KlingAI

An Introduction to Kling and Our Research towards More Powerful Video Generation Models

Pengfei Wan
Kuaishou Technology

Outline

KlingAI

Introduction

- What is Kuaishou and Kling
- Kling's main capabilities and features

Our Research

01 Advanced Model Architectures and Generation Algorithms

02 Powerful Interaction and Control Capacities

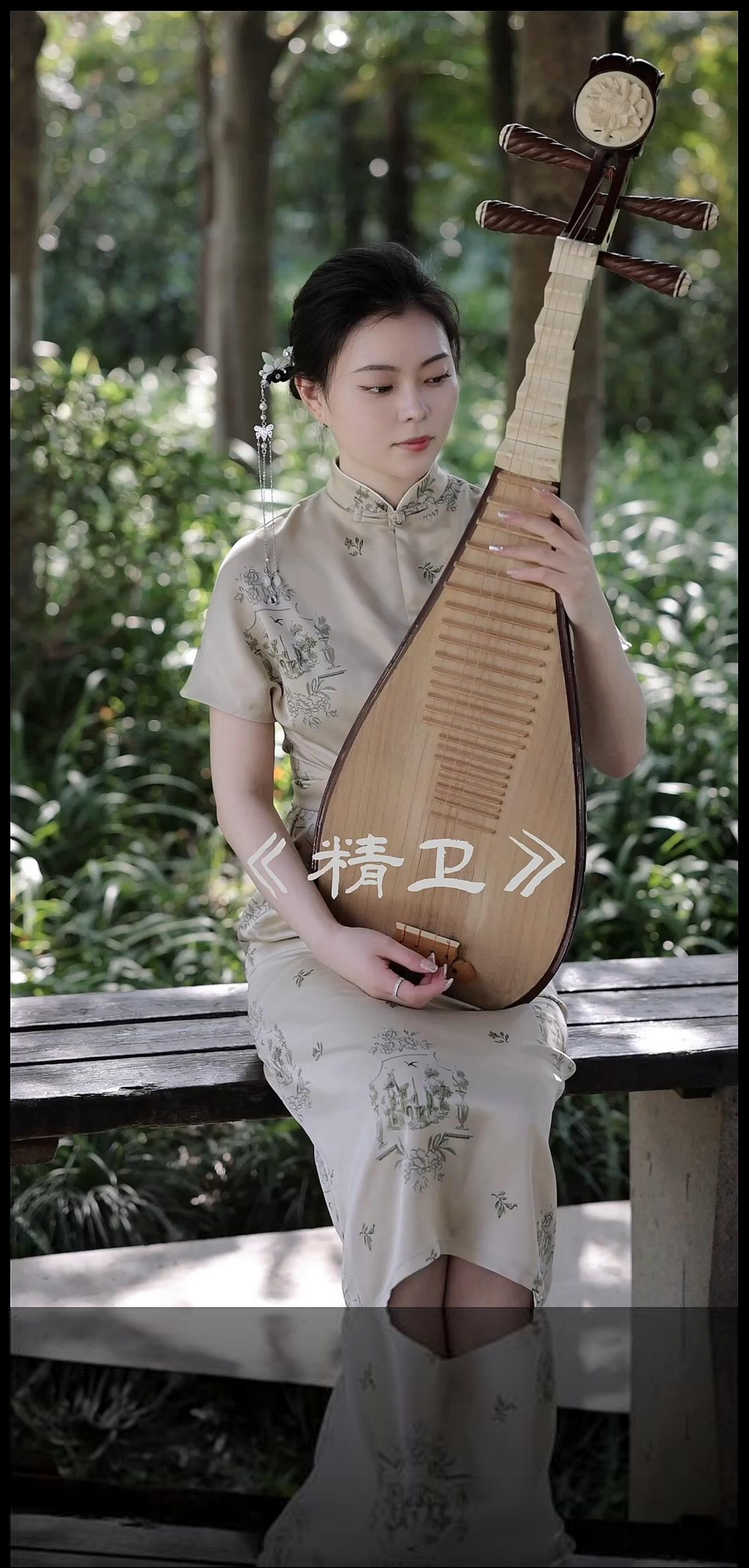
03 Accurate Evaluation and Alignment Mechanisms

04 Multimodal Perception and Reasoning

What is Kuaishou and Kling

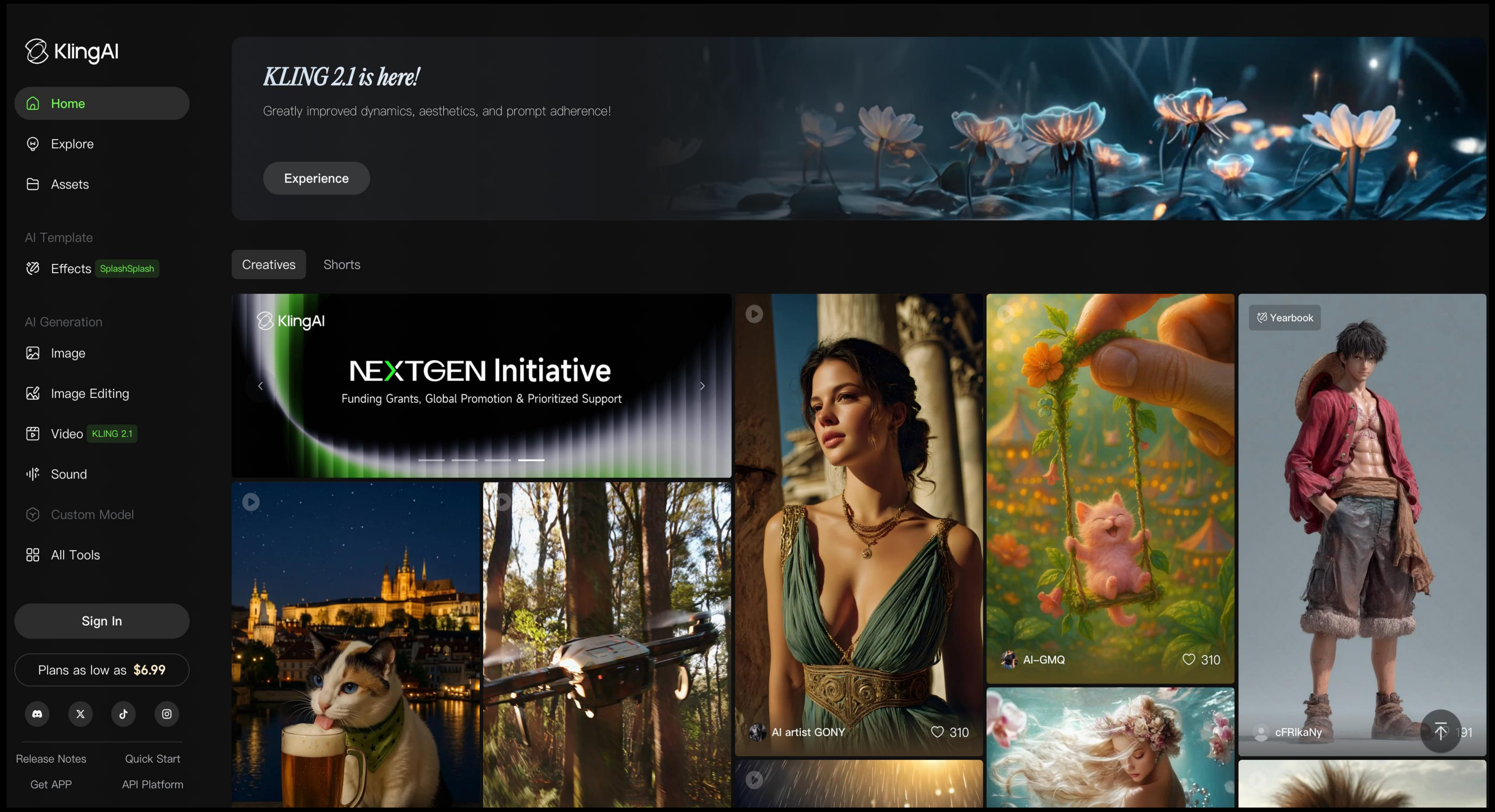
KlingAI

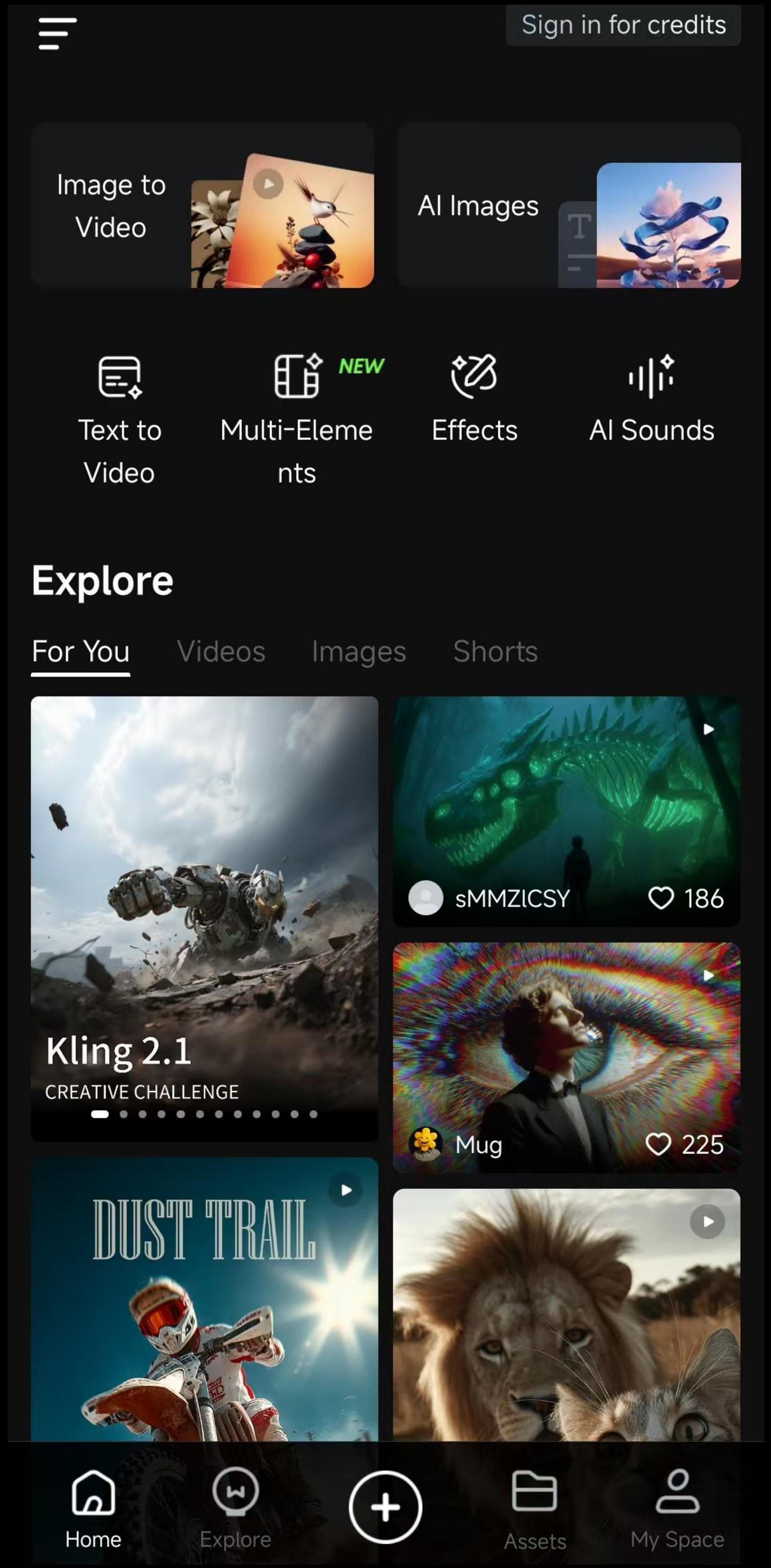
- Kuaishou is a video content community in China, with over 400 million DAUs. Kuaishou is founded in 2011, is one of the earliest short-video company.
- Embrace All Lifestyles: Kuaishou is designed to elevate the often-overlooked, yet diverse, vibrant and energetic communities and lifestyles of people.

