QCon
全球软件开发大会

InfoQ 极客传媒

AniSora—动画视频生成技术应用

蒋宇东

B站智能创作技术负责人

目录

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

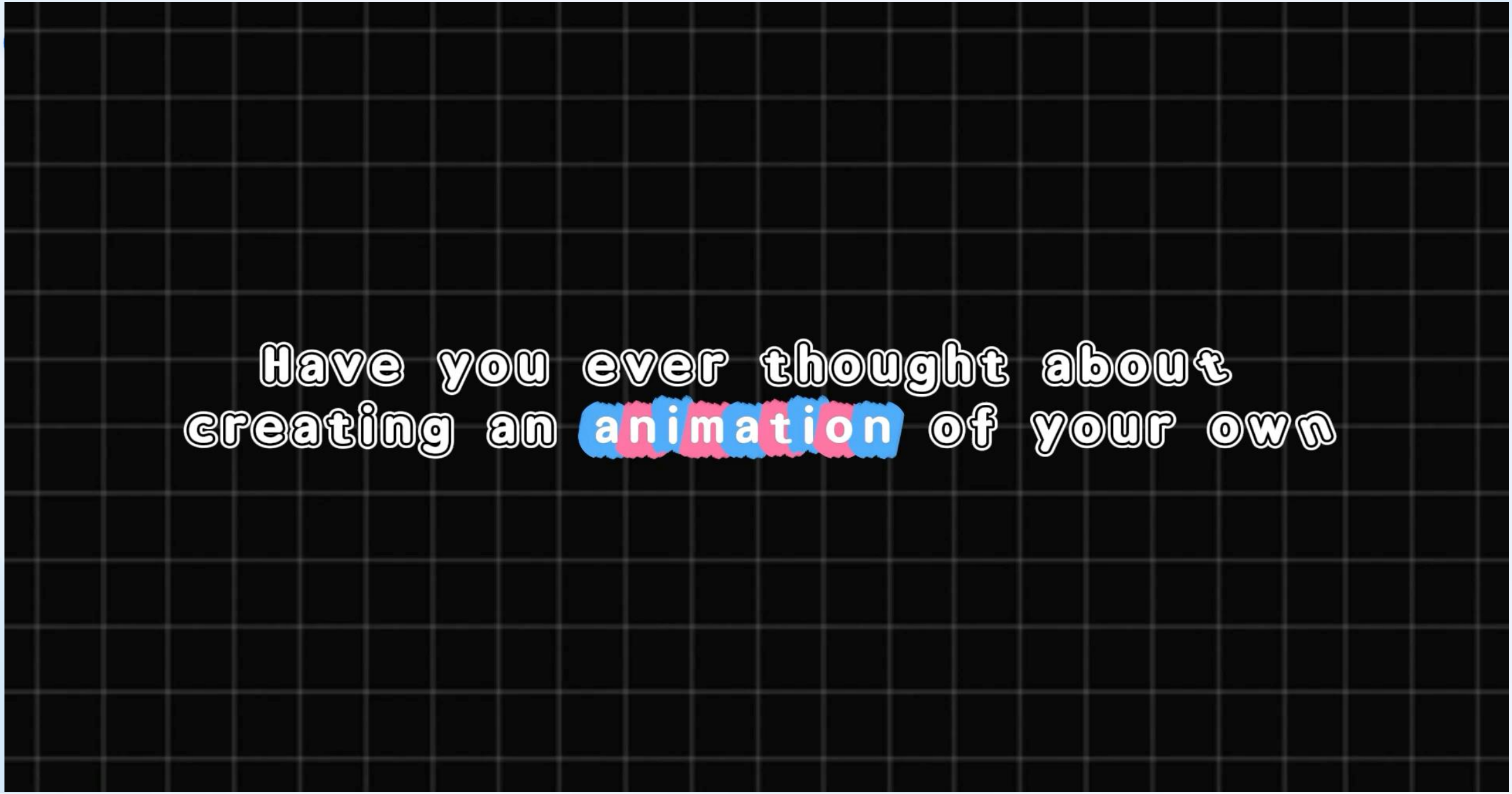
促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

01

动画视频生成的问题和挑战


Have you ever thought about
creating an animation of your own

视频生成技术发展背景

2024.02
- OpenAI Sora
- U-Net

2024.03
- Open-Sora 0.7B
- Open-Sora-Plan

2024.06
- runway Gen-3
- Kling 1.0

2024.07
- ZHIPU·AI 生数

2024.08
- MINIMAX Hailuo AI
- ByteDance TikTok PixelDance
- Seaweed
- 通义万相 Wanxiang
- LUMA AI D.M 1.6
- Kuaishou Kling 1.5
- CogVideoX-5B

2024.09
- AniSora 1.0

2024.10
- Meta Pika

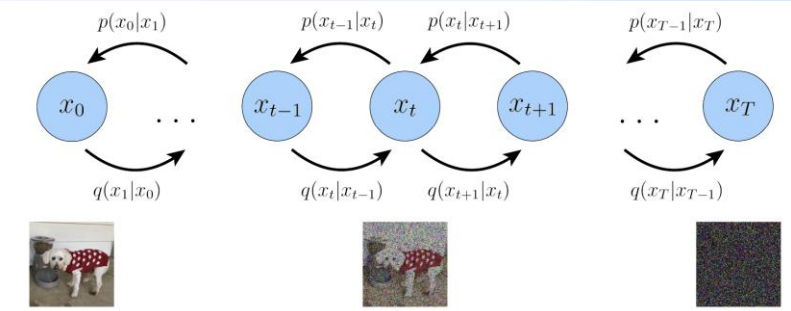
2024.12
- Sora Turbo
- Google Veo2
- Kling1.6
- Hunyuan-13B

