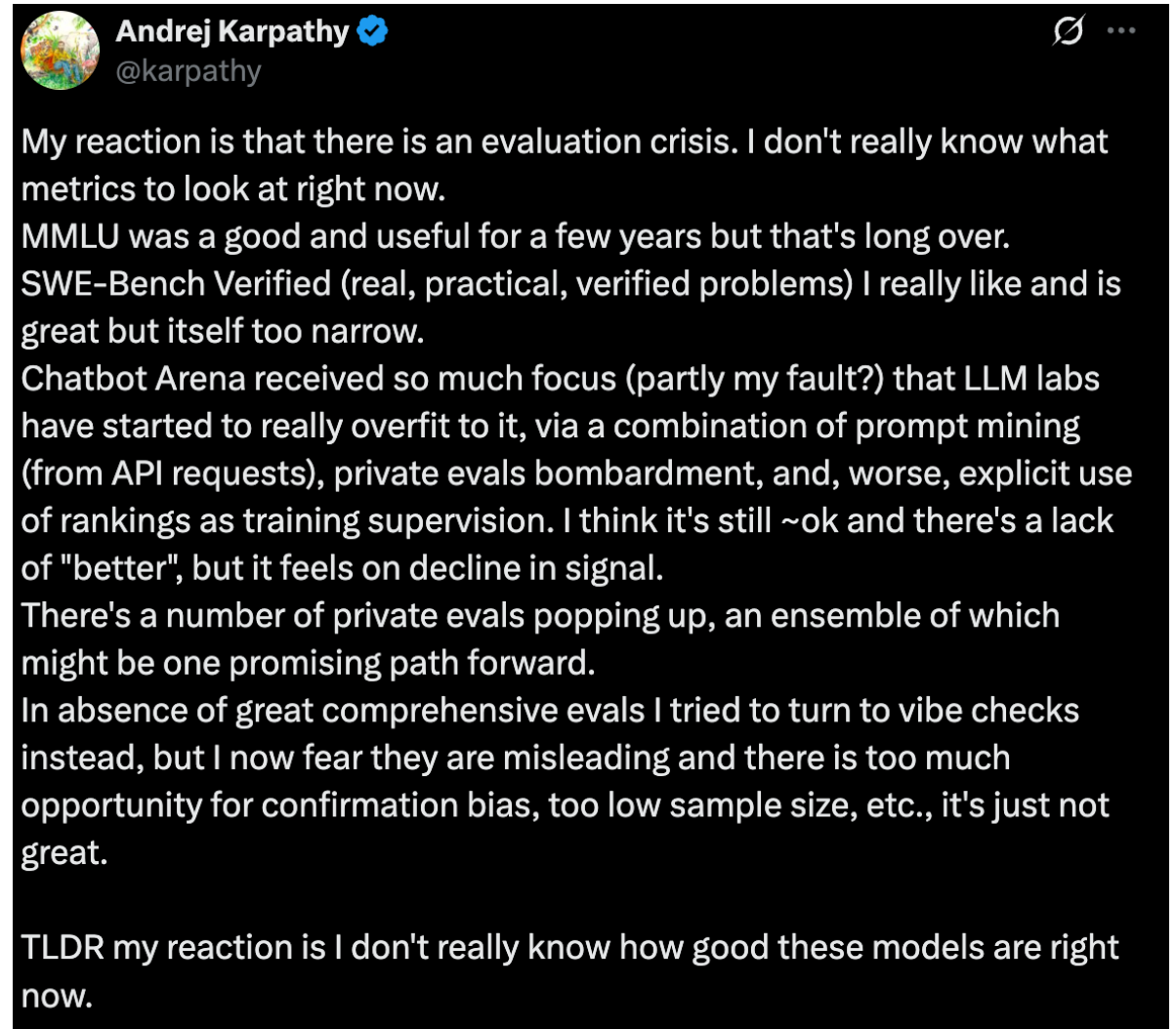


Fig. 1. A Twitter post from Andrej Karpathy

| Background – Agent

Definition

- Agent Framework (POMDP):

$$\mathcal{A}_{\text{LLM}} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, M, \pi_{\theta}, G \rangle$$

- Policy Function: mapping the context to the distribution over actions

$$a_t \sim \pi_{\theta}(\cdot \mid m_t, G)$$

- Memory / State: stores interaction histories till t

$$m_t = M(m_{t-1}, o_t, a_t),$$

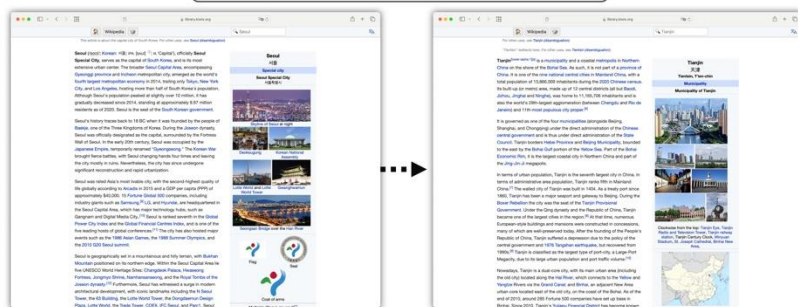
- Action Space $\mathcal{A}$ & Observation Space $\mathcal{O}$, $o_t = \Omega(s_t)$