What is Kuaishou and Kling

- Kling is the overall name of Kuaishou's video generation models and related capabilities.
- Our product "KlingAI" has over 20 million users globally, people can access our service through web and apps.





Recent Release: Kling 2.1 Model

Kling's capabilities: Text2Video, Image2Video

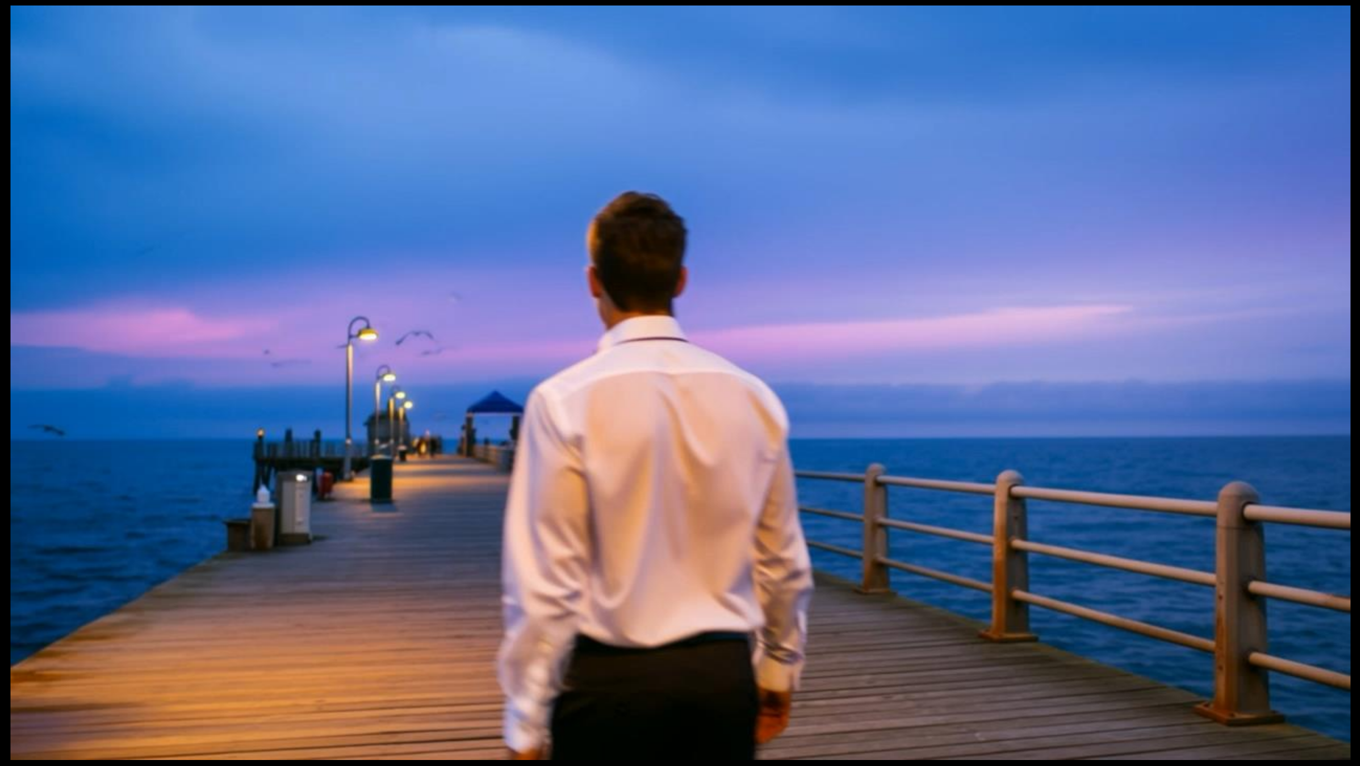
The dinosaur charges towards the camera, with motion blur and the camera shaking.

Kling's capabilities: Elements2Video

KlingAI

A standing cat character wearing a jacket and sunglasses strikes a pose towards the camera on the stage.

Kling's capabilities: Versatile Video Editing



Swap the panda from @Image for the man from @Video

Kling's capabilities: Versatile Video Editing

KlingAI

Seamlessly add the toy from @Image to the box from @Video

Kling's capabilities: Generating Image / Audio

Explosion, running

Outline

Introduction

- What is Kuaishou and Kling
- Kling's main capabilities and features

Our Research

01 Advanced Model Architectures and Generation Algorithms

02 Powerful Interaction and Control Capacities

03 Accurate Evaluation and Alignment Mechanisms

04 Multimodal Perception and Reasoning

KlingAI

01 Advanced Model Architectures and Generation Algorithms

- Scaling Laws for Video Generation

  Towards Precise Scaling Laws For Video Diffusion Transformers

- MoE Architecture for Visual Generation

  DiffMoE: Dynamic Token Selection For Scalable Diffusion Transformers

Scaling Laws for Video Generation

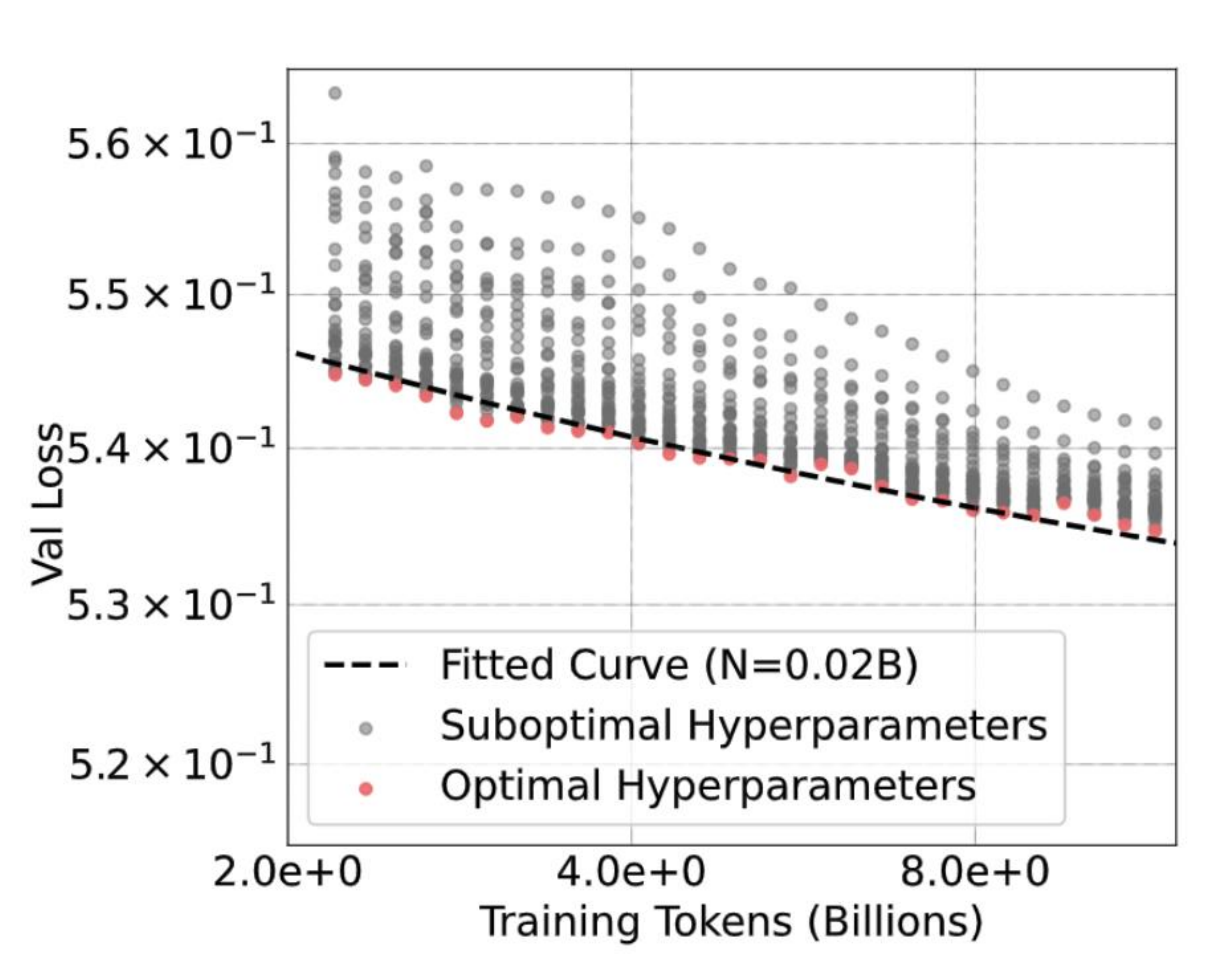
😟 Scaling Laws for LLM have been well-studied, but precise scaling laws for video generation models basically not exist.

Towards Precise Scaling Laws for Video Diffusion Transformers

Yuanyang Yin[1*], Yaqi Zhao[2*], Mingwu Zheng[3], Ke Lin[3]
Jiarong Ou[3], Rui Chen[3], Victor Shea-Jay Huang[2], Jiahao Wang[3]
Xin Tao[3], Pengfei Wan[3], Di Zhang[3], Baoqun Yin[1†], Wentao Zhang[2†], Kun Gai[3]
[1]University of Science and Technology of China [2]Peking University [3]Kuaishou Technology

😃 We established a precise relationship among validation loss, model size, and compute budget, confirming the existence of scaling laws in video diffusion transformers.

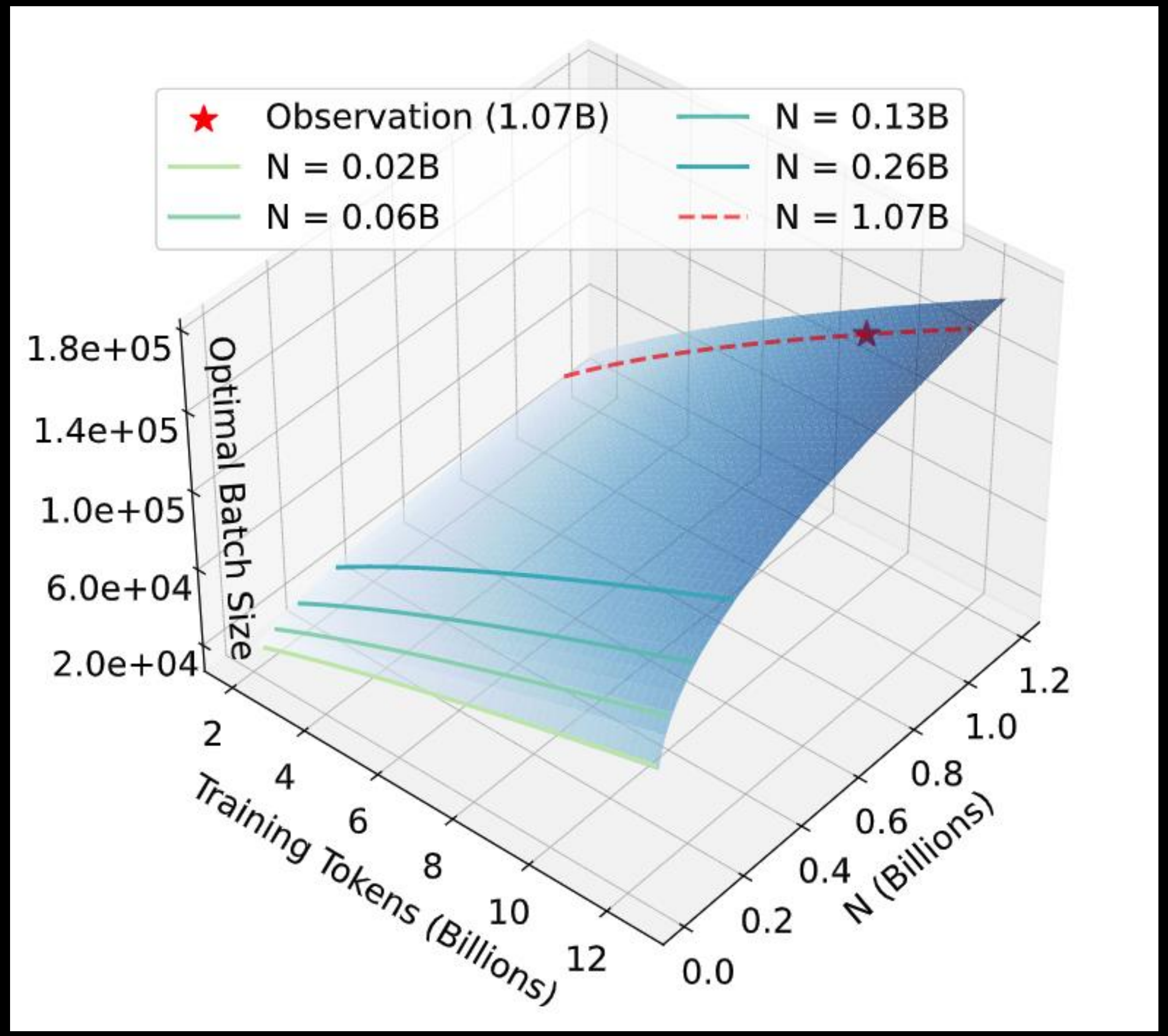
Scaling Laws for Video Generation



- Video diffusion models are highly sensitive to hyperparameters, such as learning rate and batch size.
- Therefore, in order to have a precise scaling law for video generation, identifying optimal hyperparameters is essential.