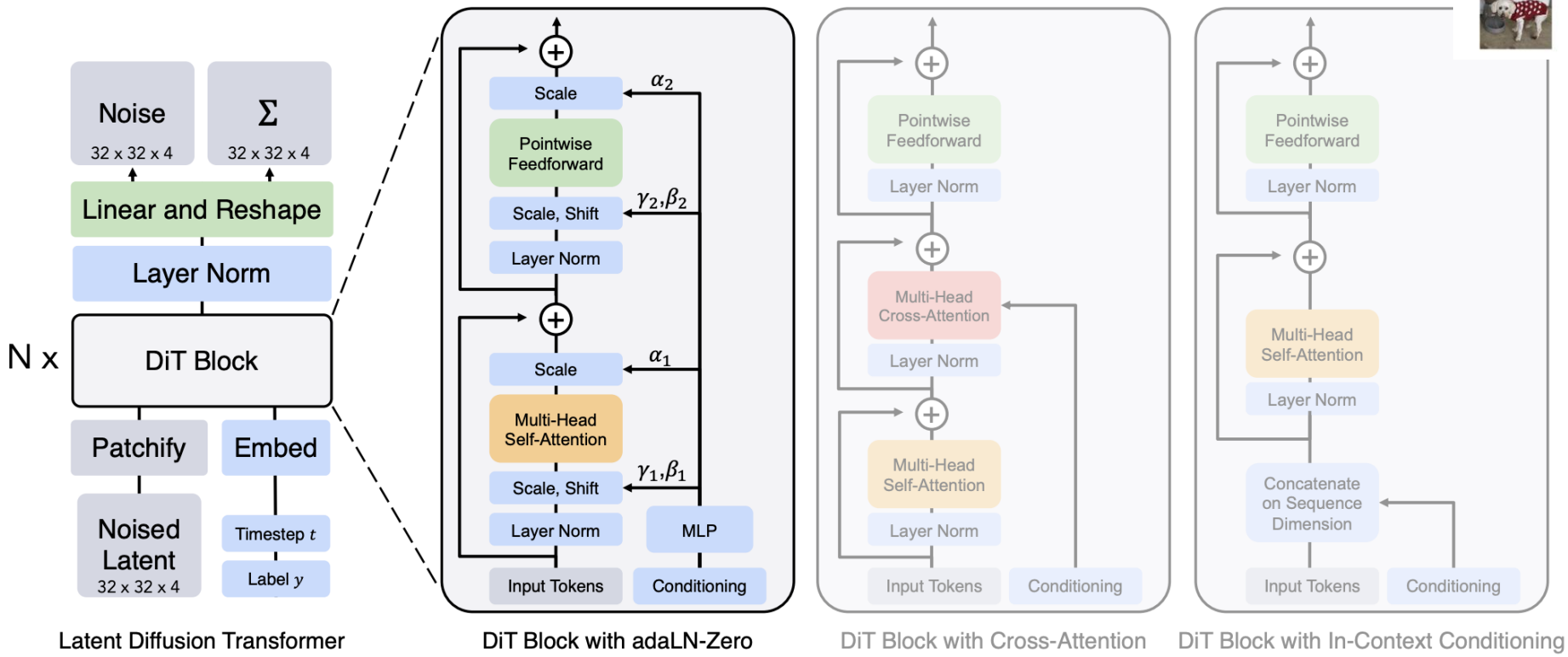
2025.02
- WanX2.1-14B
- AniSora 2.0

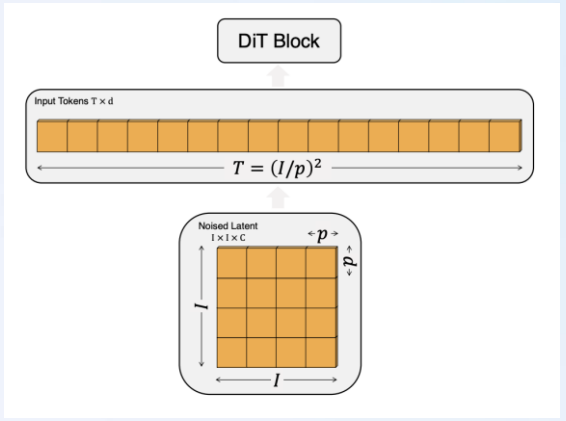
2025.07
- Jimeng2.0
- Kling2.0
- WanX2.2