$$\begin{aligned} \underbrace{m_t}_{\text{memory}} &= M(m_{t-1}, o_t, a_{t-1}) \\ \underbrace{a_t}_{\text{action}} &\sim \pi_{\theta}(a \mid m_t, G) \\ \underbrace{s_{t+1}}_{\text{state}} &= T(s_t, a_t) \\ \underbrace{o_{t+1}}_{\text{observation}} &= \Omega(s_{t+1}) \end{aligned}$$

| Example – Agent

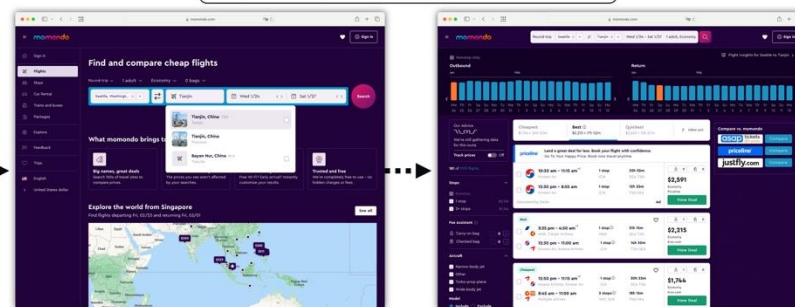
Which city features a Ferris wheel at its center, Seoul or Tianjin? Please assist me in booking a flight to the city with the Ferris wheel, followed by arranging hotel accommodation there. Additionally, could you find some YouTube videos showcasing this city?

Hop 1 <https://www.wikipedia.org/>



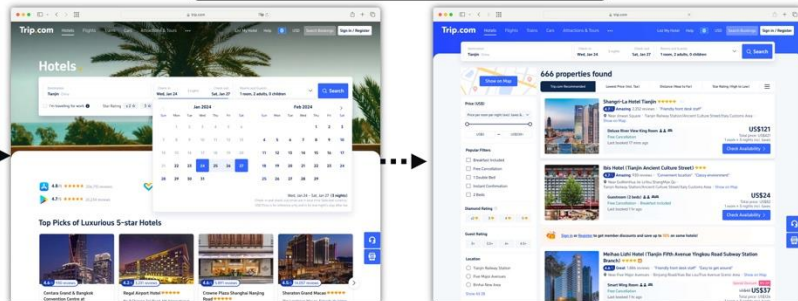
Information seeking Search for Seoul and **Tianjin** ✓ on Wikipedia

Hop 2 <https://www.momondo.com/>



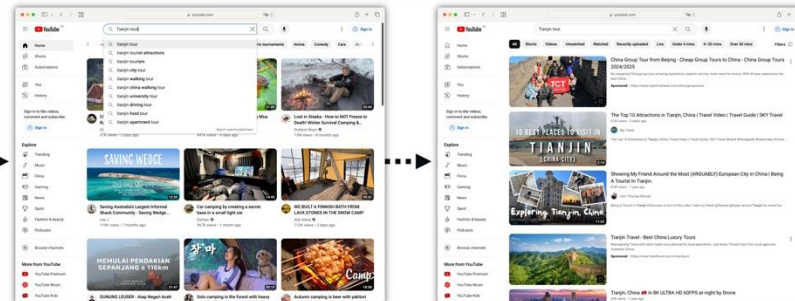
Action Book round trip flights to Tianjin

Hop 3 <https://www.trip.com/>



Action Arrange hotel accommodation in Tianjin

Hop 4 <https://www.youtube.com/>



Action Search for YouTube videos featuring tours of Tianjin

Fig. 2. A sample workflow of a multimodal web agent.

I Background – Agent

Why do we need agents?

LLM / VLM

- Model-level
- Stateless
- Lack of action interface

Agents

- System-level
- Long-term memory
- Stay within an environment
 - e.g. Web agent, embodied agent



