CVPR
JUNE 17-21, 2024
SEATTLE, WA

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

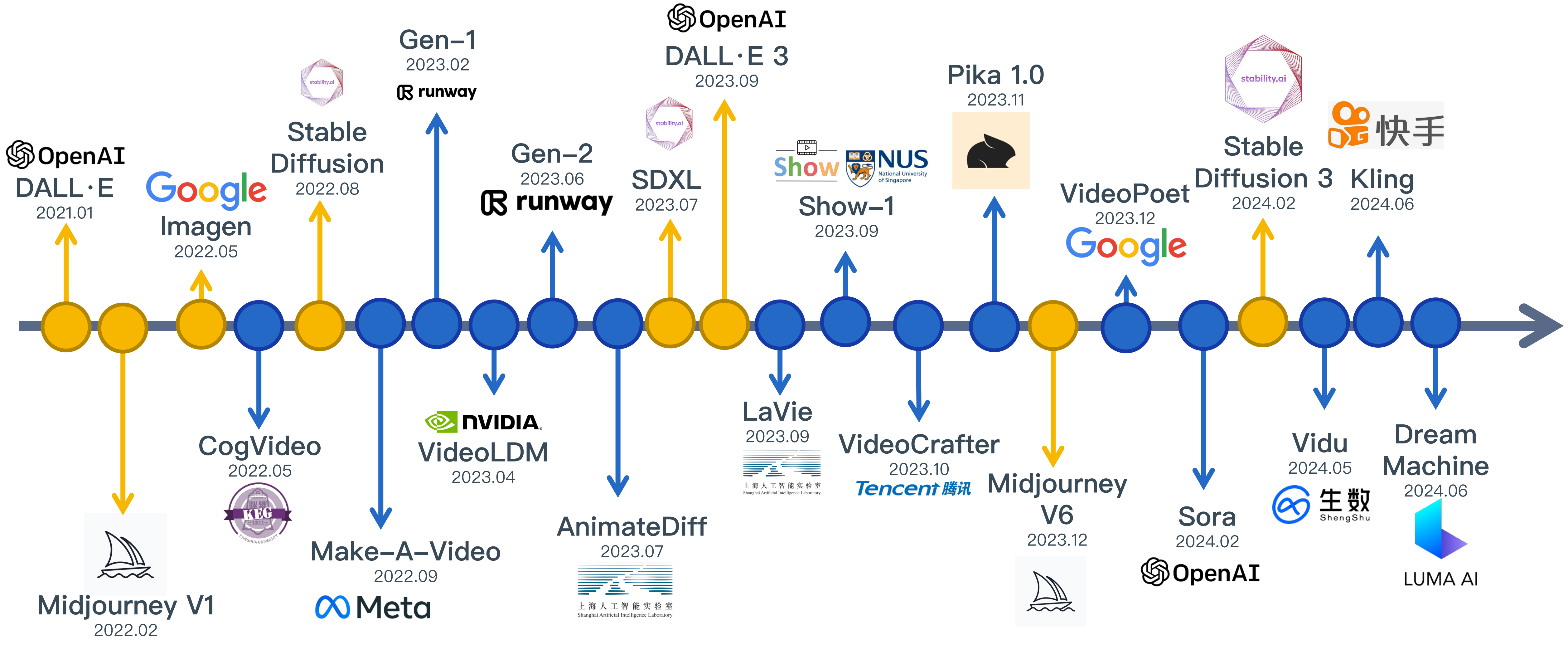
VBENCH

Comprehensive Benchmark Suite for Video Generative Models

Ziqi Huang

MMLab@NTU | S-Lab, Nanyang Technological University

Video generation is developing rapidly



OpenAI DALL·E 2021.01

Midjourney V1 2022.02

Google Imagen 2022.05

CogVideo 2022.05 KEG

stability.ai Stable Diffusion 2022.08

Make–A–Video 2022.09 Meta

Gen–1 2023.02 runway

NVIDIA VideoLDM 2023.04

Gen–2 2023.06 runway

AnimateDiff 2023.07 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

stability.ai SDXL 2023.07

OpenAI DALL·E 3 2023.09

LaVie 2023.09 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Show NUS National University of Singapore Show–1 2023.09