2025.09
- AniSora 3.0

视频生成关键技术框架--DIT







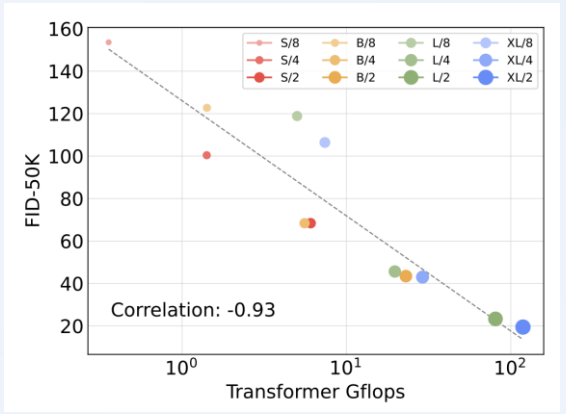
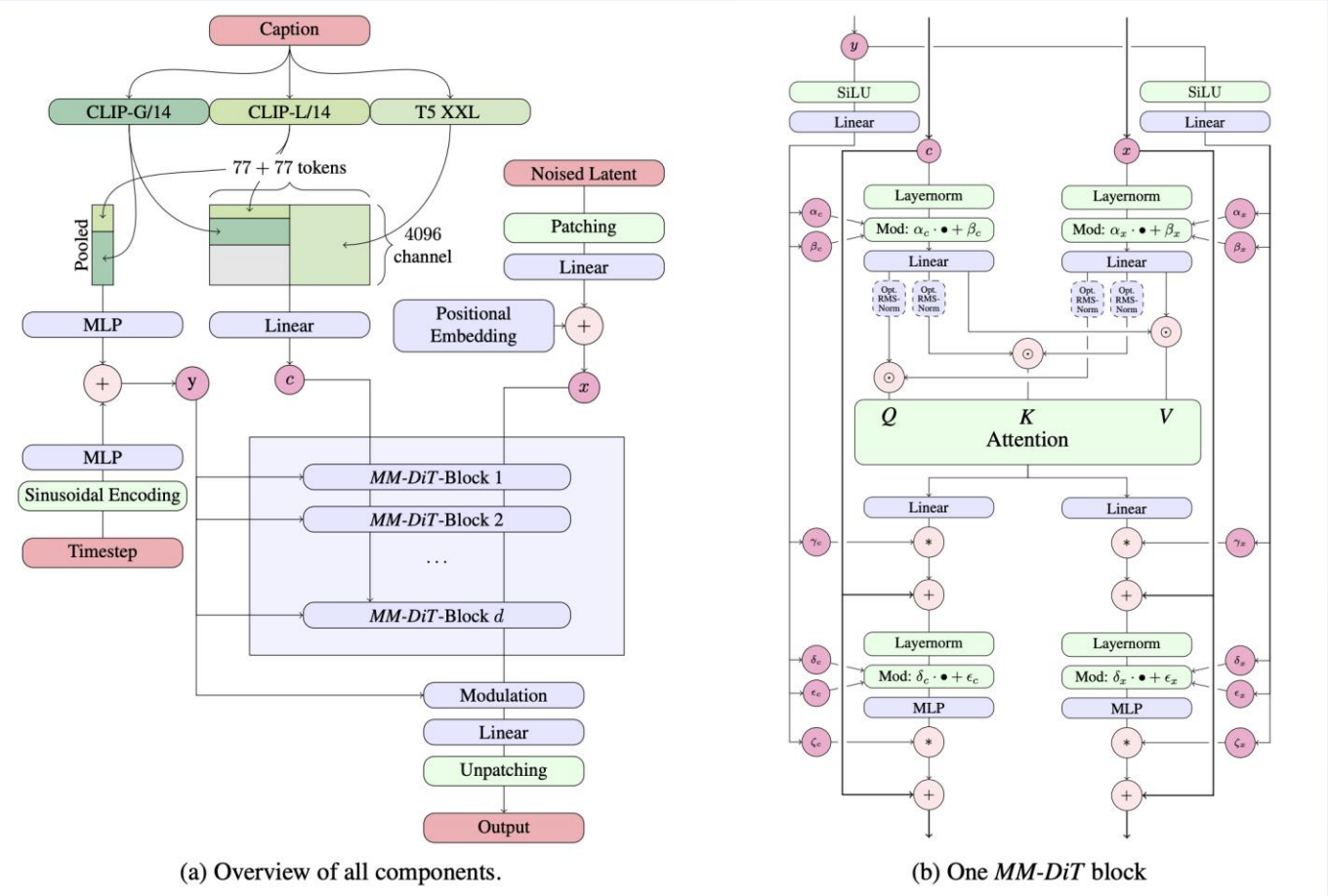


Figure 3. **The Diffusion Transformer (DiT) architecture.** *Left:* We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. *Right:* Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