Automation

Example – Agent

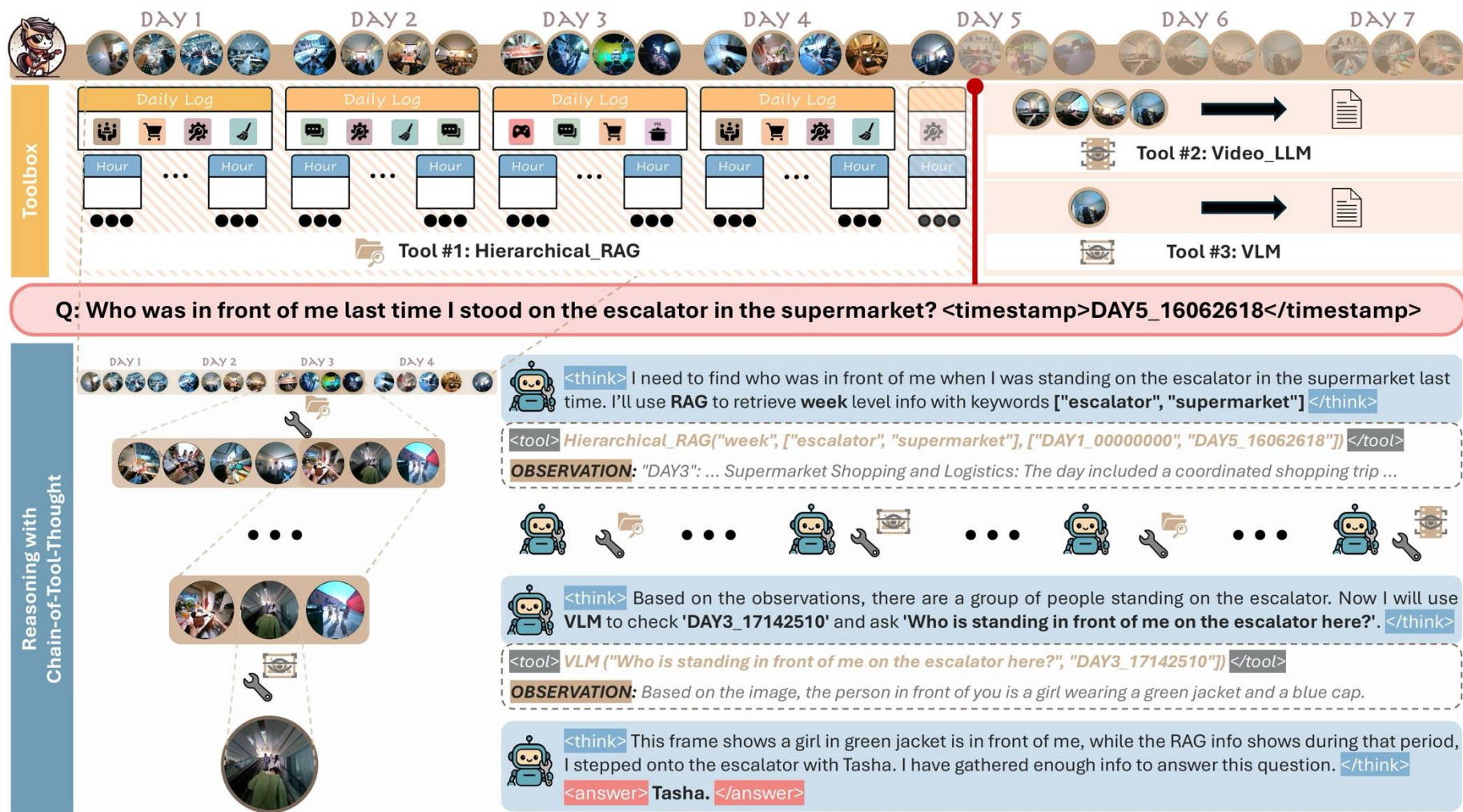
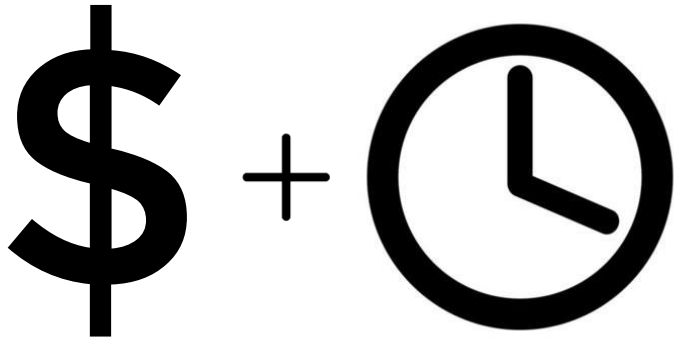


Fig. 3. A sample workflow of a multimodal web agent.