VideoCrafter 2023.10 Tencent 腾讯

Pika 1.0 2023.11

Midjourney V6 2023.12

VideoPoet 2023.12 Google

Sora 2024.02 OpenAI

stability.ai Stable Diffusion 3 2024.02

Vidu 2024.05 生数 ShengShu

快手 Kling 2024.06

Dream Machine 2024.06 LUMA AI

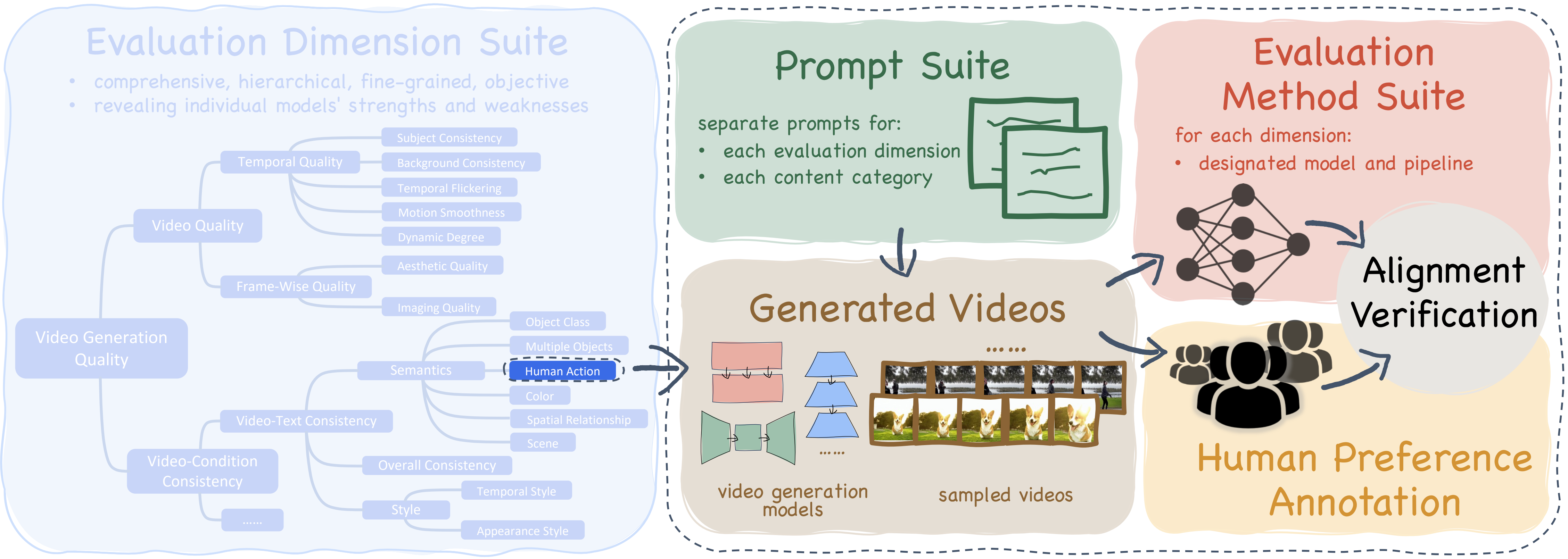
Why Need VBench?

- Video generation is developing rapidly.
- How to evaluate these models? What’s each v-gen model good/bad at?

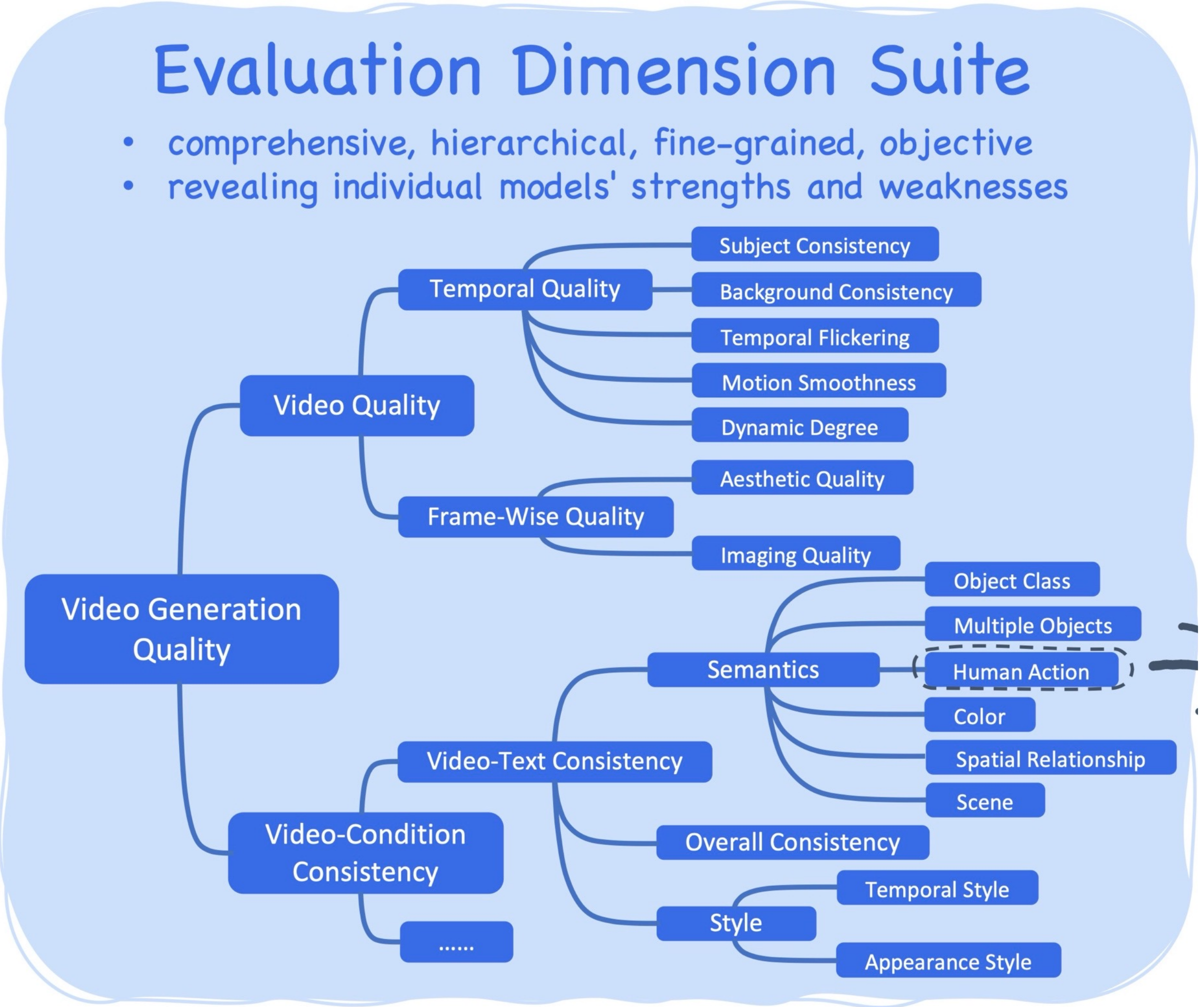
Existing Metrics	***What We Need***
a single number (FVD, CLIP) can’t reveal individual model’s strengths and weaknesses	*multiple dimensions for detailed insights*
not well-aligned with human (FVD)	*high alignment with human*
not catered for AIGC (e.g., Quality Assessment)	*focus on AIGC artifacts*

- We propose VBench to comprehensively benchmark and evaluate video generative models.