视频生成关键技术框架--MMDIT



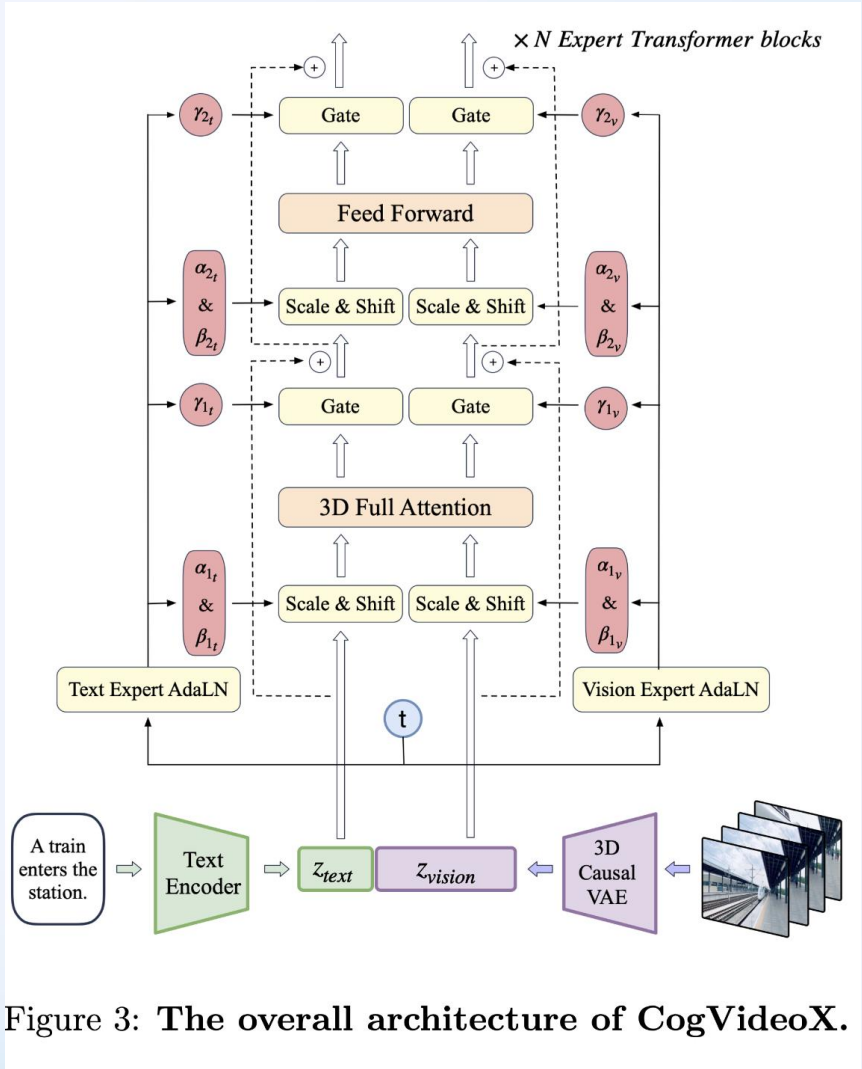


Figure 3: **The overall architecture of CogVideoX.**

动画视频生成的核心技术问题

- 多样的艺术风格
- 现实物理 Vs 动画物理
- 针对动画的benchmark构建

动画领域

- 人物一致性
- 场景一致性

一致性

- 多模态引导
- 控制动态大小、运镜、运动区域等

可控性

- 质量*信息密度*时长
- AI Director

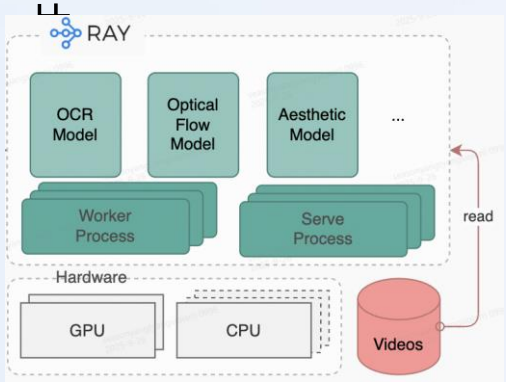
长视频生成

02

Anisora技术架构

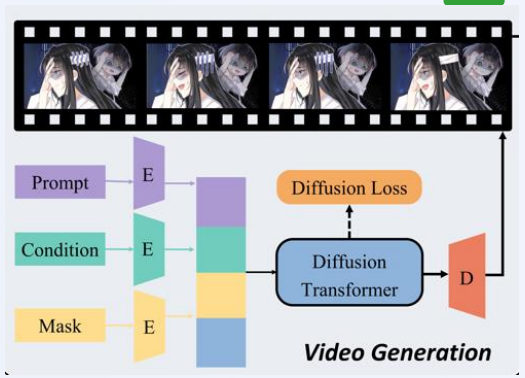
Anisora技术架构概览

10M高质量动画切



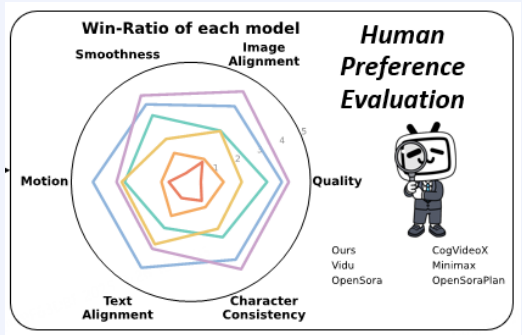
数据集构建

多控制合一训练



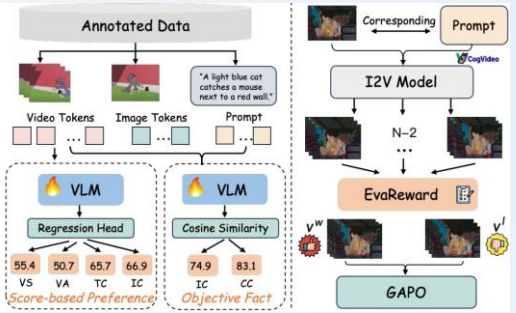
模型训练

6个维度立体评估



Benchmark构建

强化学习后训练调优



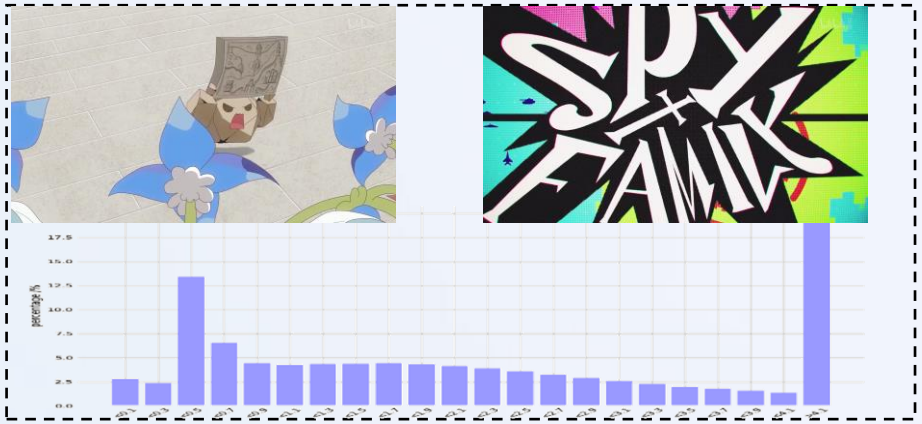
强化学习调优

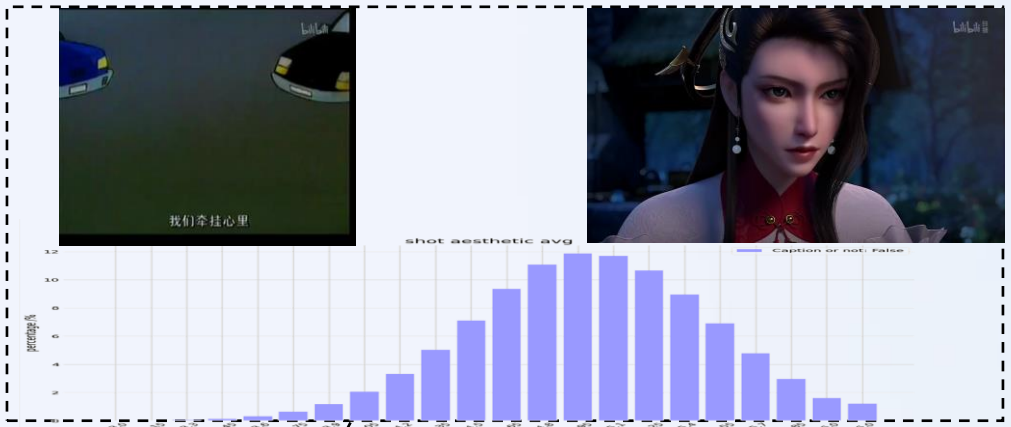