| Problems



Expensive

Computation & Time

- VBench - 4730 samples
- T2I-CompBench – 18000 samples



Fixed

Metrics & Pipelines

- Not flexible to user needs

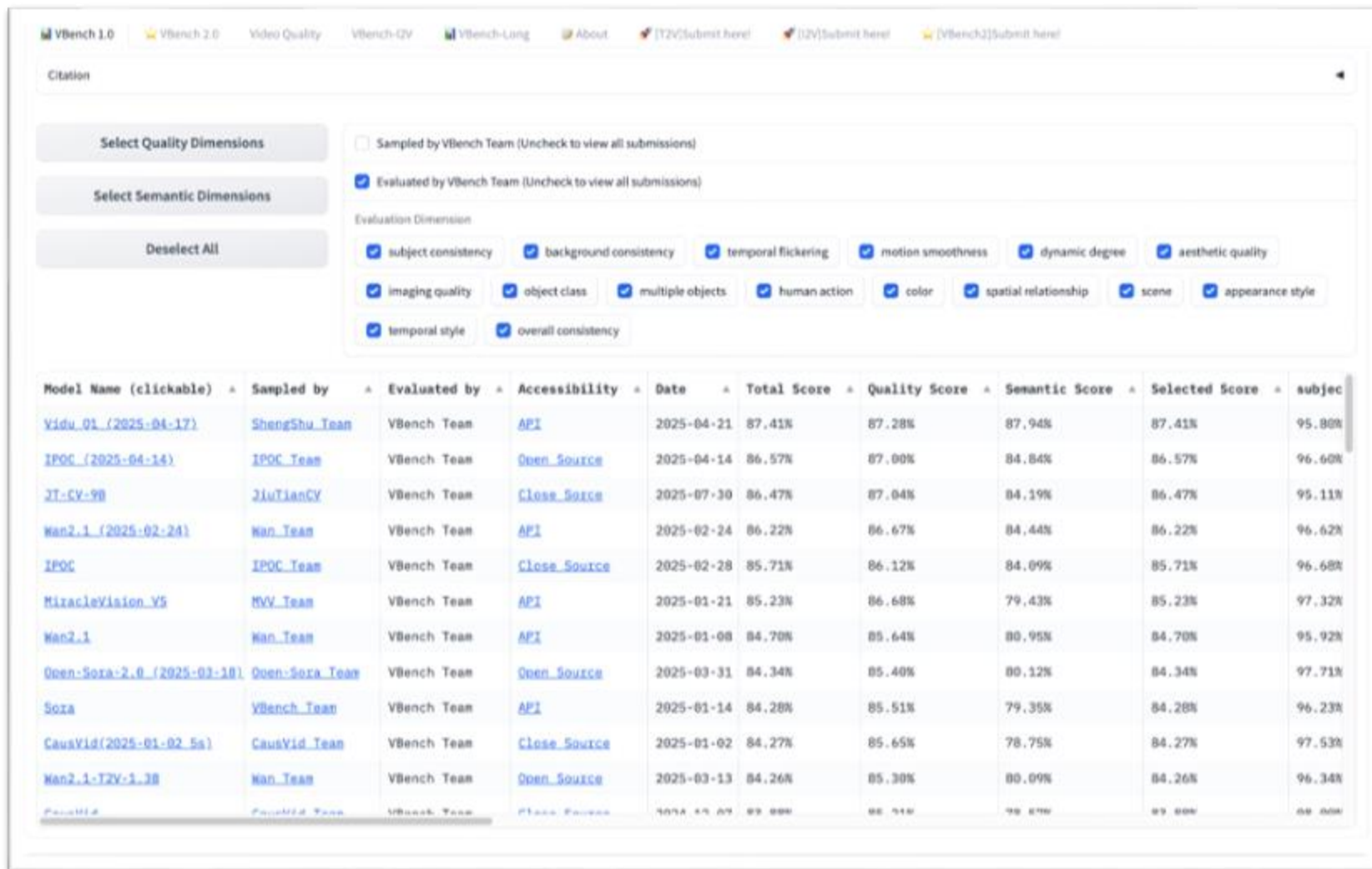


Blackbox

Process & Results

- Single / multiple numerical values
- No explainability

| Problems



The screenshot displays the VBench live results interface. At the top, there are navigation links for VBench 1.0, VBench 2.0, Video Quality, VBench-G2V, VBench-Long, About, and submission links for T2V, I2V, and V2V. Below these are buttons for 'Select Quality Dimensions', 'Select Semantic Dimensions', and 'Deselect All'. A section for 'Evaluation Dimension' includes checkboxes for various metrics: subject consistency, background consistency, temporal flickering, motion smoothness, dynamic degree, aesthetic quality, imaging quality, object class, multiple objects, human action, color, spatial relationship, scene, appearance style, temporal style, and overall consistency. The main table lists model names, who sampled them, who evaluated them, accessibility, date, total score, quality score, semantic score, selected score, and subject score.

Model Name (clickable)	Sampled by	Evaluated by	Accessibility	Date	Total Score	Quality Score	Semantic Score	Selected Score	subject
Vidu_01_(2025-04-17)	ShengShu_Team	VBench Team	API	2025-04-21	87.41%	87.28%	87.94%	87.41%	95.80%
IPOC_(2025-04-14)	IPOC_Team	VBench Team	Open_Source	2025-04-14	86.57%	87.00%	84.84%	86.57%	96.60%
JT-CV-90	JiuTianCV	VBench Team	Close_Source	2025-07-30	86.47%	87.04%	84.19%	86.47%	95.11%
Wan2.1_(2025-02-24)	Wan_Team	VBench Team	API	2025-02-24	86.22%	86.67%	84.44%	86.22%	96.62%
IPOC	IPOC_Team	VBench Team	Close_Source	2025-02-28	85.71%	86.12%	84.09%	85.71%	96.68%
MiracleVision_V5	MVV_Team	VBench Team	API	2025-01-21	85.23%	86.68%	79.43%	85.23%	97.32%
Wan2.1	Wan_Team	VBench Team	API	2025-01-08	84.70%	85.64%	80.95%	84.70%	95.92%
Open-Sora-2.0_(2025-03-18)	Open-Sora_Team	VBench Team	Open_Source	2025-03-31	84.34%	85.40%	80.12%	84.34%	97.71%
Sora	VBench_Team	VBench Team	API	2025-01-14	84.28%	85.51%	79.35%	84.28%	96.23%
CausVid(2025-01-02_5s)	CausVid_Team	VBench Team	Close_Source	2025-01-02	84.27%	85.65%	78.75%	84.27%	97.53%
Wan2.1-T2V-1.3B	Wan_Team	VBench Team	Open_Source	2025-03-13	84.26%	85.30%	80.09%	84.26%	96.34%
Pexels	Pexels_Team	VBench Team	Open_Source	2025-03-03	83.88%	85.71%	79.87%	83.88%	96.00%

Fig. 4. A screenshot of VBench live results.

Evaluation is all you need.

| Motivation

We want the **Evaluation Agent** framework, which enables *efficient*, *dynamic*, and *promptable* evaluations while providing detailed, user-tailored analyses.