Overview of VBench

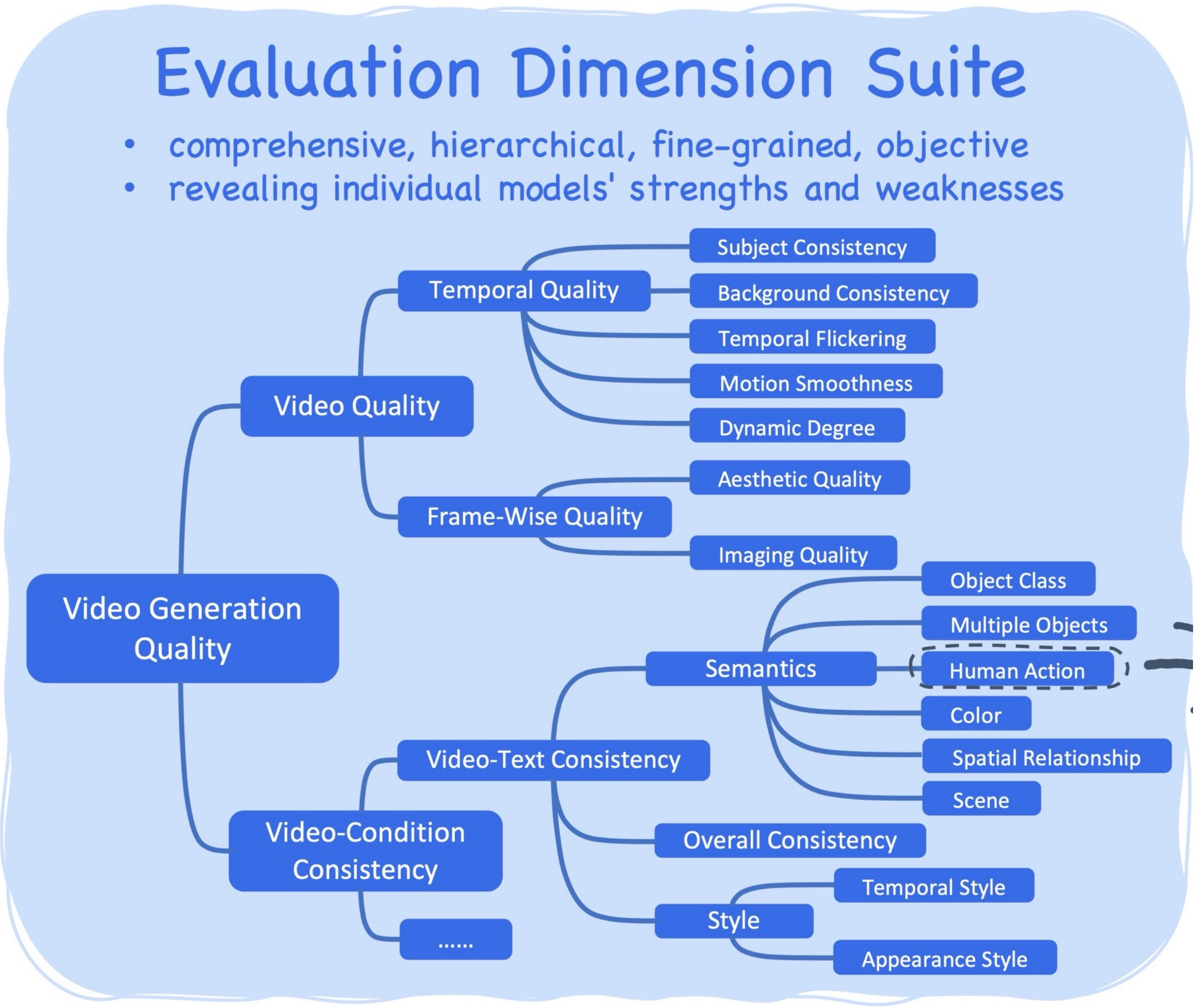


Dimension Suite



- 16 ability dimensions, hierarchical and disentangled
- each dimension assesses one aspect of video generation quality

Why Multiple Dimensions?



Evaluation Dimension Suite

- comprehensive, hierarchical, fine-grained, objective
- revealing individual models' strengths and weaknesses