AniSora: Exploring the Frontiers of Animation Video Generation in the Sora Era. Yudong Jiang, Baohan Xu et. al, ***Accepted by IJCAI25***

Aligning Anime Video Generation with Human Feedback. Yidi Wu, Bingwen Zhu et. al Under Review

GitHub - bilibili/Index-anisora SOTA Animation Video Generator, 2.1K Star

Data Pipeline





Raw Clips

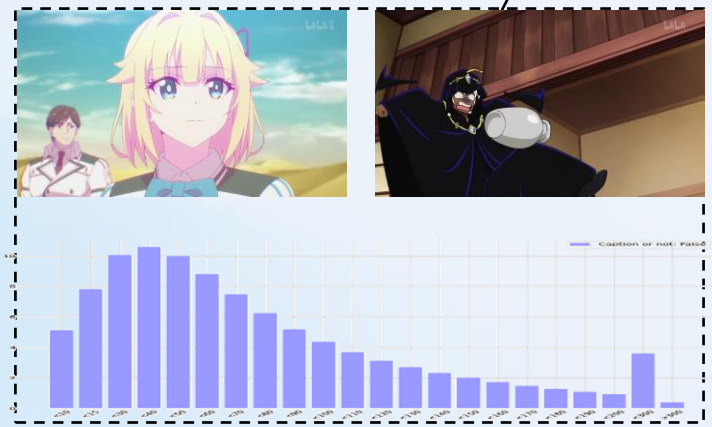
Optical Flow

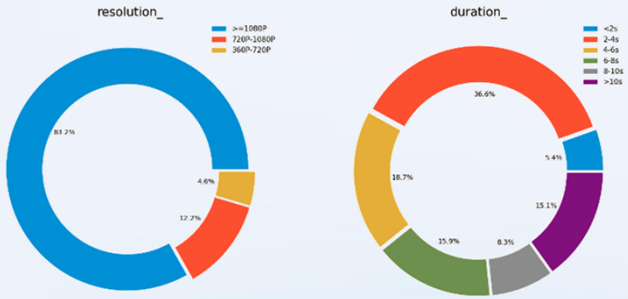
OCR

Aesthetics

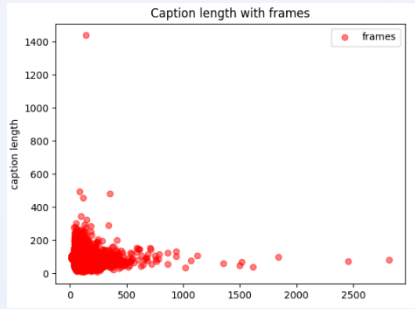
Caption

AniSora DataSet





Distribution of Resolution and Duration



Caption Length and Frames

Dataset & Benchmark

Training Set

Raw Data: 1M long animation videos

Filtering Criteria:

Text-overlay, optical flow, aesthetic scores, duration (2-20s)

Resulting Dataset:

10M high-quality clips for training

Evaluation Set

948 animation clips are collected and labeled with different actions

- *100 common action labels: talking, walking, running, eating, and so on. 10-30 clips/label*
- *857 2D animation, 91 3D animation*
- *Human manual correct prompt*

Point at things

Walk

Punch

Wave hand

Rub

Brace

- ***Visual Smoothness***

$$S_{smooth} = Reg(E_v(I_1, \ldots, I_N))$$

where I_i, N denote the single frame and the total frames, and Reg denotes the regression head, and E_v denotes the vision encoder.

- ***Visual Motion***

The magnitude of primary motion in anime videos.

$$S_{motion} = Cos(MCLIP(V), MCLIP(T_m))$$

where $MCLIP$ denotes the finetuning action model. V represents the generation video and T_m denotes the designed motion prompt.

- ***Visual Appeal***

The fundamental quality of video generation.

$$S_{appeal} = Aes(SigLIP(I_{0,1,\ldots,K}))I_i \in KeyFrm(V)$$

where$KeyFrm$, $SigLIP$ and Aes denote the key frame extraction method, feature encoder method and aesthetic evaluation method, and K denotes the number of the keyframes.

- ***Text-Video Consistency***

$$S_{tvc} = Reg(E_v(V), E_t(T))$$

where Reg denotes the regression head, and E_v, E_t denote the vision and text encoder, respectively.

- ***Image-Video Consistency***

$$S_{ivc} = Reg(E_v(V), E_v(I_p))$$

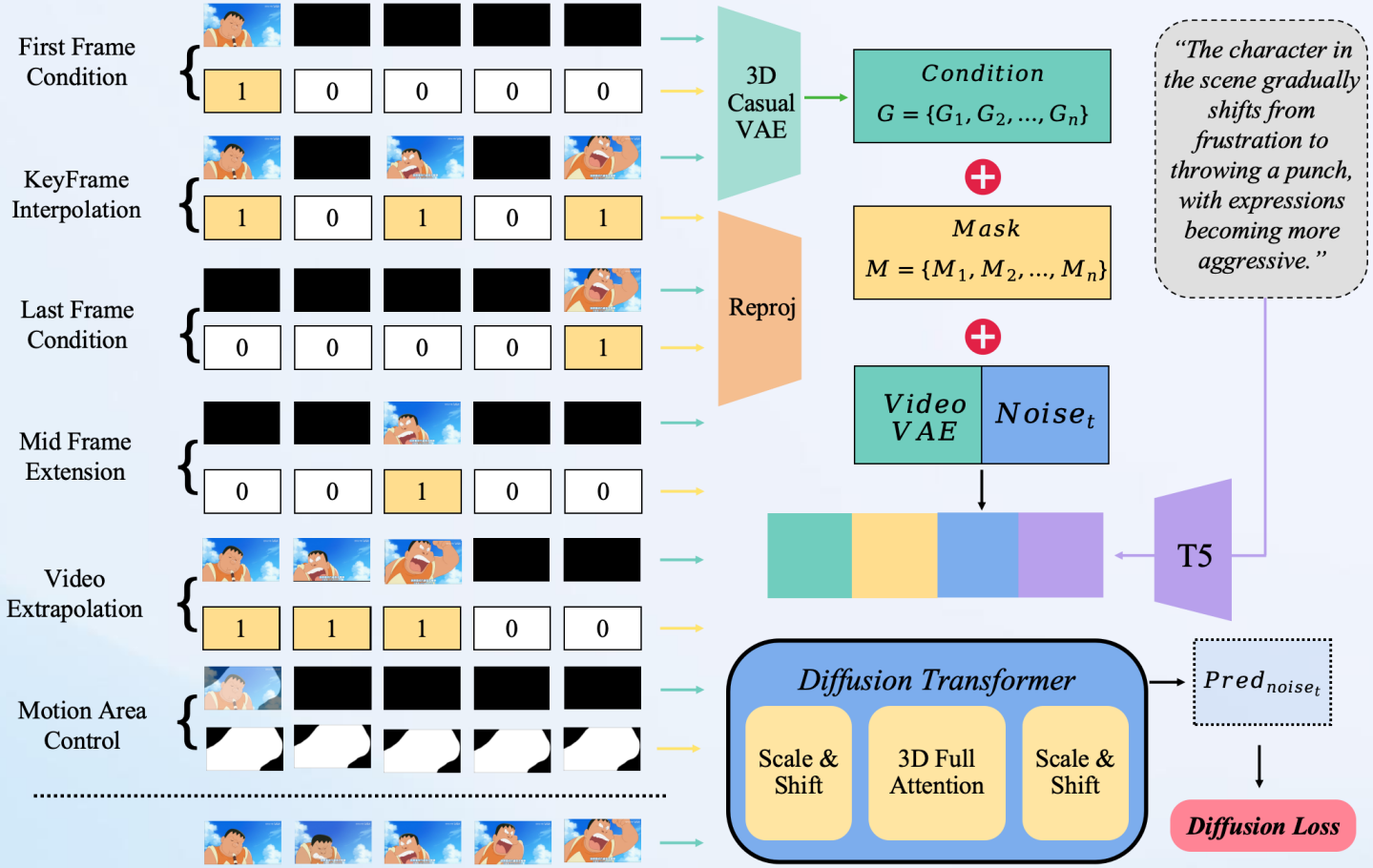
where $V and I_p$ denote the participant video clip and the input image.