Overview

Task Definition

Inputs:

- User queries (w/ targeted Gen model)

Outputs:

- Evaluation results with explanation in natural language format

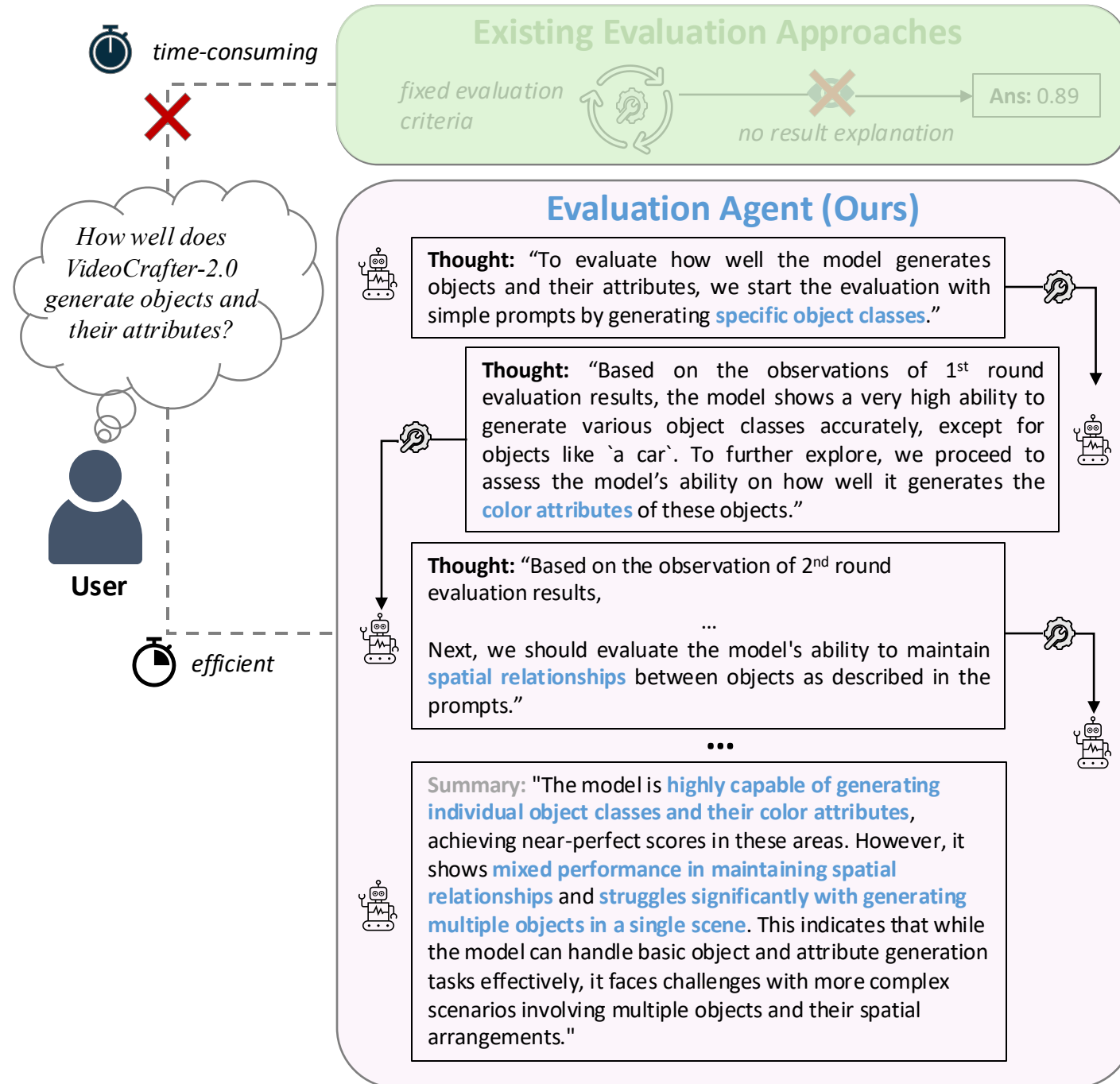


Fig.5 A Use Case of Evaluation Agent.

| Pipeline

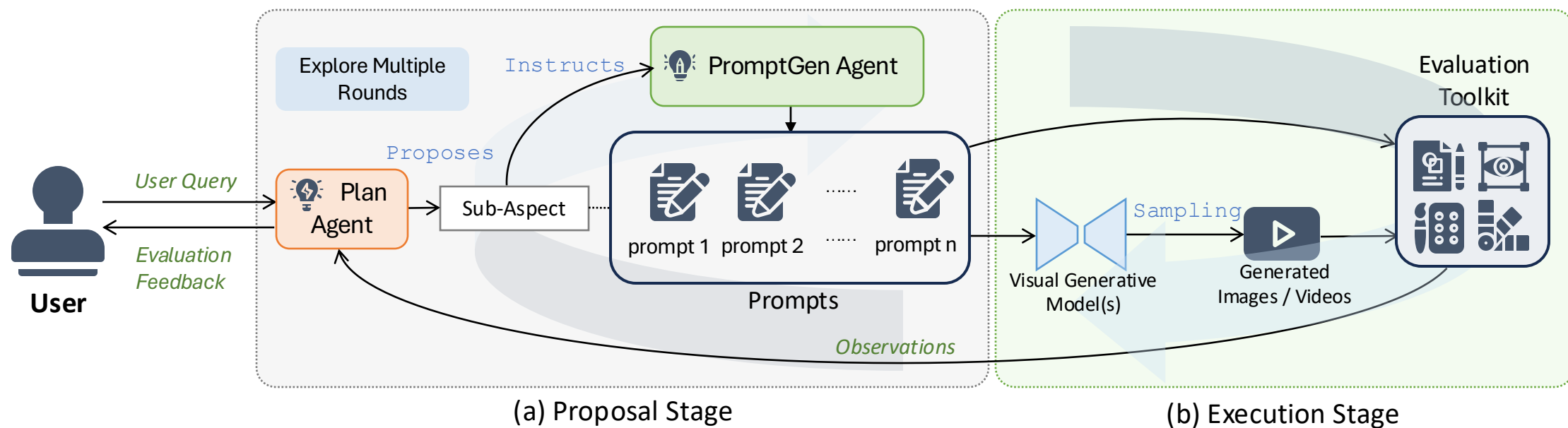


Fig.6 The pipeline of the proposed Evaluation Agent Framework. Our Evaluation Agent can handle free-form user query, and requires sampling only a few visuals to provide a reliable summary with explanations.