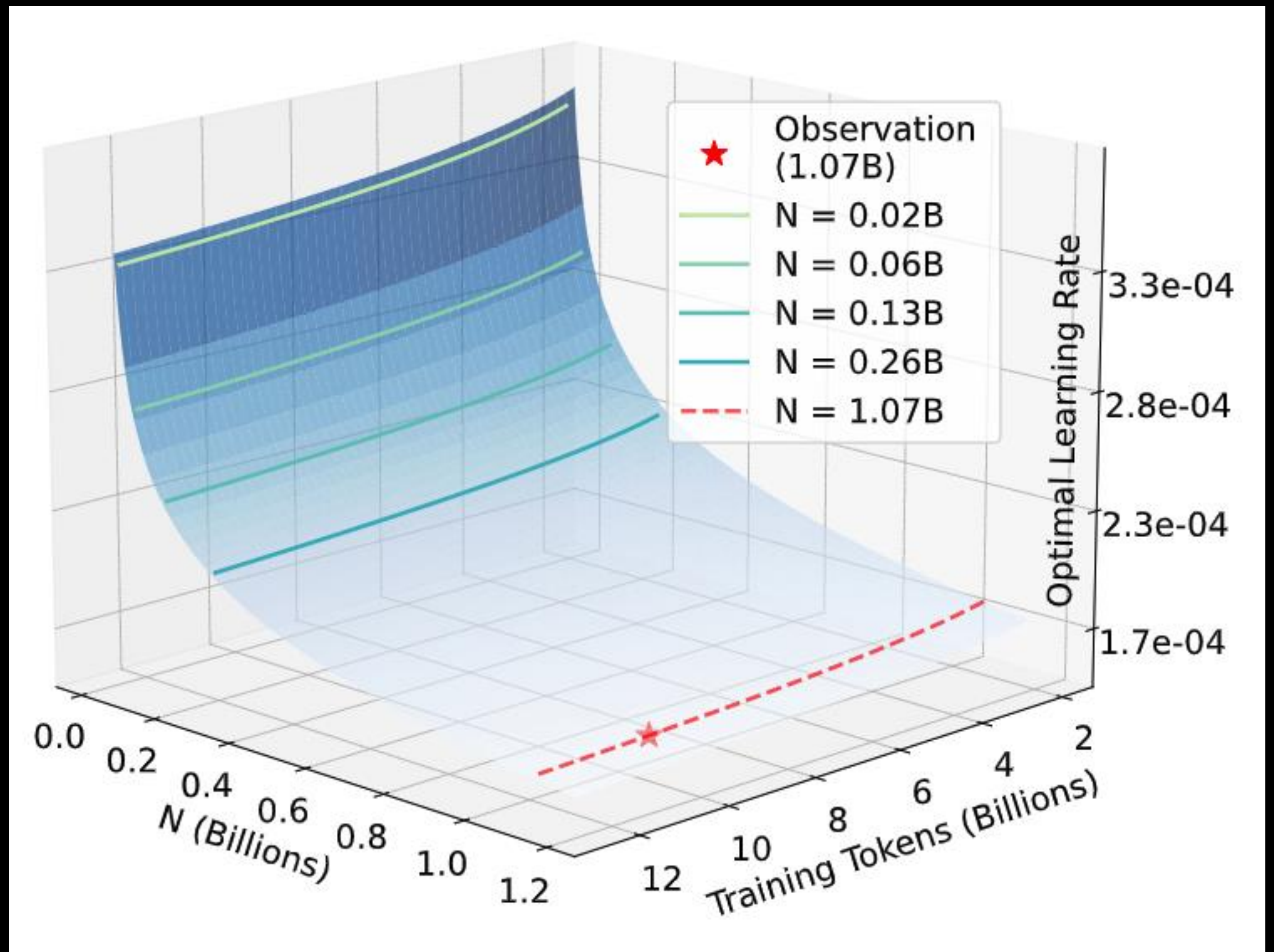
Scaling Laws for Video Generation

- Through analyses and experiments, we presented explicit equations for optimal learning rate and batch size w.r.t. model size and training tokens.



Optimal batch size
w.r.t. model size & training tokens

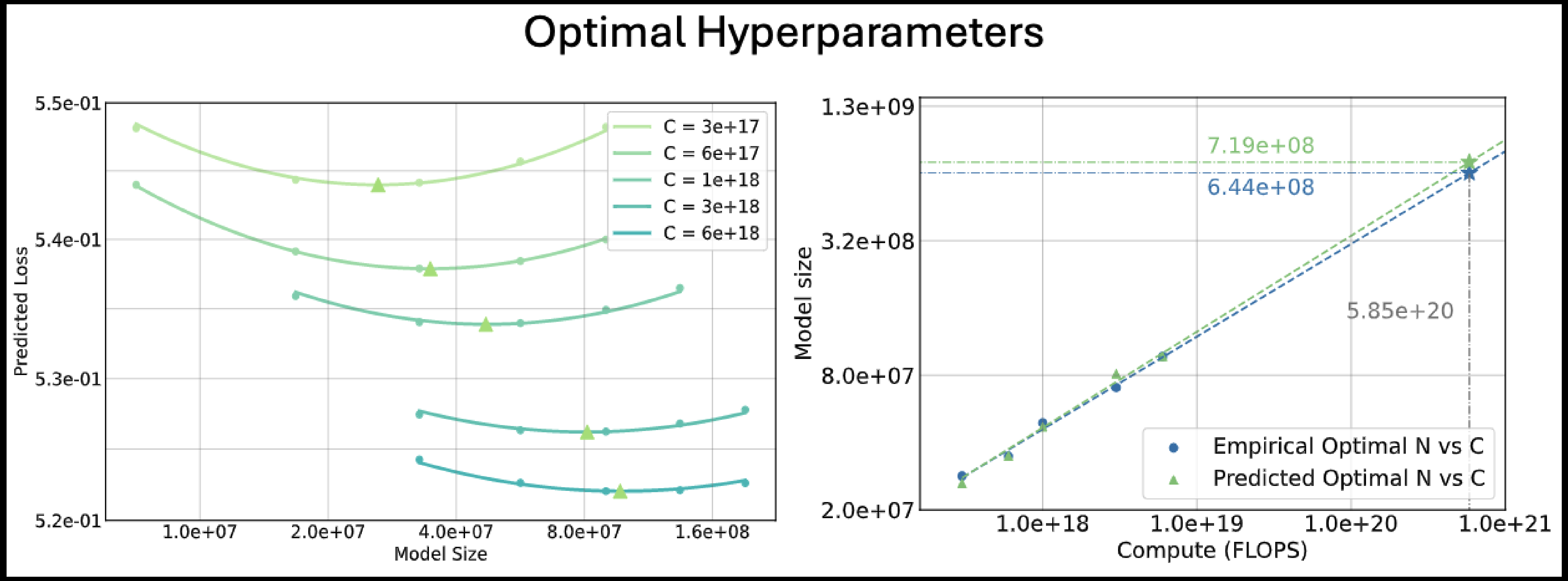
$$B_{\text{opt}} = \alpha_B T^{\beta_B} N^{\gamma_B}$$

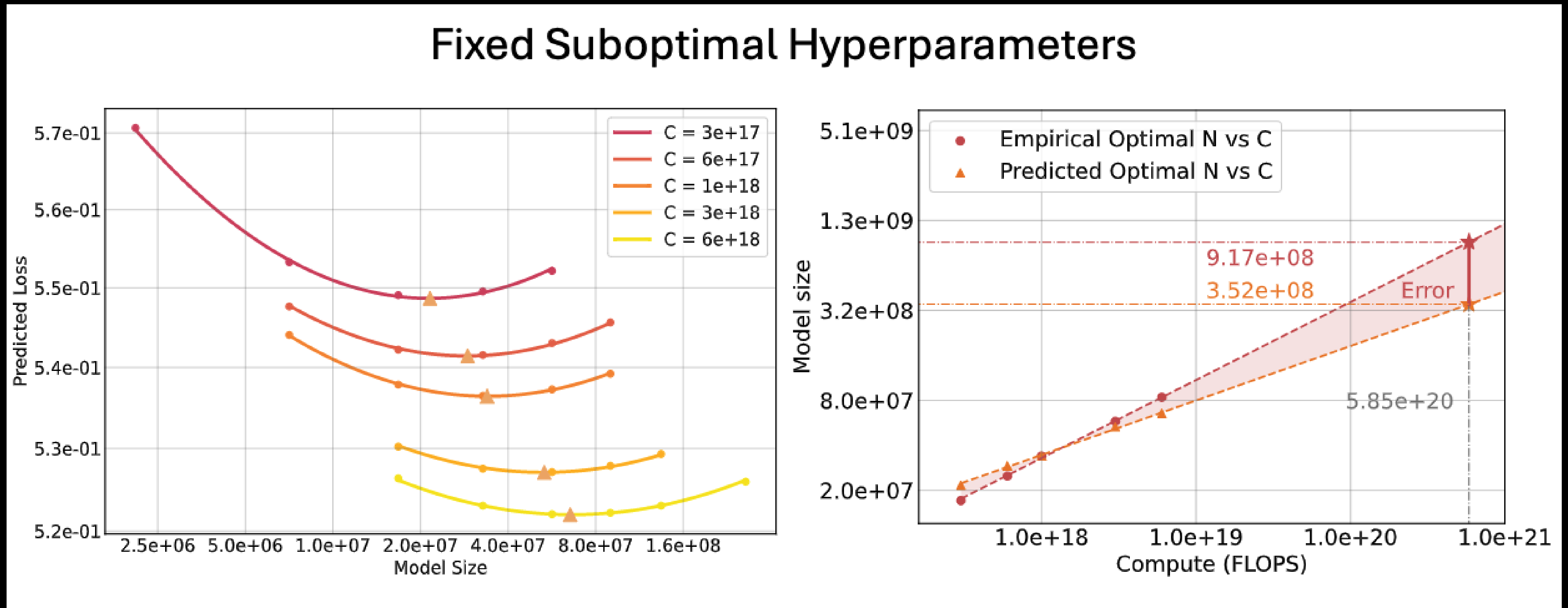


Optimal learning rate
w.r.t. model size & training tokens

$$\eta_{\text{opt}} = \alpha_\eta T^{\beta_\eta} N^{\gamma_\eta}$$

Scaling Laws for Video Generation





- Based on optimal hyperparameters, we then establish an accurate relationship among validation loss, model size, and training compute budget.

$$L(T, N) = \left(\frac{T_c}{T}\right)^{\alpha_T} + \left(\frac{N_c}{N}\right)^{\alpha_N} + L_\infty$$

$$\hat{N}_{\text{opt}} = 0.8705 \cdot C^{0.4294}$$

- We confirmed the existence of scaling laws in video diffusion transformers. It can be used to guide the design of better models and training strategies.