Video Generation Quality

Video Quality

Temporal Quality

Subject Consistency

Background Consistency

Temporal Flickering

Motion Smoothness

Dynamic Degree

Frame-Wise Quality

Aesthetic Quality

Imaging Quality

Video-Condition Consistency

Video-Text Consistency

Semantics

Object Class

Multiple Objects

Human Action

Color

Spatial Relationship

Scene

Overall Consistency

Style

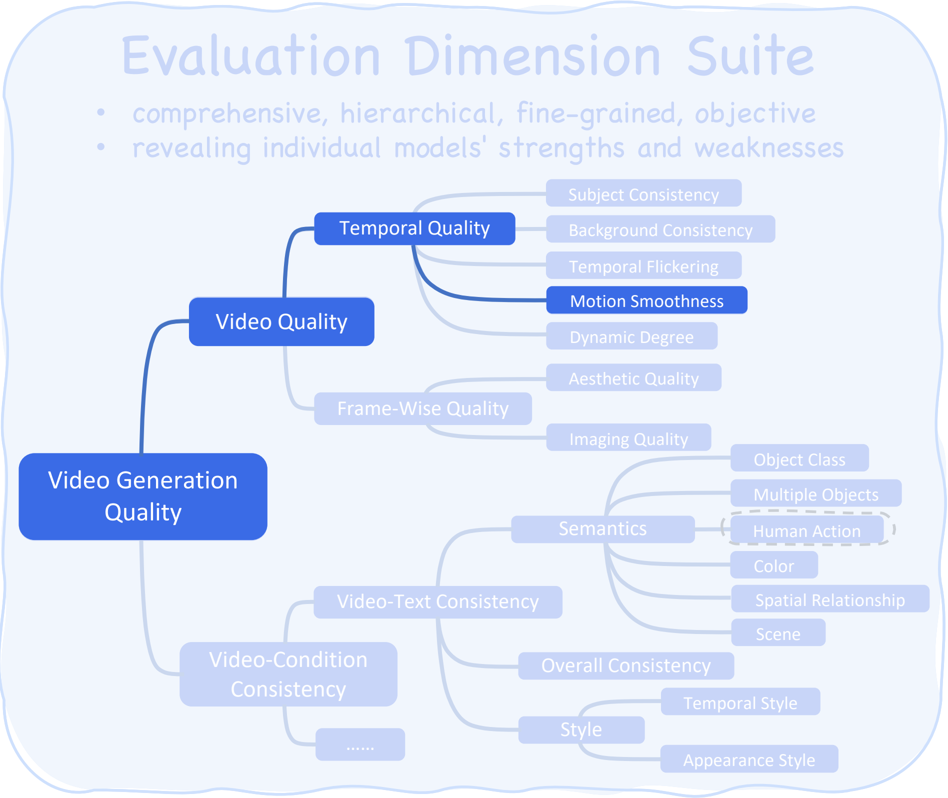
Temporal Style

Appearance Style

......

- reveal individual model's strengths and weaknesses
- different people prioritize each ability dimension differently

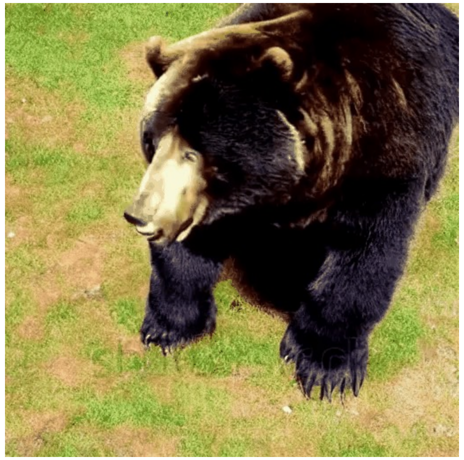
Evaluation Dimension: Motion Smoothness



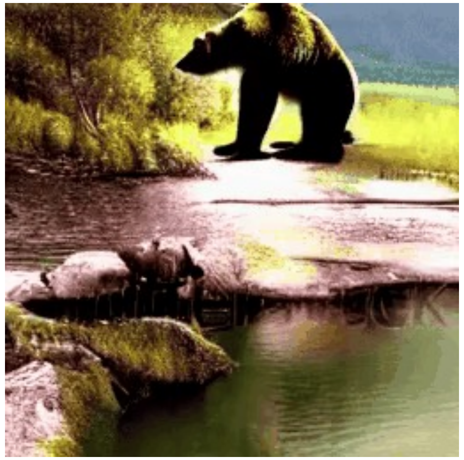
Evaluation Dimension Suite

- comprehensive, hierarchical, fine-grained, objective
- revealing individual models' strengths and weaknesses