| Pipeline

(I) Proposal Stage:

- User queries (w/ observations)
- Two agents
- Proposed evaluation prompts

(II) Execution Stage:

- Proposed evaluation prompts
 - Targeted Gen model(s)
 - Evaluation toolkit

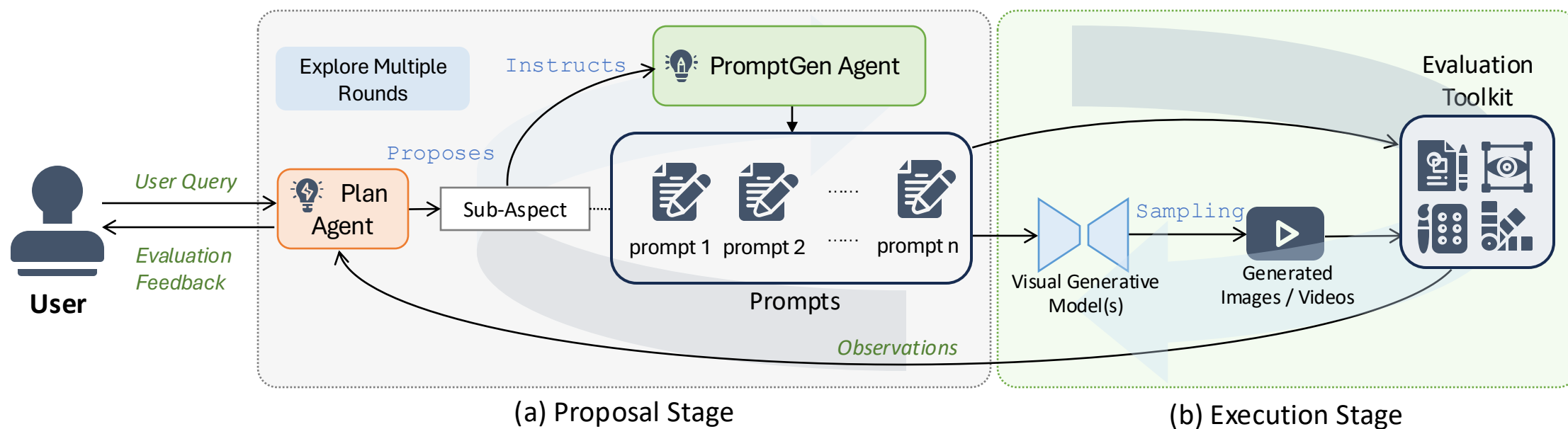


Fig.6 The pipeline of the proposed Evaluation Agent Framework. Our Evaluation Agent can handle free-form user query, and requires sampling only a few visuals to provide a reliable summary with explanations.

| Efficiency Comparison

Comparison of Video Generation Model Evaluation with VBench

Models	VBench (Total Cost) ↓	VBench (Avg. Cost per Dimension) ↓	Evaluation Agent (Ours) ↓
Latte-1 (Ma et al., 2024)	2557 min, 4355 samples	170 min, 290 samples	15 min, 25 samples
ModelScope (Wang et al., 2023)	1160 min, 4355 samples	77 min, 290 samples	6 min, 23 samples
VideoCrafter-0.9 (He et al., 2022)	1459 min, 4355 samples	97 min, 290 samples	9 min, 24 samples
VideoCrafter-2 (Chen et al., 2024a)	4261 min, 4355 samples	284 min, 290 samples	24 min, 23 samples

Comparison of Image Generation Model Evaluation with T2I-CompBench

Models	T2I-Comp (Total Cost) ↓	T2I-Comp (Avg. Cost per Dimension) ↓	Evaluation Agent (Ours) ↓
SD1.4 (Rombach et al., 2022)	563 min, 12000 samples	141 min, 3000 samples	5 min, 26 samples
SD2.1 (Rombach et al., 2022)	782 min, 12000 samples	196 min, 3000 samples	6 min, 26 samples
SDXL (Podell et al., 2023)	1543 min, 12000 samples	386 min, 3000 samples	8 min, 26 samples
SD3.0 (Esser et al., 2024)	1410 min, 12000 samples	353 min, 3000 samples	7 min, 25 samples

| Effectiveness Comparison

Experiment Results - T2V

Models	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
Latte-1 (Ma et al., 2024)	50%/80%	0%/30%	40%/70%	30%/70%	60%/100%	70%/100%	40%/50%
ModelScope (Wang et al., 2023)	80%/80%	80%/90%	60%/80%	60%/100%	60%/100%	100%/100%	0%/50%
VideoCrafter-0.9 (He et al., 2022)	100%/100%	80%/100%	70%/100%	80%/100%	90%/100%	20%/100%	20%/60%
VideoCrafter-2 (Chen et al., 2024a)	10%/100%	60%/100%	30%/90%	30%/80%	80%/100%	50%/100%	70%/100%
Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
40%/100%	10%/10%	30%/70%	10%/80%	20%/40%	70%/90%	40%/100%	70%/100%
50%/100%	10%/40%	0%/20%	10%/30%	20%/100%	90%/100%	50%/90%	20%/100%
80%/100%	10%/30%	10%/40%	20%/100%	30%/100%	60%/100%	80%/100%	0%/80%
20%/60%	10%/90%	90%/100%	0%/70%	0%/10%	80%/100%	80%/100%	60%/100%