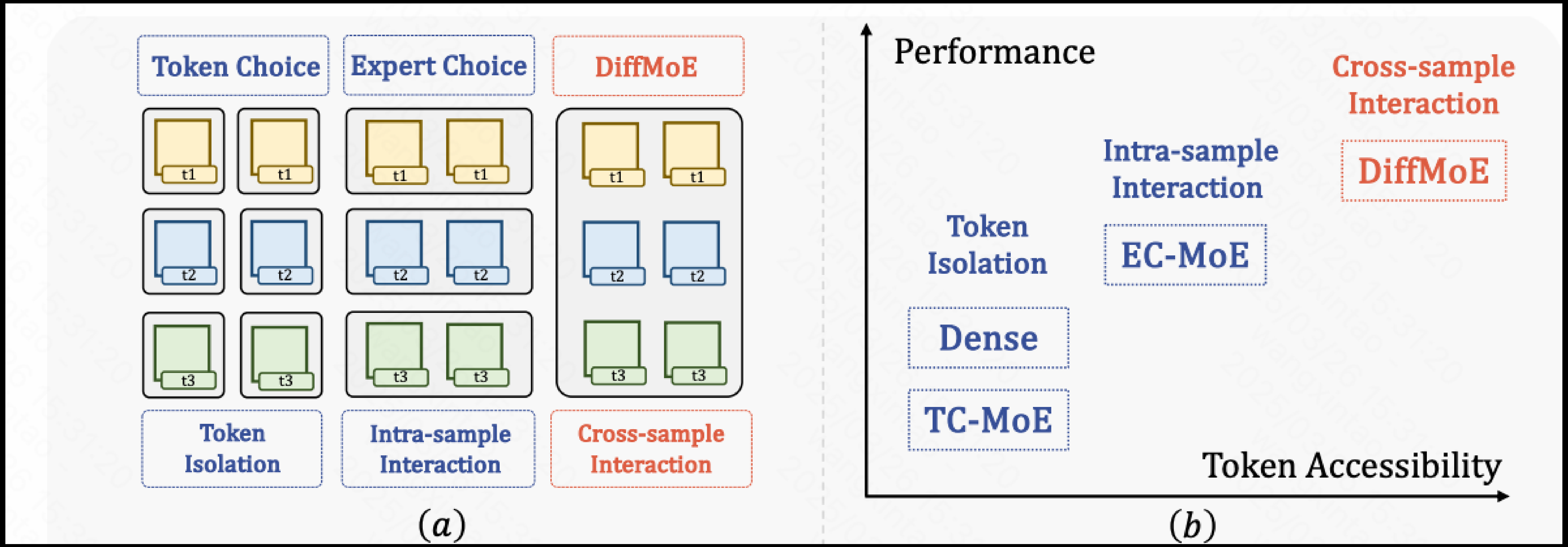
MoE for Visual Generation

😟 Mixture-of-Experts (MoE) is widely used in LLM, but is less popular in diffusion / flow-based visual generation models, due to poor performance.

😃 We present a MoE-based architecture for DiT that enables experts to access global token distributions, outperforming competing MoE approaches.

MoE for Visual Generation

- We identify the importance of global token distribution accessibility (choose from as much tokens as possible) for MoE diffusion models.
- Proposed a batch-level token pool, enabling experts to access a global token distribution spanning different noise levels and conditions.



MoE for Visual Generation

- DiffMoE **outperformed dense architectures with 3× activated parameters** in image generation experiments.

Diffusion Models (3000K)	# Avg. Activated Params.	FID↓	IS↑	Precision↑	Recall↑
Dense-DiT-XL-FlowG (cfg=1.5, ODE)	675M (1.5x)	2.52	273.78	0.84	0.56
Dense-DiT-XXL-Flow-G (cfg=1.5, ODE)	951M (2x)	2.41	281.96	0.84	0.57
Dense-DiT-XXXL-Flow-G (cfg=1.5, ODE)	1353M (3x)	2.37	291.29	0.84	0.57
DiffMoE-L-E8-Flow-G (cfg=1.5, ODE)	458M (1x)	2.40	280.30	0.83	0.57
DiffMoE-L-E16-Flow-G (cfg=1.5, ODE)	458M (1x)	2.36	287.26	0.83	0.58
DiffMoE-XL-E16-Flow-G (cfg=1.5, ODE)	675M (1.5x)	**2.30**	291.23	0.83	0.58

Results of ImageNet Class-Conditional Generation

02 Powerful Interaction and Control Capacities

- ### Unified Multi-Task Video Generative Model

  FullDiT: Multi-Task Video Generative Foundation Model with Full Attention

- ### Interactive Generative Video for Gaming

  GameFactory: Creating New Games with Generative Interactive Videos

Unified Multi-Task Video Generation Model

😟 Typically, people need to develop different models for different controllable video generation tasks, using various types of adapters.

FullDiT

Multi-Task Video Generative Foundation Model with Full Attention

Xuan Ju[1,2], Weicai Ye[1*], Quande Liu[1], Qiulin Wang[1], Xintao Wang[1],
Pengfei Wan[1], Di Zhang[1], Kun Gai[1], Qiang Xu[2*]
[1]Kuaishou Technology, [2]The Chinese University of Hong Kong, [*]Corresponding Author

😃 We proposed an unified multi-task video generative model, that synergistically integrates multiple input conditions. FullDiT reduces parameter overhead, avoids conditions conflict, and shows scalable and emergent ability.

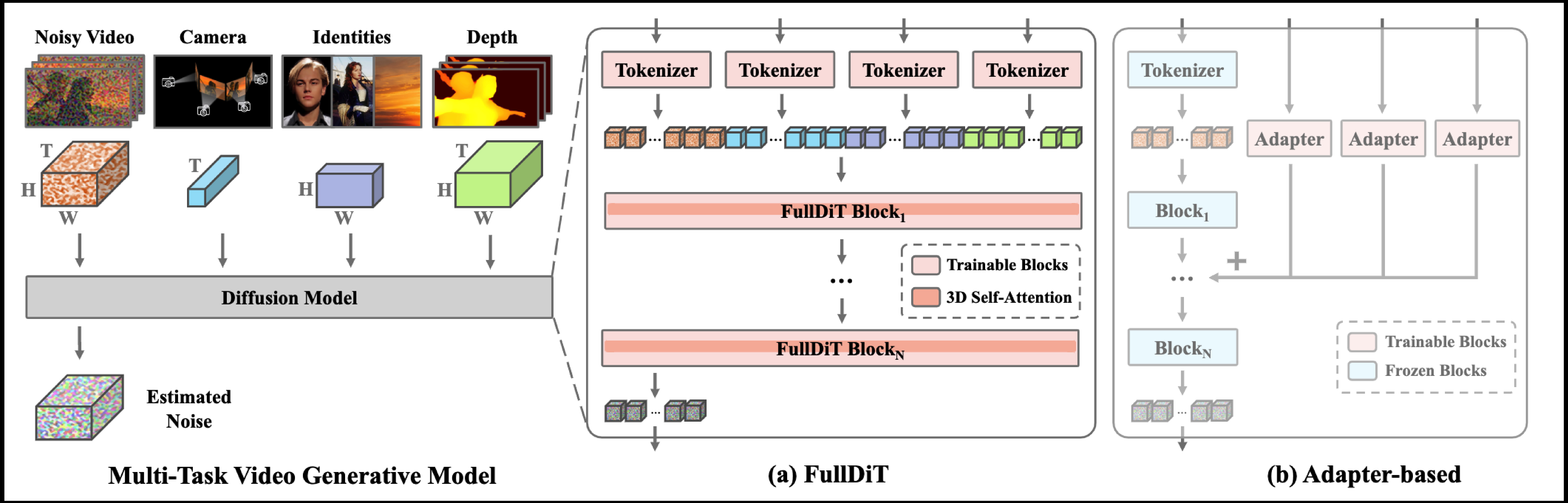
Unified Multi-Task Video Generation Model

Powerful Interaction and Control Capacities

KlingAI

- Conditions as multimodal context: tokenize all spatial-temporal conditions into tokens and then concatenate them with the noisy video tokens.

- "Full attention": all tokens are jointly modeled in DiT without any modification to the model architecture, which is different in nature with adapter-based methods.

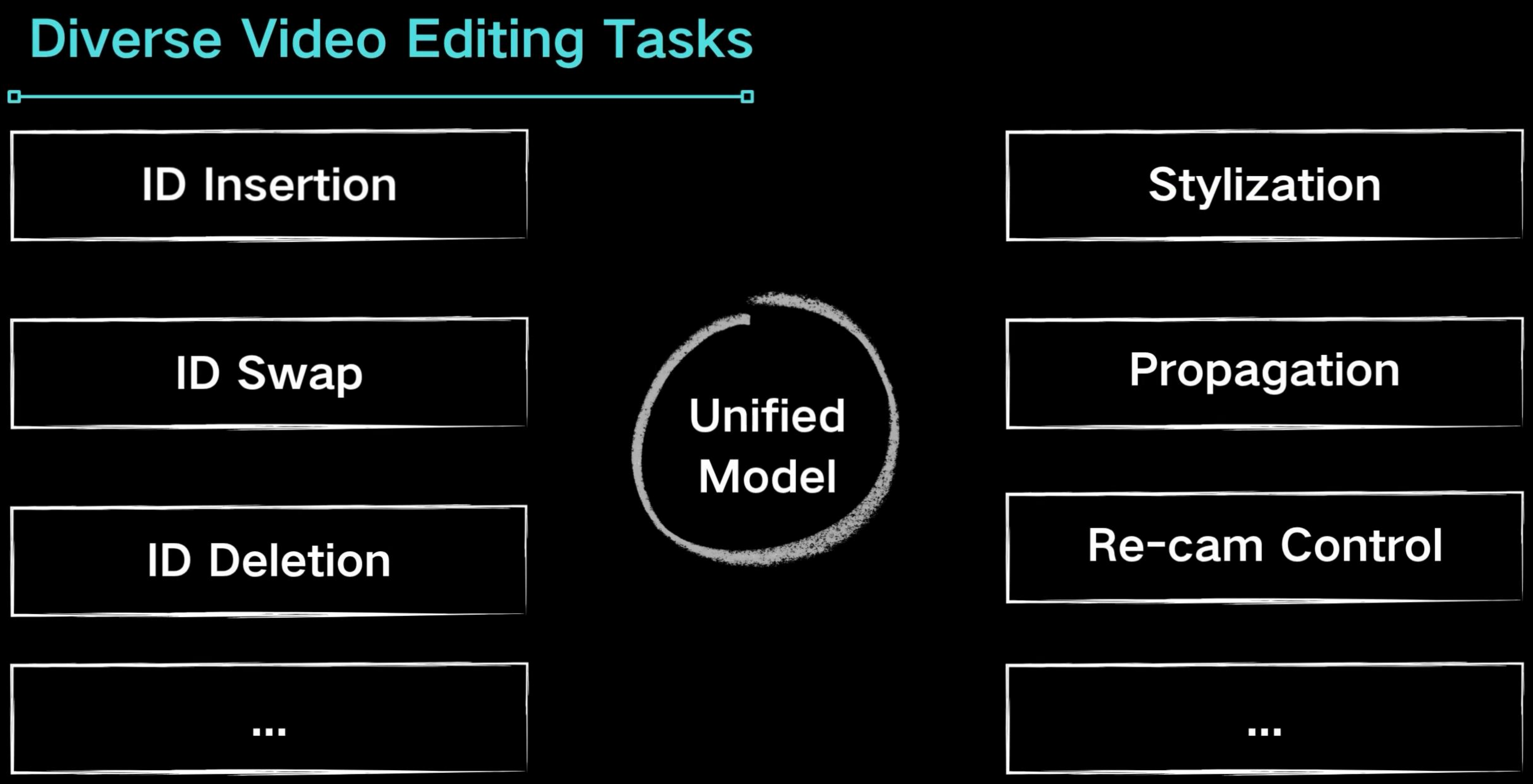


Unified Multi-Task Video Generation Model

Powerful Interaction and Control Capacities

KlingAI

- Because of the long-context learning ability, FullDiT can flexibly take different combinations of input to generate desired videos with good generalization performance.