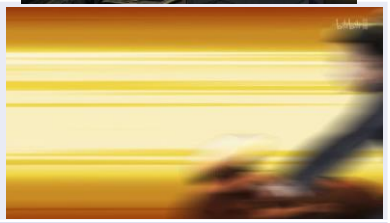
- ***Character Consistency***

$$S_{IPc} = \frac{1}{S}\sum_{i}^{S} Cos(BLIP(M_i), fea_c)$$

where S denotes the number of sample frames, M_i denotes the mask obtained from GroudingDino and SAM methods, and fea_c denotes the stored character's features.

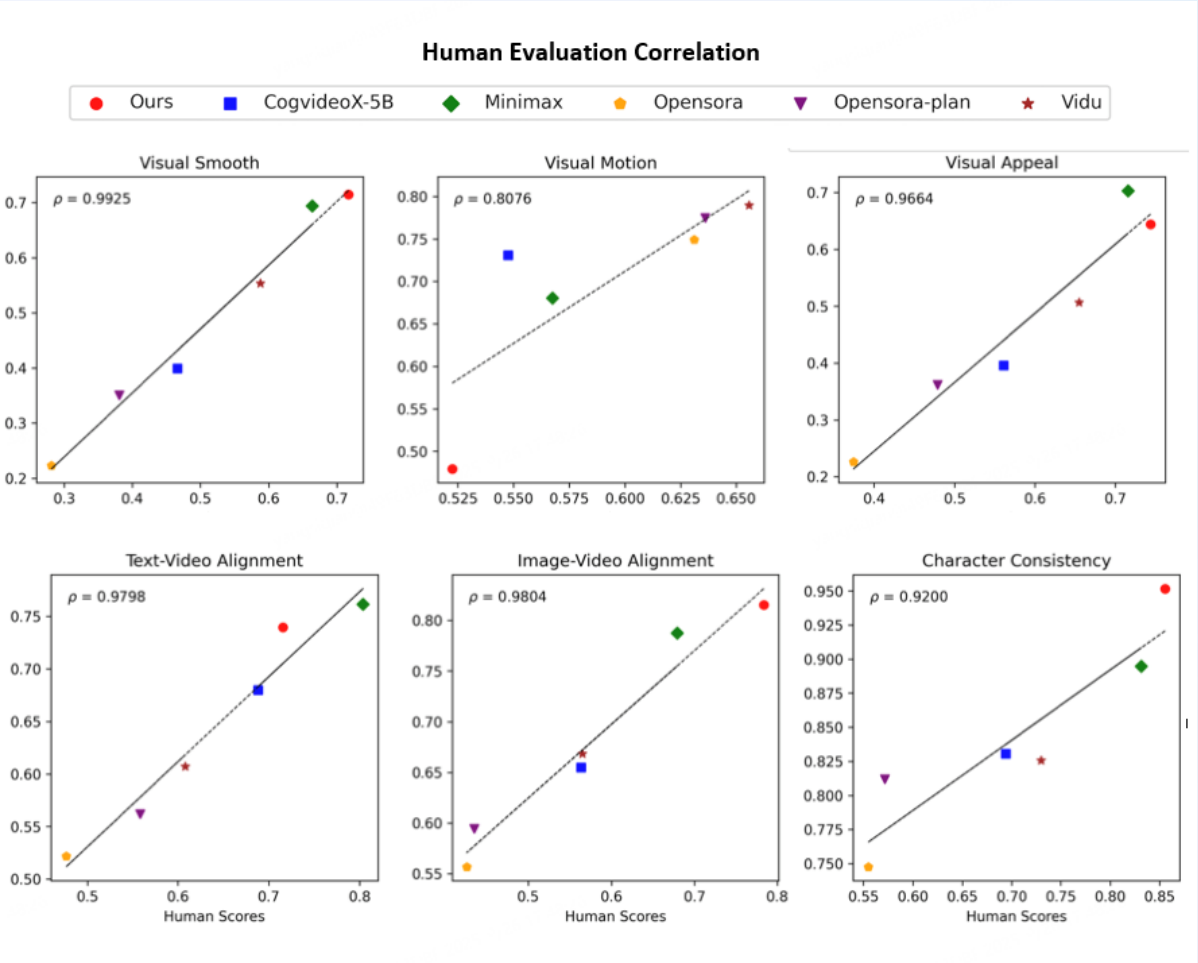
模型训练



引导图	位置	生成效果
	首帧	
	中间帧	
	尾帧	
		

AniSora Benchmark

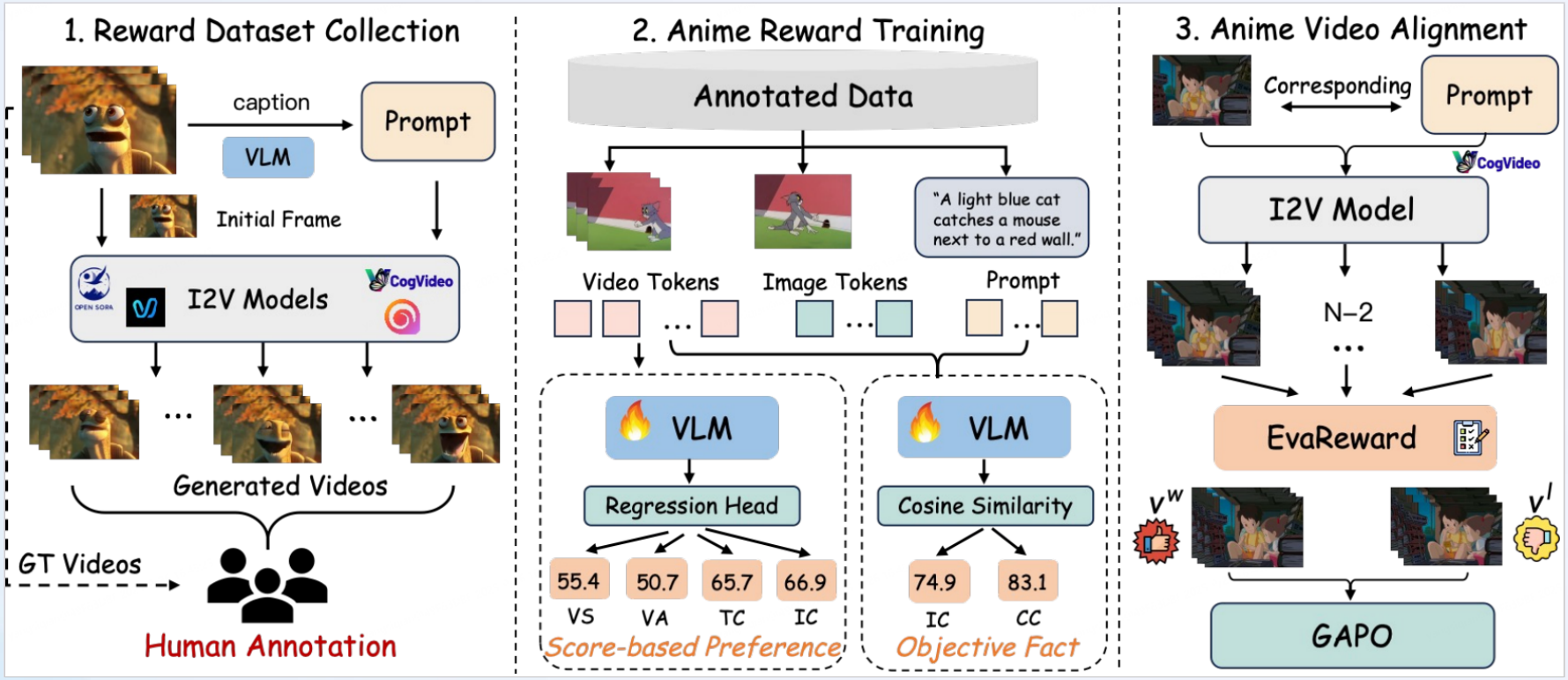
Method	Human Evaluation	Visual Smooth	Visual Motion	Visual Appeal	Text-Video Consistency	Image-Video Consistency	Character Consistency
Vidu-1.5	60.98	55.37	**78.95**	50.68	60.71	66.85	82.57
Opensora-V1.2	41.1	22.28	74.9	22.62	52.19	55.67	74.76