score 96.04% *(better)*



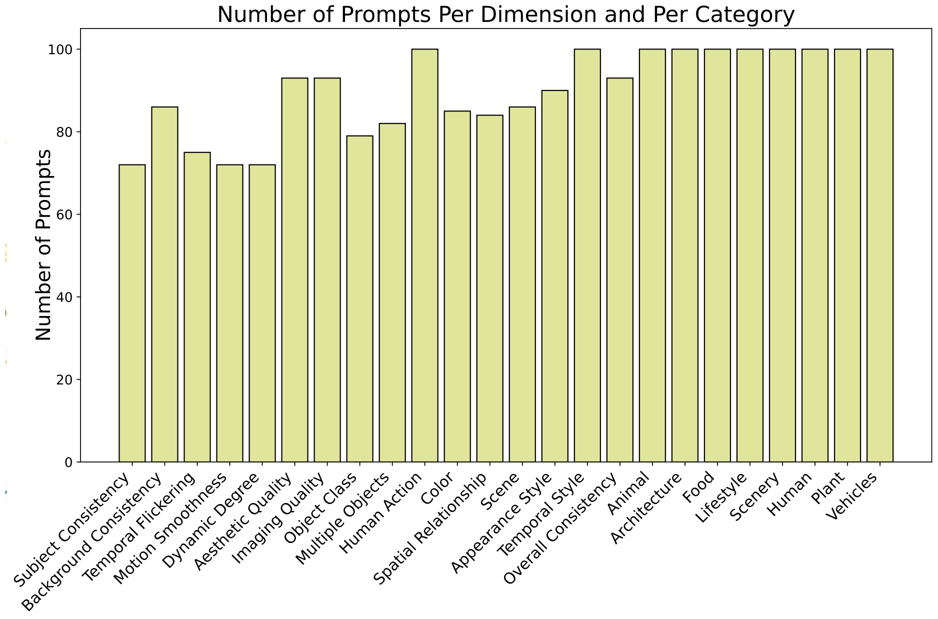
score 88.47%



whether the motion in the generated video is smooth

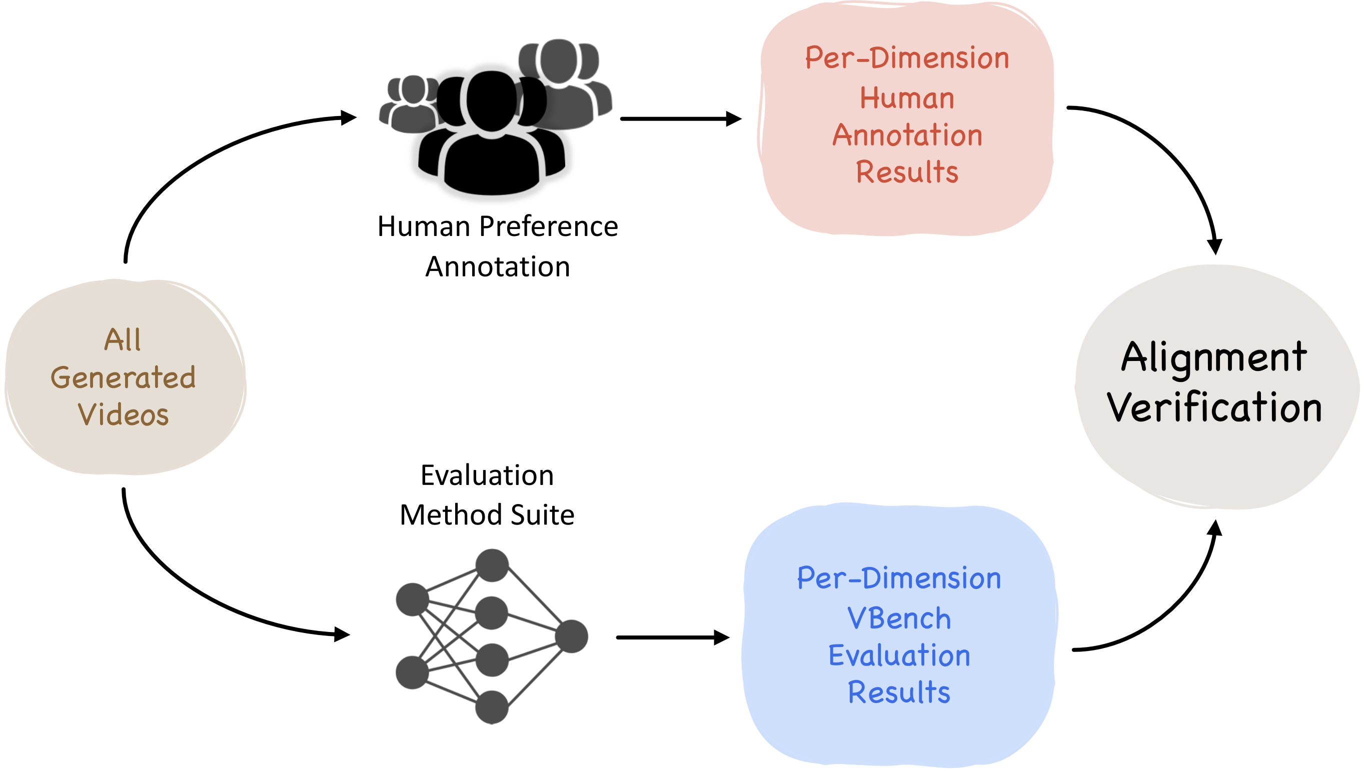
Prompt Suite

- diverse → comprehensive evaluation
- compact → efficient evaluation
- prompt suites for each dimension and each content category → multi-perspective insights
- per ability dimension: ~100 prompts
- per content category: ~100 prompts