In a subsequent work, we show the potential of FullDiT to unify diverse video editing tasks and the emergent task composition ability.

Interactive Generative Video for Gaming

Powerful Interaction and Control Capacities

KlingAI

😟 Video generation models can serve as generative game engines, but generalizing the (diverse) action control abilities to new games (scenes) remains a problem.

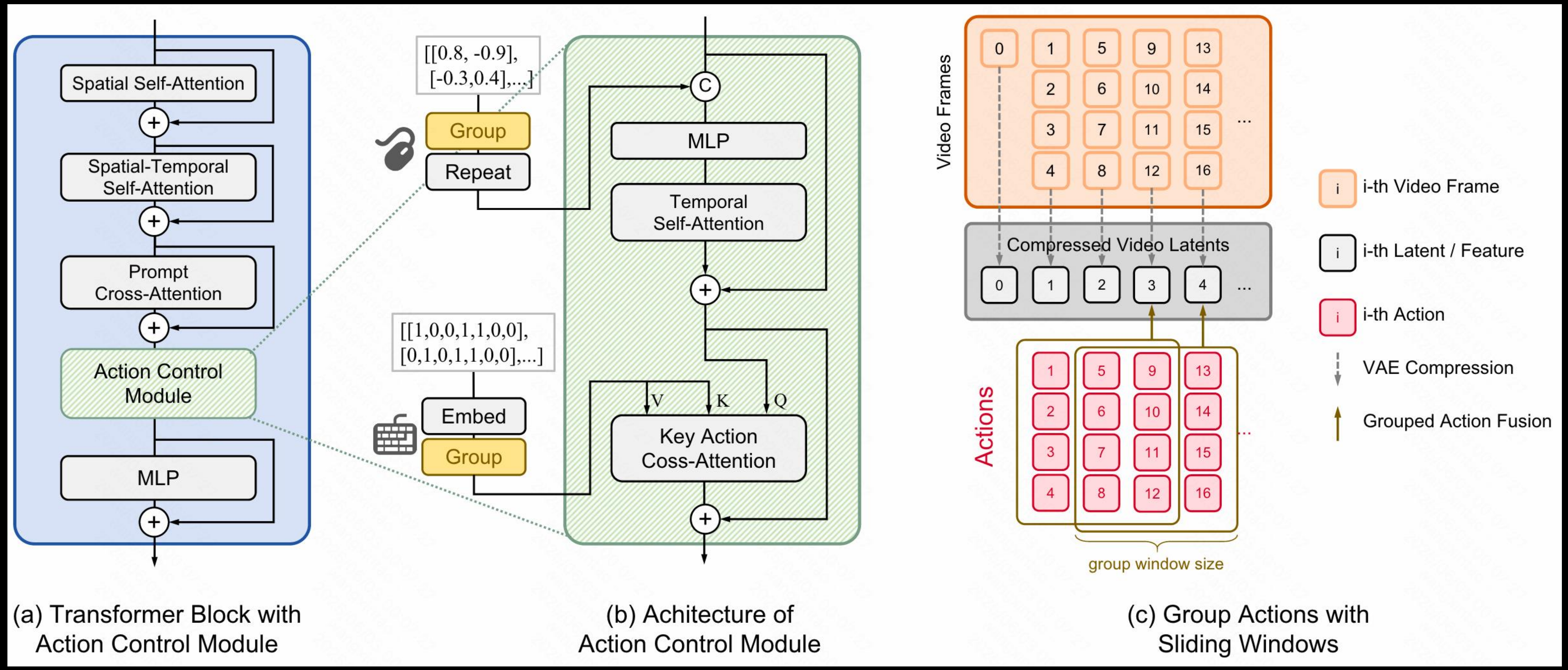
😃 We explored keyboard & mouse control for interactive generative videos, and tackled the challenge of scene generalization in game video generation.

Interactive Generative Video for Gaming

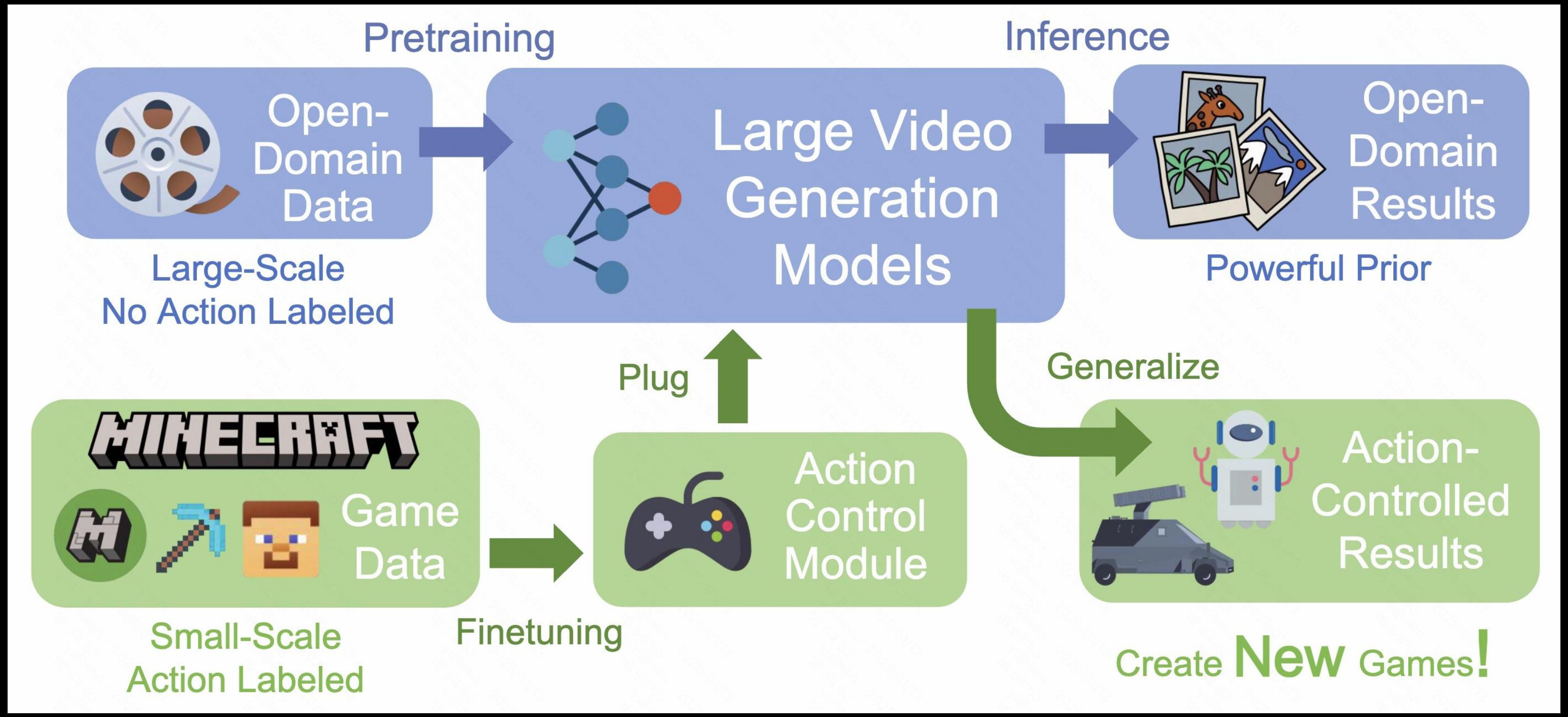
Powerful Interaction and Control Capacities

KlingAI

- Proposed an effective control mechanisms for continuous actions and discrete actions.
- Through multi-phase training, the action control module, learned from a small amount of game data, become generalizable. It can be plugged into any video models to create new games.



Action control module



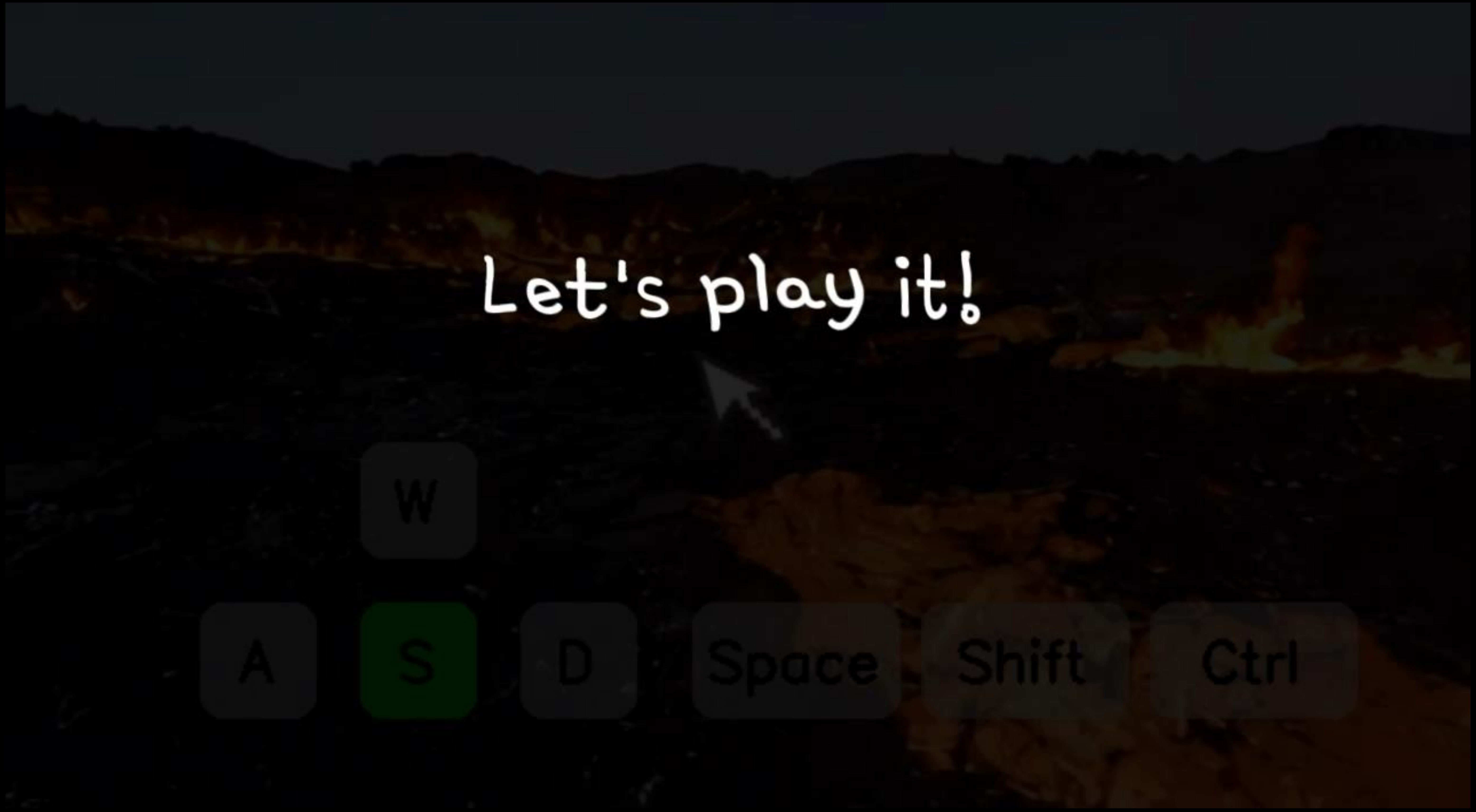
Multi-phase training strategy

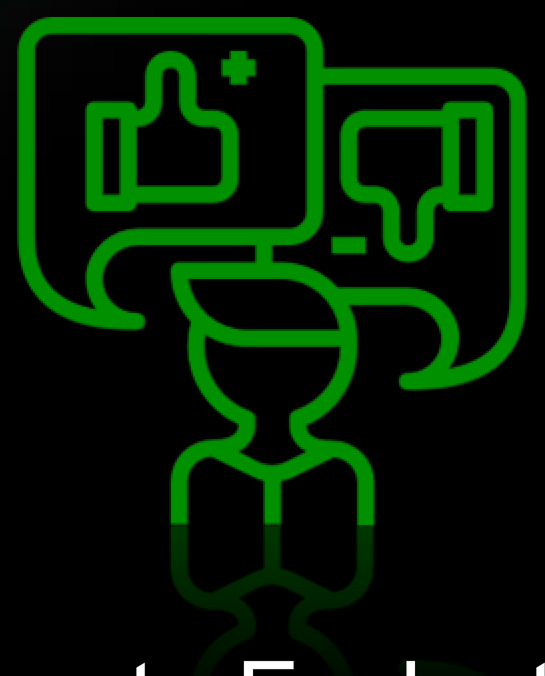
Interactive Generative Video for Gaming

Powerful Interaction and Control Capacities

KlingAI

- Scene-generalizable action control is the core contribution of GameFactory.