Tab.1 Experiment results of baseline models on T2V

Experiment Results - T2I

Models	Color Binding	Shape Binding	Texture Binding	Non-Spatial Relationships
SD-1.4 (Rombach et al., 2022)	50%/100%	100%/100%	0%/100%	50%/100%
SD-2.1 (Rombach et al., 2022)	100%/100%	60%/100%	80%/100%	60%/100%
SDXL (Podell et al., 2023)	100%/100%	20%/100%	80%/100%	60%/100%
SD-3.0 (Esser et al., 2024)	20%/90%	0%/90%	0%/70%	80%/90%

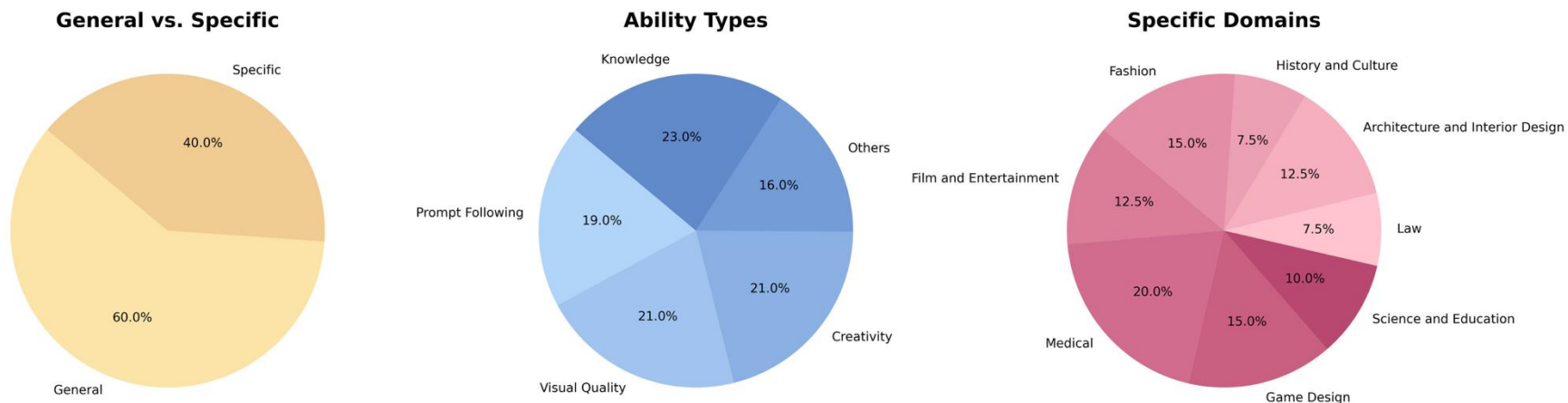
Tab.2 Experiment results of baseline models on T2I

T2V Toolkit:
VBench (Huang et al., 2024)

T2I Toolkit:
T2I-CompBench (Huang et al., 2023)

| Open-Ended Query Evaluation

- The Need for Customized Evaluation
 - A user study revealed that users prioritize models that address their specific needs over general performance, such as:
 - Can the model generate variations of existing artwork while maintaining the origin style
 - Does the model understand and apply photography concepts such as focal length, aperture, and ISO?
- We built an Open-Ended Query Evaluation Dataset by collecting, cleaning, filtering, expanding and tagging user's specific needs from a user study



| Open-Ended Query Evaluation



User

Can the model generate variations of existing artwork while maintaining the original style?

Sub-aspect: *"Replicating Basic Art Styles"*

Thought: To begin assessing the model's capabilities, I will first explore its ability to generate artwork that replicates foundational or well-known art styles, such as Impressionism, Cubism, or Surrealism. This will provide a baseline understanding of how well the model can maintain stylistic elements when generating new artwork.

Evaluation Process

sampled images



Question A: "Does the artwork reflect the Impressionism style with emphasis on light and color?"

Question B: "Are the brushstrokes and overall texture similar to Monet's work?"

Answer: "Yes. The artwork reflects the Impressionism style with its emphasis on light and color. The depiction of the landscape, with bright greens, blues, and yellows, along with the visible brush strokes, creates an impression of a moment in time, typical of Impressionism."

Answer: "Yes. The brushstrokes and overall texture in this painting closely resemble the style typical of Claude Monet's work. Monet, a leading figure in the Impressionist movement, was known for his short, visible brushstrokes that captured the changing qualities of light. In this image, the brushstrokes are loose and textured, creating a sense of movement and capturing the light in a way that is reminiscent of Monet's landscapes."

Prompt: Create a new landscape painting in the style of Impressionism, focusing on light and color play typical of Monet's work.