Human Alignment

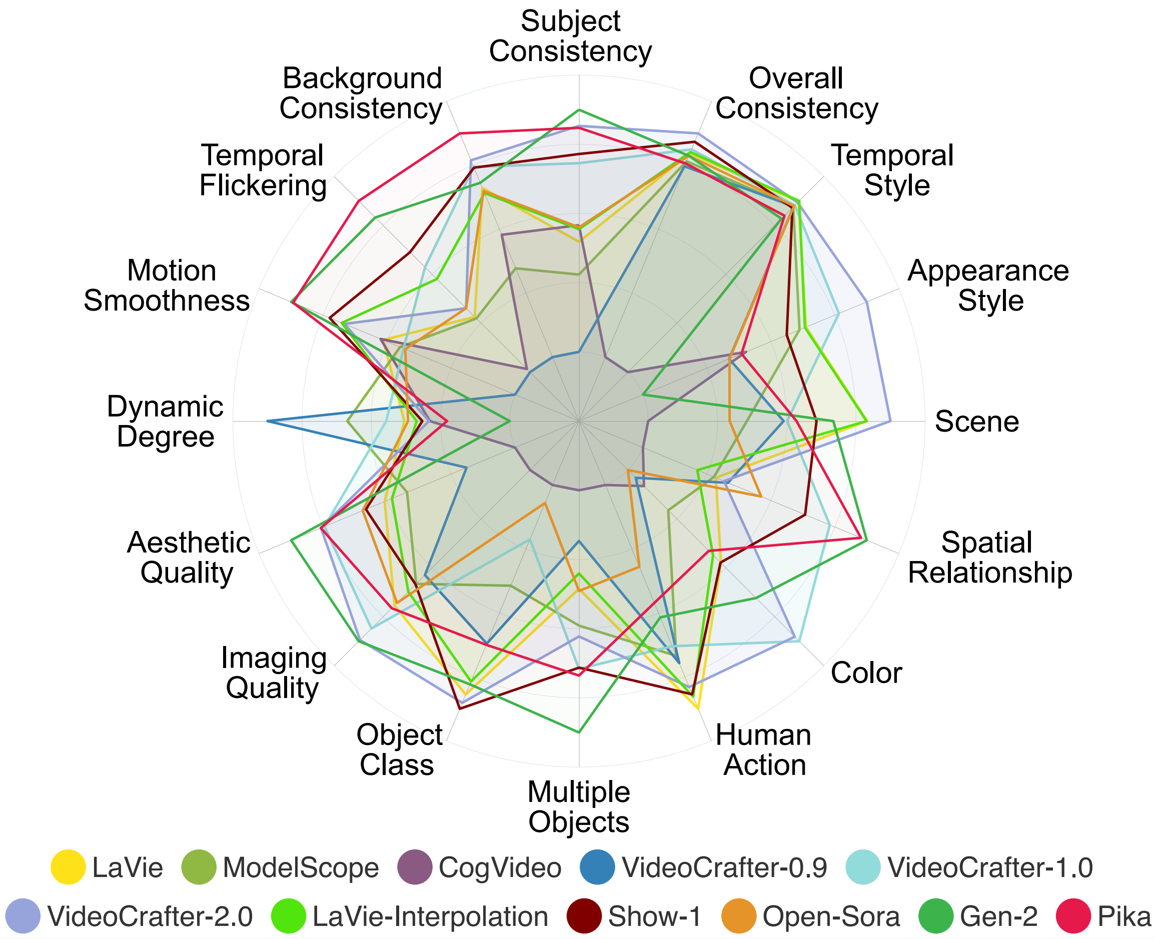
- VBench evaluation is well-aligned with human perception in each of the 16 dimensions.



All Generated Videos
Human Preference Annotation
Per-Dimension Human Annotation Results
Evaluation Method Suite
Per-Dimension VBench Evaluation Results
Alignment Verification