03 Accurate Evaluation and Alignment Mechanisms

- RLHF Pipeline for Video Generation

  Improving Video Generation with Human Feedback

- Online RL Algorithm for Visual Generation

  Flow-GRPO: Training Flow Matching Models via Online RL

RLHF Pipeline for Video Generation

KlingAI

😟 Human alignment is important in both LLM and video generation. But it is still challenging to make RLHF for video generation effective.

Improving Video Generation with Human Feedback

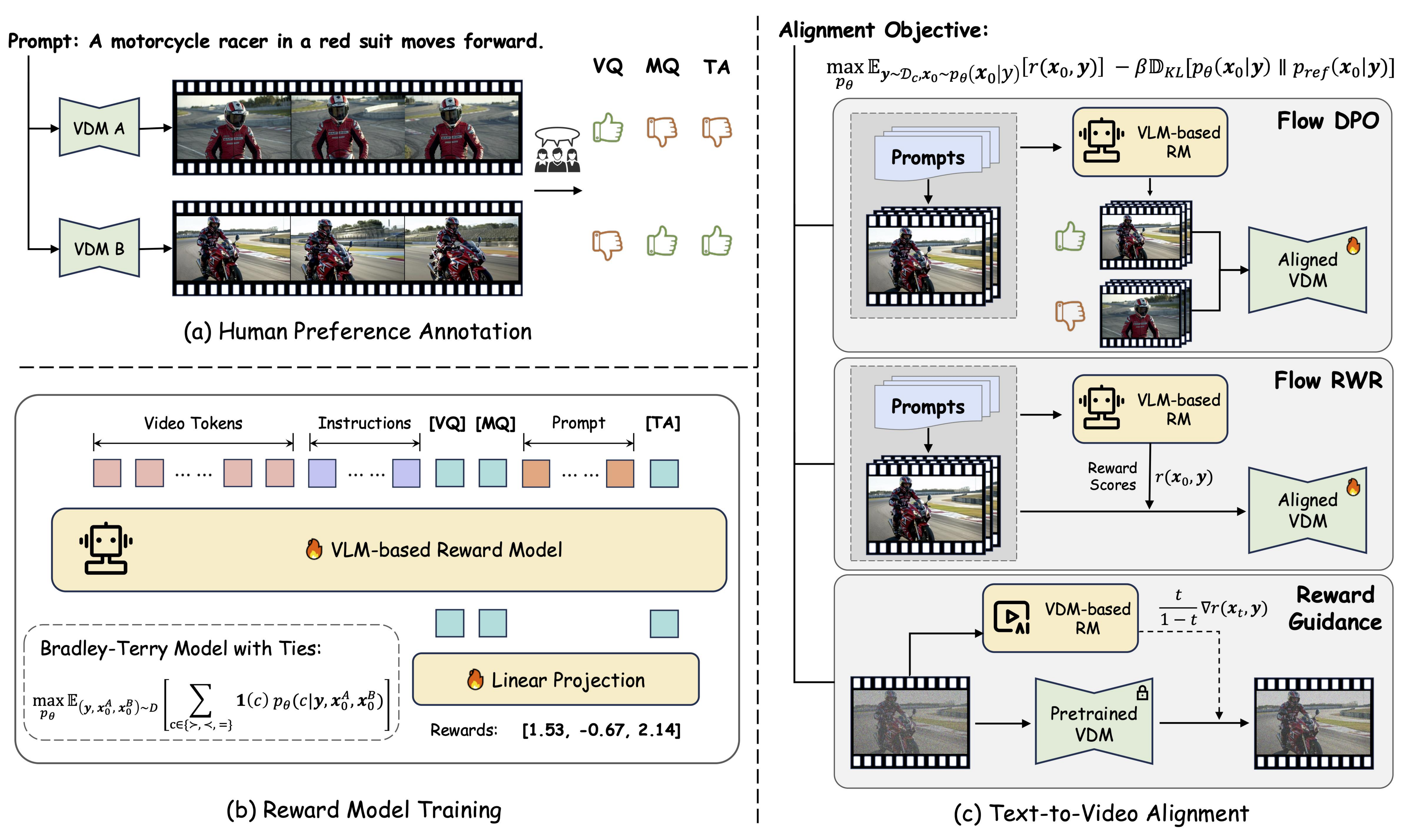
Jie Liu[1,3,5*], Gongye Liu[2,3*], Jiajun Liang[3†], Ziyang Yuan[2,3], Xiaokun Liu[3], Mingwu Zheng[3], Xiele Wu[3,4], Qiulin Wang[3], Wenyu Qin[3], Menghan Xia[3], Xintao Wang[3], Xiaohong Liu[4], Fei Yang[3], Pengfei Wan[3], Di Zhang[3], Kun Gai[3], Yujiu Yang[2✉], Wanli Ouyang[1,5],
[1]The Chinese University of Hong Kong, [2]Tsinghua University, [3]Kuaishou Technology, [4]Shanghai Jiao Tong University, [5]Shanghai AI Laboratory
*Equal contribution †Project Leader ✉Corresponding Author

ArXiv | Code | VideoReward | VideoGen-RewardBench

Eval Dataset | Online Demo(Reward) | Qualitative Results

😃 We presented one of the earliest systematic approach for incorporating human feedback in video generation.

RLHF Pipeline for Video Generation

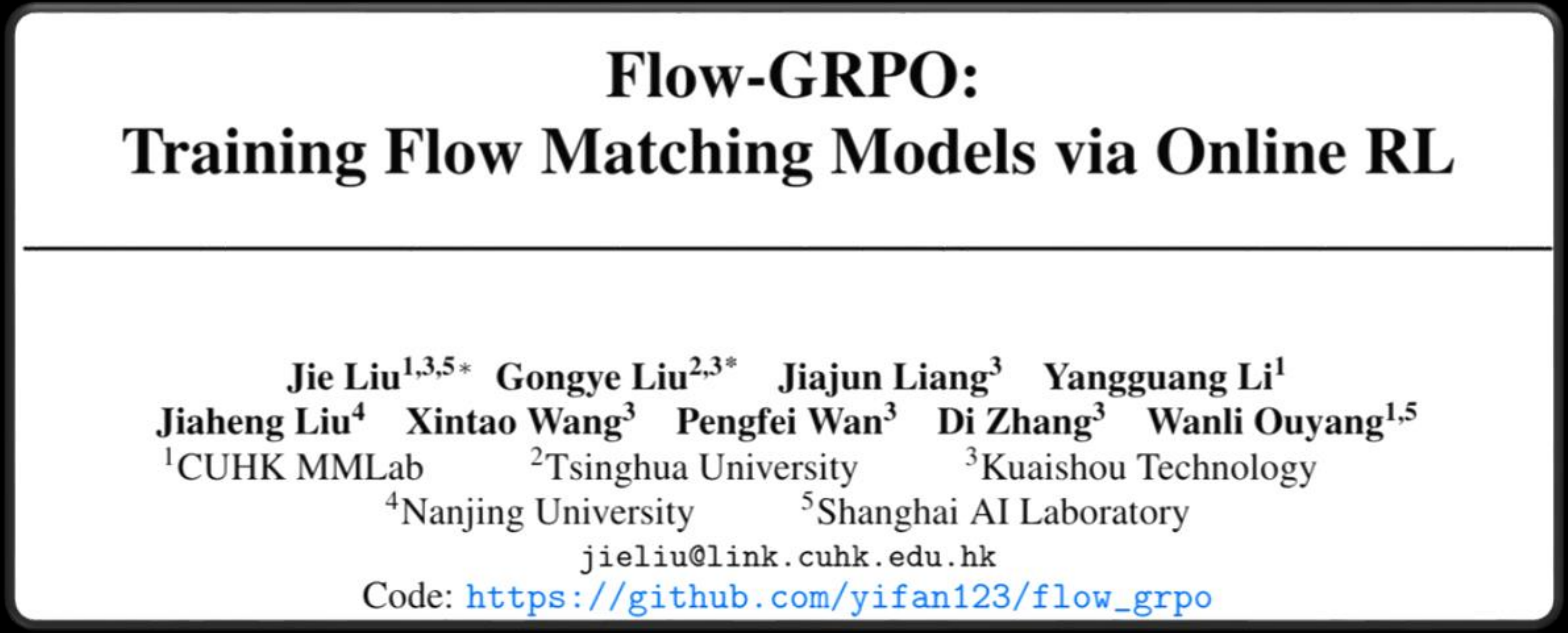


We analyzed and provided a systematic pipeline, including:

- preference dataset
- reward models & benchmark
- various alignment algorithms

Online RL Algorithm for Visual Generation

😟 The potential of online RL for flow matching generative models remains largely unexplored, due to several key technical challenges.



Flow-GRPO:
Training Flow Matching Models via Online RL

Jie Liu[1,3,5]* Gongye Liu[2,3]* Jiajun Liang[3] Yangguang Li[1]
Jiaheng Liu[4] Xintao Wang[3] Pengfei Wan[3] Di Zhang[3] Wanli Ouyang[1,5]
[1]CUHK MMLab [2]Tsinghua University [3]Kuaishou Technology
[4]Nanjing University [5]Shanghai AI Laboratory
jieliu@link.cuhk.edu.hk
Code: https://github.com/yifan123/flow_grpo

😃 We proposed the first method to introduce GRPO to flow matching models, showing that online RL is highly effective for visual generation tasks.

Online RL Algorithm for Visual Generation

Challenge 1: the need for stochasticity in RL conflicts with the deterministic nature of flow matching models.

- We transform a deterministic ODE in flow matching into an equivalent SDE that matches the original model's marginal probability density function at all timesteps, enabling statistical sampling for RL exploration.

$$\mathrm{d}\boldsymbol{x}_t = \left(\boldsymbol{v}_t(\boldsymbol{x}_t) - \frac{\sigma_t^2}{2}\nabla \log p_t(\boldsymbol{x}_t)\right)\mathrm{d}t + \sigma_t \mathrm{d}\boldsymbol{w},$$

Challenge 2: online RL depends on efficient sampling to collect training data, but flow models typically require many iterative steps to generate each sample .

- We find that online RL for flow matching models does not require the standard long timesteps for training sample collection. So we propose to reduce the training denoising steps, improving sampling efficiency.