Sub-aspect: *"Style Consistency in Detail-Oriented Artworks"*

Thought: Having established that the model can replicate well-known art styles effectively, the next step is to explore if it can maintain style consistency in more detail-oriented artworks. This will involve testing the model's ability to introduce variations while preserving intricate stylistic details, which is crucial for generating authentic variations of existing artworks.

Evaluation Process

sampled images



Question A: "Does the image include a city skyline?"

Question B: "Are the brush strokes and color palette consistent with van Gogh's style?"

Answer: "Yes. The image includes a city skyline. It features a series of tall buildings with various architectural designs set against a stylized sky, reminiscent of Vincent van Gogh's style."

Answer: "Yes. The brush strokes and color palette in the image are consistent with Vincent van Gogh's style. The swirling, dynamic brush strokes are reminiscent of his famous works such as "Starry Night." The bold, vibrant colors, especially the blues and yellows, are signature elements of van Gogh's approach to painting."

Prompt: Create a variation of Vincent van Gogh's 'Starry Night' that includes a city skyline while maintaining the original style.





Sub-aspect: *"Blending Styles: Introducing New Elements While Preserving Original Aesthetic"*

Thought: The model demonstrates a strong ability to replicate foundational styles and introduce variations while preserving stylistic elements. To push the boundaries further, I will evaluate the model's ability to blend styles by incorporating elements from different periods or cultures into existing artworks while maintaining the core aesthetic. This will test the model's flexibility and creativity in maintaining stylistic integrity when merging diverse artistic elements.

Evaluation Process

sampled images



Question A: "Does the image preserve the interplay of light and color typical of Monet's 'Water Lilies'?"

Question B: "How are indigenous Australian dot painting techniques reflected in the image?"

Prompt: Integrate elements of indigenous Australian dot painting into the Impressionist style of Claude Monet's 'Water Lilies', while keeping the focus on the interplay of light and color.

Answer: "No. The image does not preserve the interplay of light and color typical of Monet's 'Water Lilies'. Monet's work is known for its soft, impressionistic style, often featuring subtle shifts in light and color that create a dreamy, atmospheric effect. In contrast, this image uses bright, bold colors and has a more digital or stylized appearance, with distinct lines and defined edges. The light is vivid and direct, rather than the diffused, nuanced lighting typical of Monet's paintings."

Answer: "The image does not reflect traditional Indigenous Australian dot painting techniques. The painting primarily features impressionistic techniques to depict a vibrant water lily scene. It uses small dabs and strokes of color rather than the systematically arranged dots common in Indigenous Australian dot paintings."





Sub-aspect: *"Complex Style Integration: Merging Multiple Cultural Art Styles"*

Thought: The model has shown potential in maintaining stylistic elements while introducing new cultural motifs. The next step is to test its ability to integrate multiple cultural art styles in a single piece, maintaining the core elements of each style. This will assess the model's proficiency in complex style synthesis and its ability to create cohesive artworks that respect and reflect diverse artistic traditions.

Evaluation Process

sampled images





Question A: "Does this image maintain the key characteristics of Japanese ukiyo -e style?"

Question B: "Are traditional African tribal patterns clearly integrated into the artwork?"

Answer: "Yes. The image maintains key characteristics of the Japanese ukiyo-e style: 1. Bold Lines and Flat Colors. 2. Simplified Forms. 3. Decorative Patterns. 4. Natural Themes. 5. Focus on Beauty. These aspects collectively show a strong adherence to classic ukiyo -e characteristics."

Answer: "No. The artwork does not incorporate traditional African tribal patterns. Instead, it features elements typical of Japanese art, such as a focus on natural scenes with leaves and flowers, and detailed, colorful clothing reminiscent of kimonos. The style is more in line with traditional East Asian artistic influences."

Prompt: Create an artwork that combines Japanese ukiyo -e style with traditional African tribal patterns.





Evaluation Agent

Analysis: During the exploration process, the model showed proficiency in maintaining the core elements of individual art styles when tasked with generating variations. For instance, the model effectively replicated the Impressionist style by capturing the interplay of light and movement, and it maintained the distinct swirling skies and color palette of Van Gogh's 'Starry Night'. However, when the task involved blending multiple styles, such as combining ukiyo-e with African tribal motifs or Renaissance art with Aboriginal dot paintings, the model struggled to integrate both styles meaningfully. In several cases, it defaulted to predominantly representing only one style. For instance, in attempting to merge ukiyo-e with African tribal patterns, the model successfully maintained ukiyo-e characteristics but failed to incorporate African motifs. Similarly, when synthesizing Renaissance with Aboriginal dot painting, the resulting image lacked Renaissance characteristics and primarily displayed elements of dot painting. These results demonstrate the model's current limitation in capturing and blending the essence of multiple distinct styles, especially when they differ significantly in form, color, and thematic elements.

Summary: The model can effectively generate variations of existing artwork while maintaining the original style when working within a single artistic tradition. However, it encounters challenges when tasked with blending multiple diverse art styles into a cohesive piece. This limitation is evident in its inability to consistently integrate secondary styles, suggesting a boundary in its ability to synthesize complex, cross-cultural artistic elements. Thus, while the model is adept at style replication within a single tradition, its capability to merge and innovate across disparate styles remains constrained.

Advantages of Our **Evaluation Agent**

- ***Efficient***: Fast and dynamic evaluation.
- ***Flexible***: Supports open-ended queries
- ***Interpretable***: Clear and insightful results.
- ***Scalable***: Easily integrates new tools.

For Further Details, Check Below

Project Page



Code