Evaluation Results

Video Generative Models



- trade-off across dimensions:
  - e.g., temporal consistency vs. dynamic degree

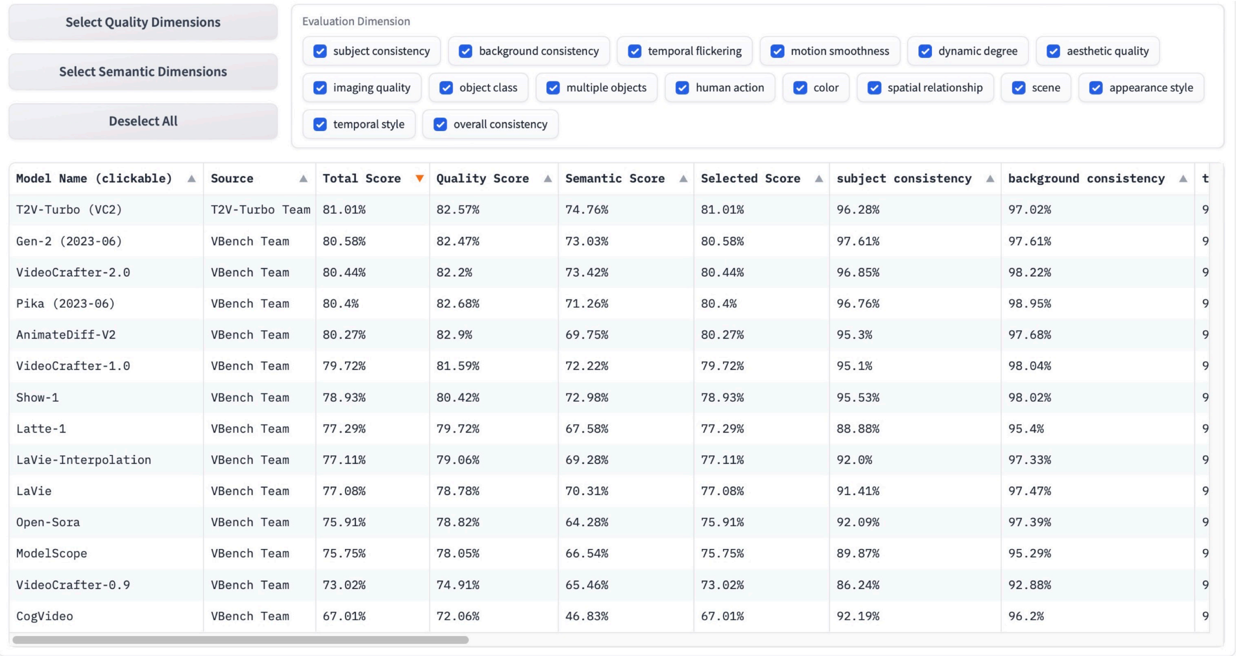
VBench Leaderboard

- 14 T2V models, 12 I2V models
- *Join our leaderboard!*

Leaderboard

VBench

Comprehensive Benchmark Suite for Video Generative Models



Select Quality Dimensions

Select Semantic Dimensions

Deselect All

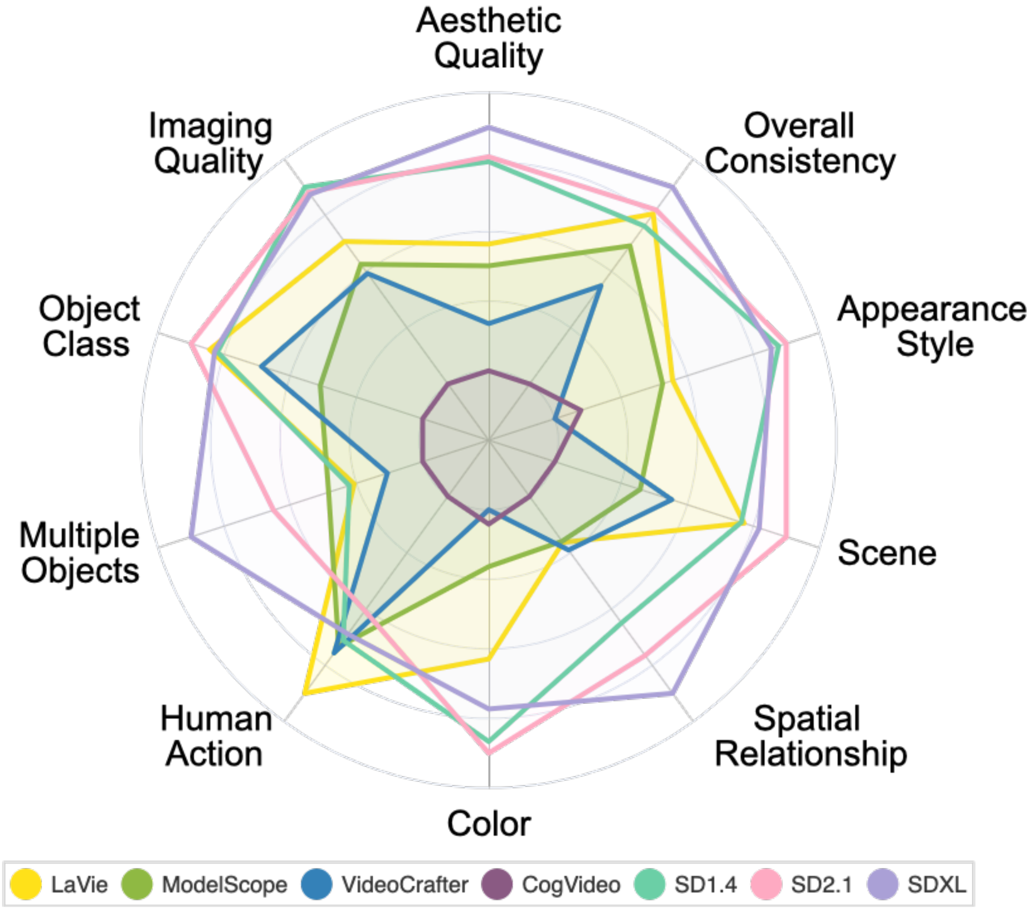
Evaluation Dimension

☑ 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 ☑ overall consistency

Model Name (clickable) ▲	Source ▲	Total Score ▼	Quality Score ▲	Semantic Score ▲	Selected Score ▲	subject consistency ▲	background consistency ▲	t
T2V-Turbo (VC2)	T2V-Turbo Team	81.01%	82.57%	74.76%	81.01%	96.28%	97.02%	9
Gen-2 (2023-06)	VBench Team	80.58%	82.47%	73.03%	80.58%	97.61%	97.61%	9
VideoCrafter-2.0	VBench Team	80.44%	82.2%	73.42%	80.44%	96.85%	98.22%	9
Pika (2023-06)	VBench Team	80.4%	82.68%	71.26%	80.4%	96.76%	98.95%	9
AnimateDiff-V2	VBench Team	80.27%	82.9%	69.75%	80.27%	95.3%	97.68%	9
VideoCrafter-1.0	VBench Team	79.72%	81.59%	72.22%	79.72%	95.1%	98.04%	9
Show-1	VBench Team	78.93%	80.42%	72.98%	78.93%	95.53%	98.02%	9
Latte-1	VBench Team	77.29%	79.72%	67.58%	77.29%	88.88%	95.4%	9
LaVie-Interpolation	VBench Team	77.11%	79.06%	69.28%	77.11%	92.0%	97.33%	9
LaVie	VBench Team	77.08%	78.78%	70.31%	77.08%	91.41%	97.47%	9
Open-Sora	VBench Team	75.91%	78.82%	64.28%	75.91%	92.09%	97.39%	9
ModelScope	VBench Team	75.75%	78.05%	66.54%	75.75%	89.87%	95.29%	9
VideoCrafter-0.9	VBench Team	73.02%	74.91%	65.46%	73.02%	86.24%	92.88%	9
CogVideo	VBench Team	67.01%	72.06%	46.83%	67.01%	92.19%	96.2%	9