Online RL Algorithm for Visual Generation

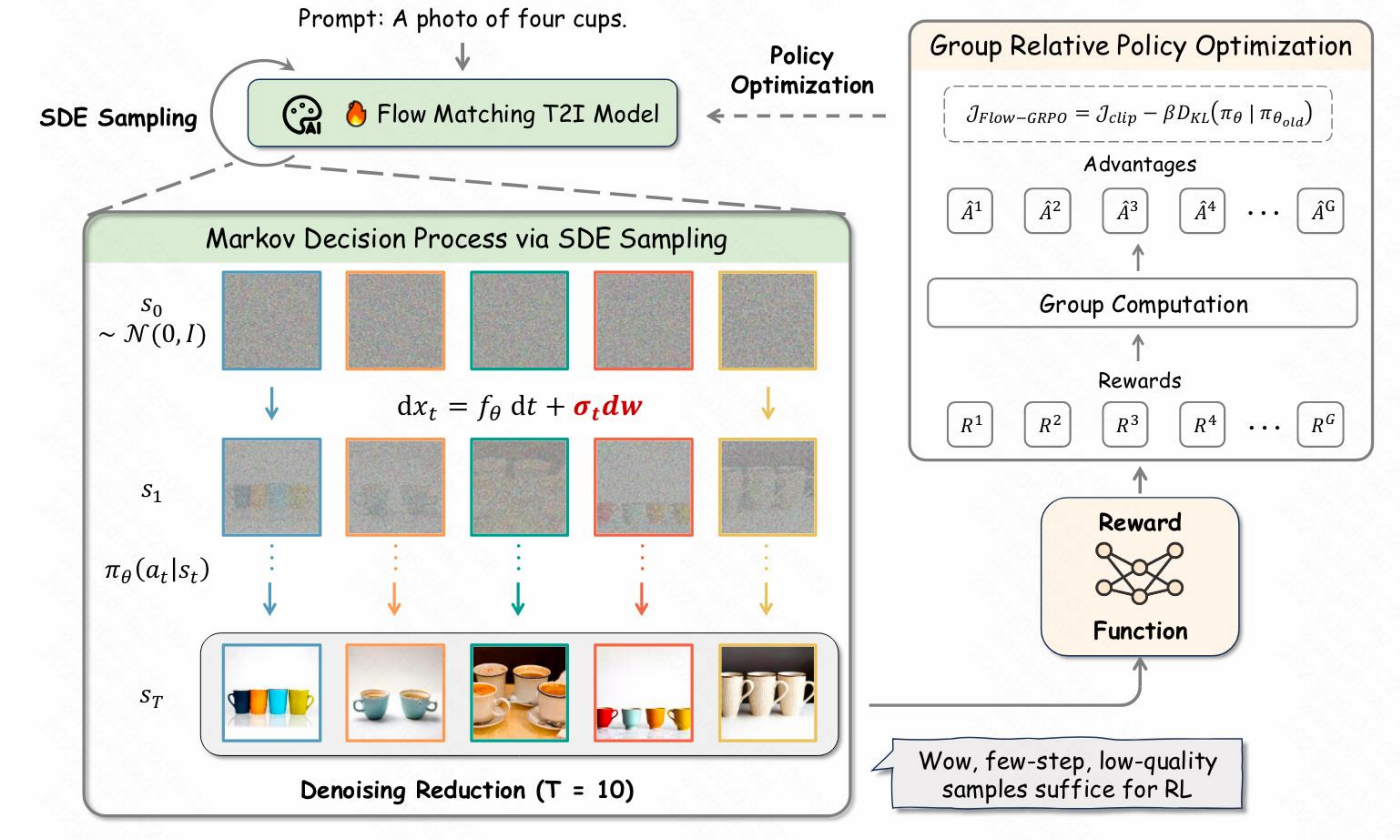
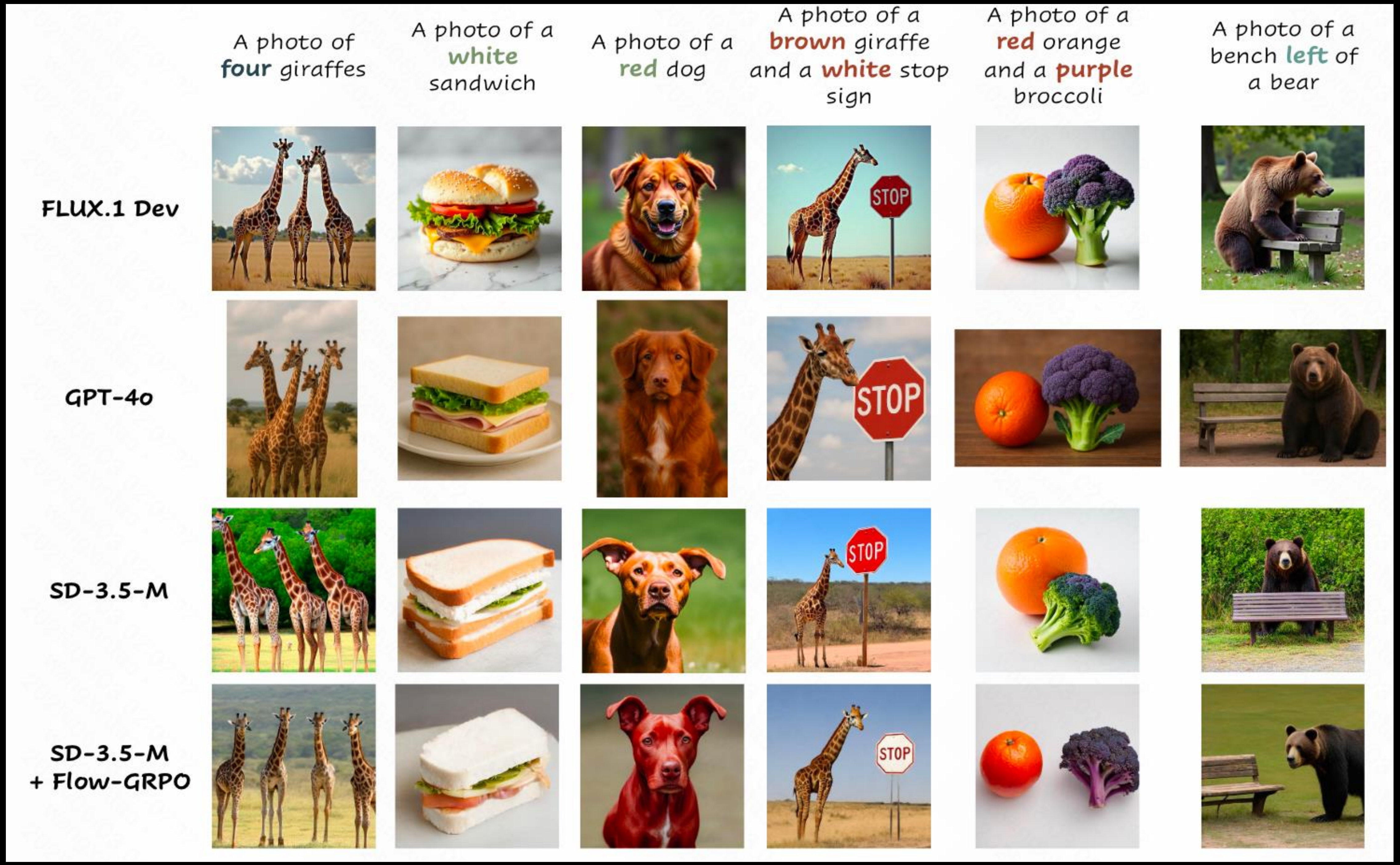


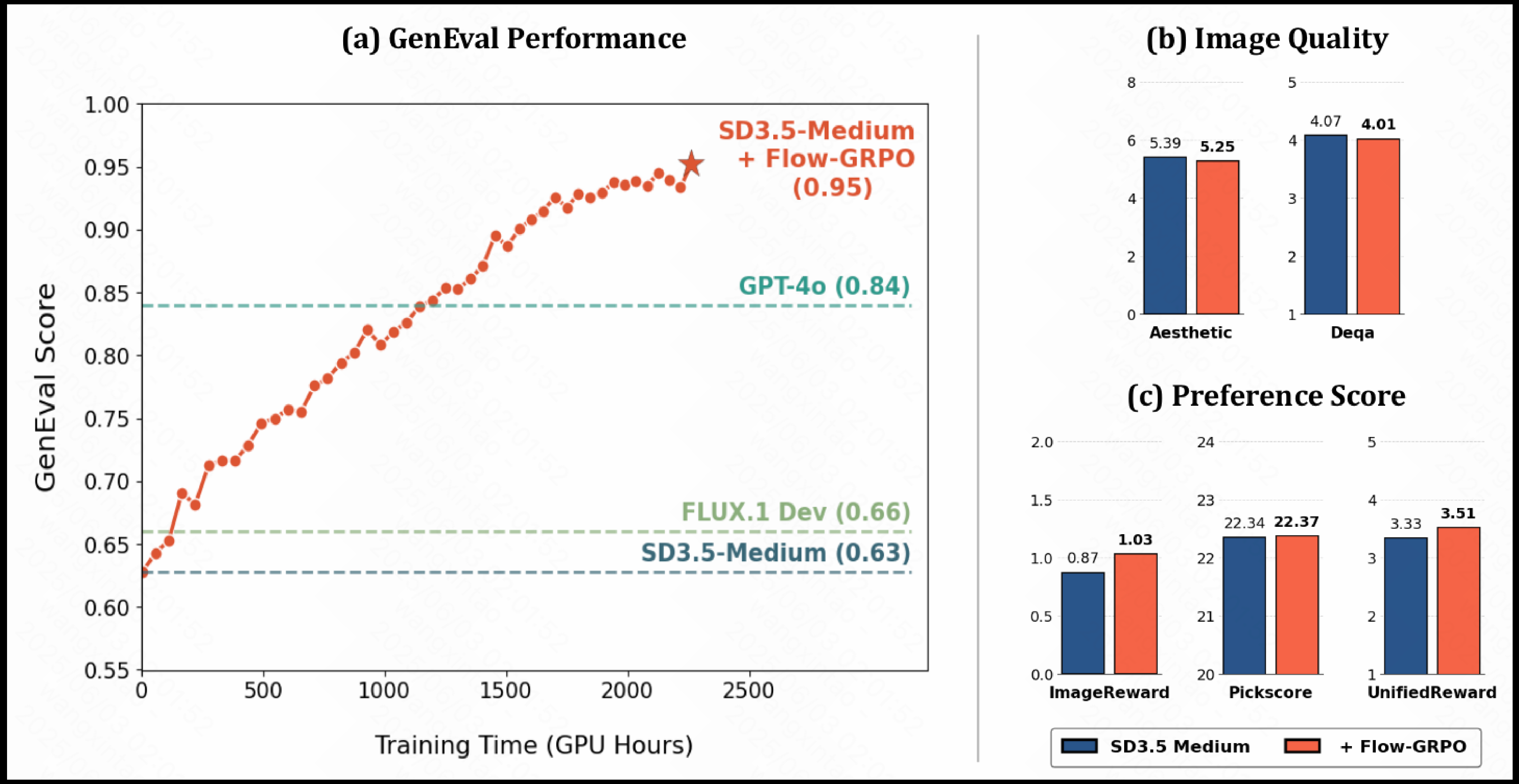
Figure 2: Overview of Flow-GRPO. Given a prompt set, we introduce an ODE-to-SDE strategy to enable stochastic sampling for online RL. With Denoising Reduction (only T = 10 steps), we efficiently gather low-quality but still informative trajectories. Rewards from these trajectories feed the GRPO loss, which updates the model online and yields an aligned policy.

- Given a prompt, we introduce an ODE-to-SDE strategy to enable stochastic sampling.
- With Denoising Reduction, we efficiently gather low-quality but still informative trajectories.
- Rewards from these trajectories feed the GRPO loss, which updates the model online and yields an aligned policy.

Online RL Algorithm for Visual Generation

- Online RL is highly effective for T2I tasks, for example, showing superior performance in Counting, Colors, Attribute Binding, and Position.





SD3.5 equipped with Flow-GRPO can surpass GPT-4o in GenEval (with minimal reward-hacking) !

04 Multimodal Perception and Reasoning

- Video Captioner and its Evaluation

  VidCapBench: A Comprehensive Benchmark of Video Captioning for Controllable Text-to-Video Generation

- Multimodality Bridge for Video Generation

  Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

Video Captioner and its Evaluation

😟 Performance of video generation relies heavily on the quality of video captions. But there lacks good evaluation benchmarks, which hinders the development of video captioners.

VidCapBench: A Comprehensive Benchmark of Video Captioning for Controllable Text-to-Video Generation

Xinlong Chen[1,2*], Yuanxing Zhang[3], Chongling Rao[3], Yushuo Guan[3], Jiaheng Liu[4], Fuzheng Zhang[3], Chengru Song[3], Qiang Liu[1,2†], Di Zhang[3], Tieniu Tan[1,2,4]
[1]New Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Kuaishou Technology [4]Nanjing University

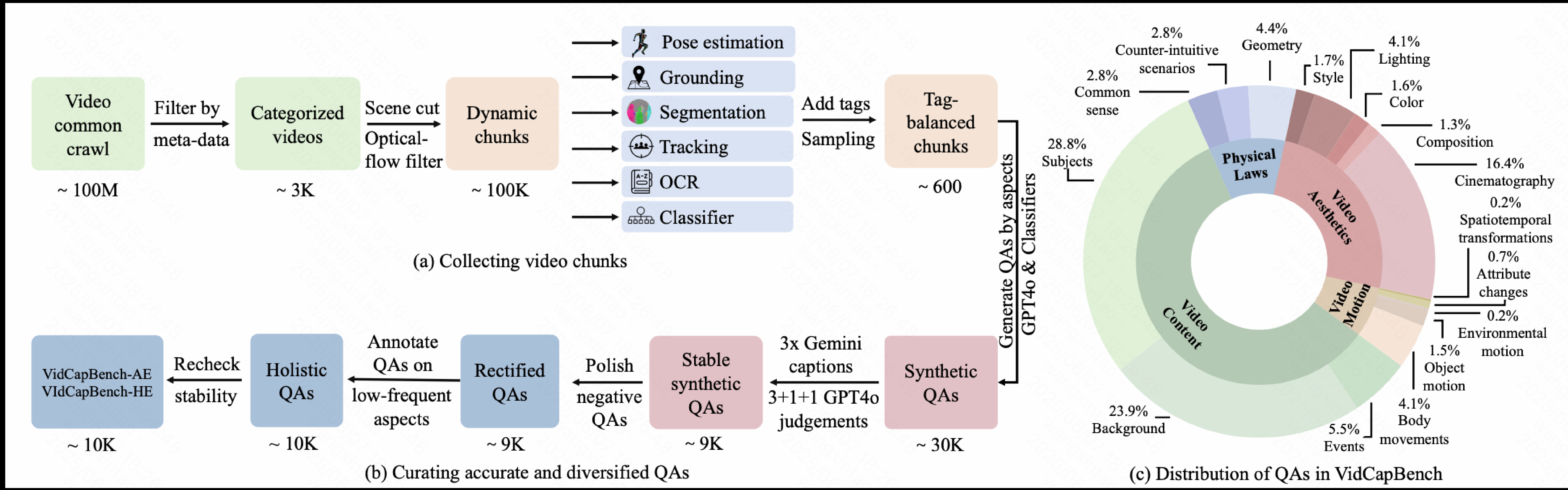
😃 We introduced a comprehensive evaluation framework for video captioning in T2V generation. Compared to existing benchmarks, VidCapBench exhibits greater stability and reliability, as well as strong correlation to the final T2V performance.

Video Captioner and its Evaluation

Multimodal Perception and Reasoning

KlingAI

- Implemented a data curation and annotation pipeline, which associates each video with key information in terms of aesthetics, content, motion, and physical laws.

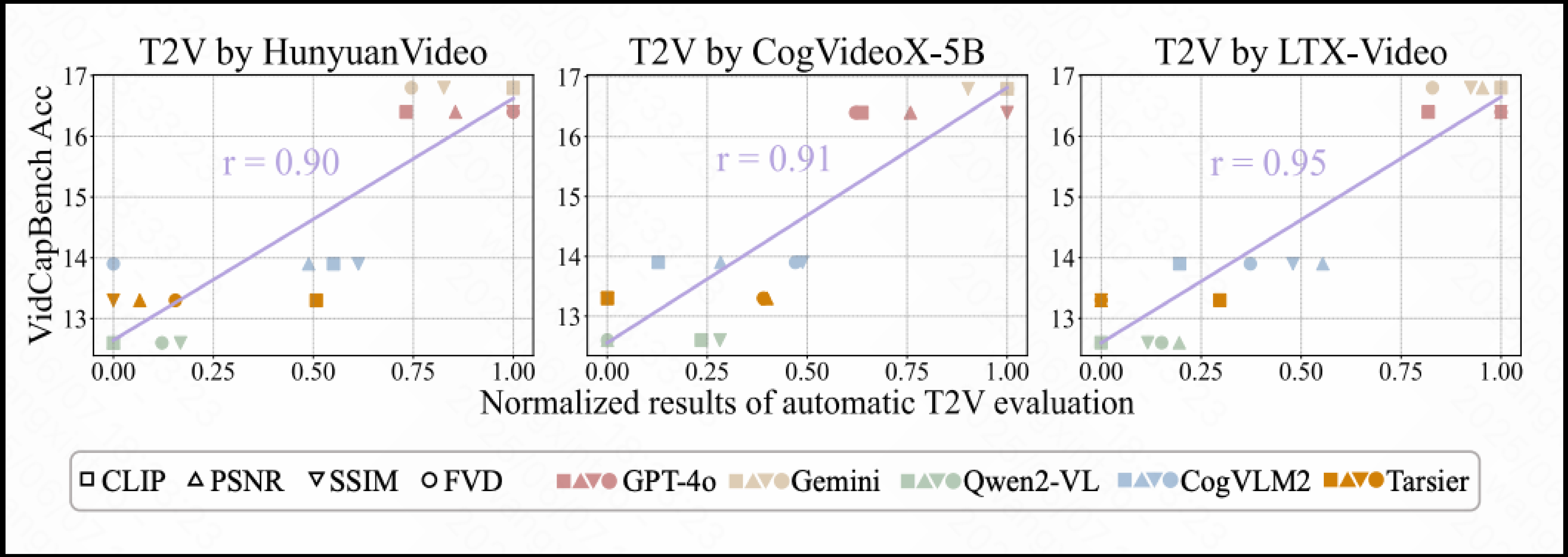


Benchmark	Metrics	# Videos	# QA pairs	Video diversity	Aesthetics	Subject	Motion	Physical law	Conciseness	Caption format
MSR-VTT (Xu et al., 2016)	CIDEr	2,990	2,990	✗	✗	✗	✗	✗	✔	Short
VATEX (Wang et al., 2019)	CIDEr	4,478	4,478	✗	✗	✗	✗	✗	✔	Short
DREAM-1K (Wang et al., 2024a)	Pre/Rec/F1	1,000	6,298	✔	✗	✔	✗	✗	✗	Unstructured
VDC (Chai et al., 2024)	Acc/VDCScore	1,027	96,902	✗	✗	✔	✔	✗	✗	Structured
VidCapBench	Acc/Pre/Cov/Con	643	10,644	✔	✔	✔	✔	✔	✔	Arbitrary

Table 1: Comparison between VidCapBench and mainstream video caption benchmarks.

Video Captioner and its Evaluation

- Experiments showed a strong positive correlation between the performance on VidCapBench and the quality of generated video. Good for T2V.



Correlations between T2V quality metrics and VidCapBench accuracy

Multimodality Bridge for Video Generation

😟 Video generation backbones have limited capacity for reasoning across different input modalities, resulting in suboptimal generation ability.

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

Shengqiong Wu[1,2*] Weicai Ye[1,✉] Jiahao Wang[1] Quande Liu[1] Xintao Wang[1] Pengfei Wan[1] Di Zhang[1] Kun Gai[1] Shuicheng Yan[2] Hao Fei[2,✉] Tat-Seng Chua[2]

[1]Kuaishou Technology [2]National University of Singapore

([*]Work done during internship at Kuaishou Technology. [✉]Correspondence)

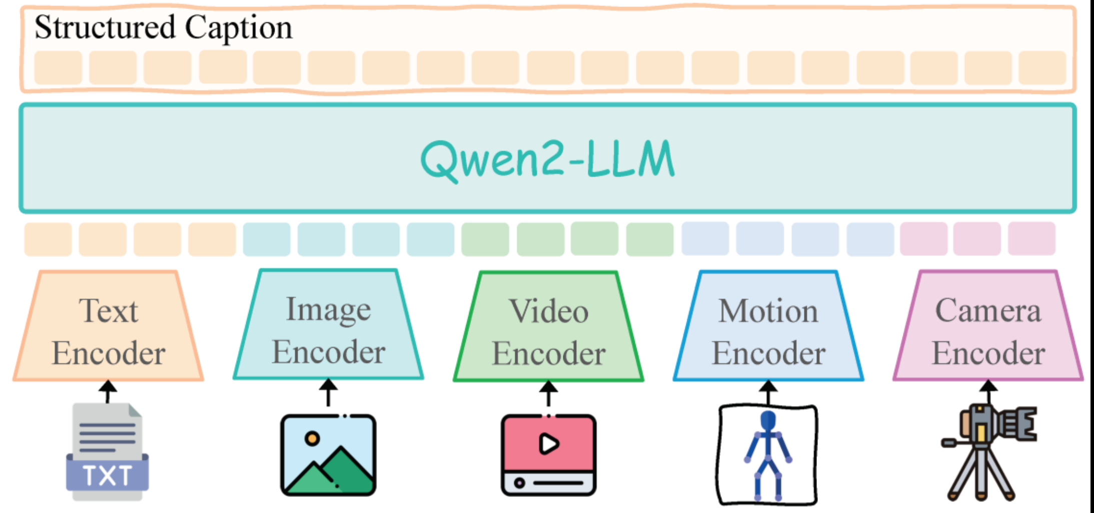
😃 We leveraged MLLMs to interpret diverse conditions into structured captions, decoupling the first job of interpreting conditions from the second job of video generation.

Multimodality Bridge for Video Generation

Multimodal Perception and Reasoning

KlingAI

- Any2Caption bridged the gap between user-provided multimodal inputs and structured video generation instructions. Simple yet effective.



(a) Overall architecture of Any2Caption

Compositional Condition	Text	Camera			Identities		Depth	Overall Quality			
	CLIP-T↑	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓	Smoothness↑	Dynamic↑	Aesthetic↑	Integrity↑
Camera+Identities	14.81	1.37	**4.04**	4.24	25.63	64.14	-	**94.43**	28.87	4.99	59.81
+ Structured Cap.	**19.03**	**1.30**	4.36	**4.03**	**26.75**	**68.45**	-	94.38	**34.99**	**5.25**	**63.02**
Camera+Depth	20.80	1.57	**3.88**	**4.77**	-	-	32.15	95.36	**30.12**	4.82	63.90
+ Structured Cap.	**21.19**	**1.49**	4.41	4.84	-	-	**25.37**	**95.40**	30.10	**4.96**	**65.05**
Depth+Identities	20.01	-	-	-	35.24	57.82	**23.00**	**93.15**	32.21	4.96	**61.21**
+ Structured Cap.	**20.76**	-	-	-	**36.25**	**63.48**	24.78	92.50	**36.43**	**5.18**	60.81
Camera+Identities+Depth	18.49	2.05	7.74	8.47	35.86	64.25	18.37	92.02	30.09	3.91	60.62
+ Structured Cap.	**19.52**	**1.57**	**7.74**	**8.20**	**38.74**	**64.37**	**17.41**	**93.03**	**32.81**	**4.99**	**61.22**

Table 6. Quantitative comparison of structured captions when handling compositional conditions. Better results are marked in **bold**.

Multimodality Bridge for Video Generation

Kling@CVPR2025

KlingAI

- Towards Precise Scaling Laws for Video Diffusion Transformers
- Koala-36M: A Large-scale Video Dataset Improving Consistency between Fine-grained Conditions and Video Content
- SketchVideo: Sketch-based Video Generation and Editing
- StyleMaster: Stylize Your Video with Artistic Generation and Translation
- GPAvatar: High-fidelity Head Avatars by Learning Efficient Gaussian Projections
- PatchVSR: Breaking Video Diffusion Resolution Limits with Patch-wise Video Super-Resolution
- Unleashing the Potential of Multi-modal Foundation Models and Video Diffusion for 4D Dynamic Physical Scene Simulation

Join Us!

Kuaishou Visual Generation and Interaction Center (aka the KLING Team), is committed to exploration and innovation for the cutting-edge technologies of multimedia content creation and interaction.

Career opportunities (internship or full-time job) are open. Feel free to contact us: kwaivgi@kuaishou.com

KlingAI

Thanks