Evaluation Results

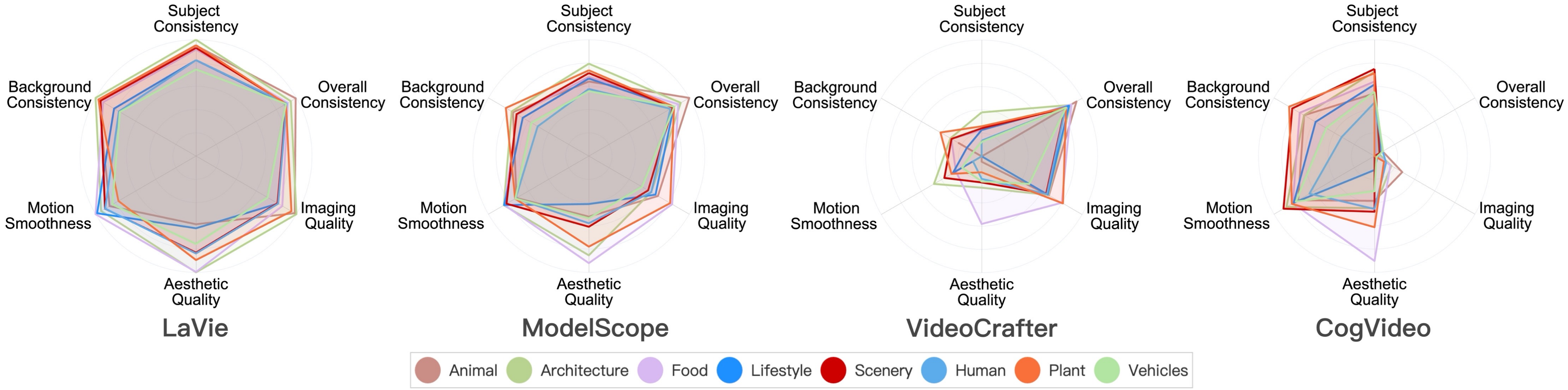
Video vs. Image Generative Models



- gap with T2I in compositionality
  - e.g., multiple objects,
  - e.g., spatial relations

Evaluation Results

Content Categories



- uncovering hidden potential of models in specific content categories
  - *e.g.,* CogVideo has strong aesthetics in Food category.
  - CogVideo's potential in aesthetics by improving such ability in other content types.
  - we recommend *evaluating video generation models not just based on ability dimensions but also considering specific content categories to uncover their hidden potential.*

Fully Open-Source

- *Evaluation Method Suite (code)*
- *Prompt Suite (text prompts)*
- *Human Preference Annotations*
- *Generated Videos (mp4)*

  LaVie,ModelScope,CogVideo,Show-1,
  VideoCrafter-0.9/1/2, Pika,Gen-2,
  OpenSora (more to be added)

```
pip install vbench
```

GitHub

Serial Works in Progress

VBENCH-I2V

Image-to-Video (I2V): multi-ratio and multi-scale image benchmark, I2V evaluation dimensions

VBENCH-Long

for longer videos (e.g., 10 sec, 20 sec, 1 min)

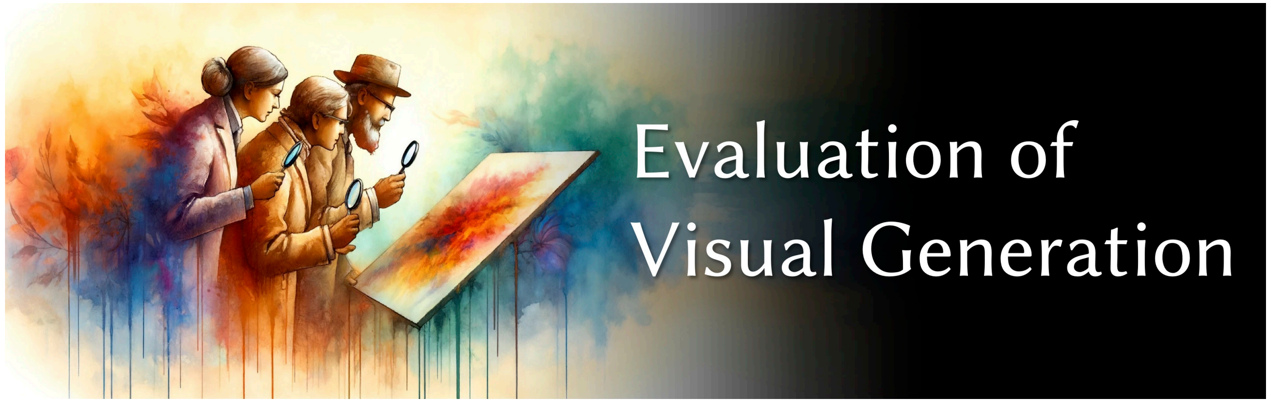
VBENCH-Trustworthiness

non-technical aspects of video generation model: culture, bias, safety

Credits: mainly developed and maintained by the team of VBench Contributors (order based on the time joining the project): Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Nattapol Chanpaisit, Xiaojie Xu, Qianli Ma, Ziyue Dong

Evaluating Visual Generation

- Towards A Better Metric for Text-to-Video Generation
- FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation
- EvalCrafter: Benchmarking and Evaluating Large Video Generation



Paper List

CVPR
JUNE 17-21, 2024
SEATTLE, WA

VBench

Comprehensive Benchmark Suite for Video Generative Models

Ziqi Huang[1], Yinan He[2]*, Jiashuo Yu[2]*, Fan Zhang[2]*, Chenyang Si[1], Yuming Jiang[1], Yuanhan Zhang[1], Tianxing Wu[1], Qingyang Jin[1], Nattapol Chanpaisit[1], Yaohui Wang[2], Xinyuan Chen[2], Limin Wang[4,2], Dahua Lin[2,3]†, Yu Qiao[2]†, Ziwei Liu[1]†*

(* equal contributions, † corresponding authors)

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

[3] The Chinese University of Hong Kong [4] Nanjing University

Q&A

Poster Session

- Friday 10:30am-12:00
- Arch 4A-E Poster #219
- Welcome any questions & discussions

CVPR
JUNE 17-21, 2024
SEATTLE, WA

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

S-LAB
FOR ADVANCED
INTELLIGENCE

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

VBENCH

Comprehensive Benchmark Suite for Video Generative Models

Code

https://github.com/Vchitect/VBench