Opensora-Plan-V1.3	46.14	35.08	77.47	36.14	56.19	59.42	81.19
CogVideoX-5B-V1	53.29	39.91	73.07	39.59	67.98	65.49	83.07
MiniMax-I2V01	69.63	69.38	68.05	70.34	76.14	78.74	89.47
AniSora (Ours)	70.13	71.47	47.94	64.44	72.92	81.54	**94.54**
AniSora (Interpolated Avg)	-	70.78	53.02	64.41	73.56	80.62	91.59
AniSora (KeyFrame Interp)	-	70.03	58.1	64.57	74.57	80.78	91.98
AniSora (KeyFrame Interp)	-	70.03	58.1	64.57	74.57	80.78	91.98
Wan2.1	-	81.70	61.88	82.05	87.81	88.50	90.65
AnisoraV2	-	86.98	50.34	85.91	90.98	91.96	92.75
AniSoraV3	-	**90.00**	62.78	**87.84**	**91.24**	**92.53**	92.58
GT	-	92.2	58.27	89.72	92.51	94.69	95.08



AniSora-Benchmark

引导指令	Keling1.6	Hailuo AI	ViduQ1	AniSoraV2.0
 画面中一个人在快速向前奔跑，他奔跑的速度很快使得人物有些模糊				
 画面中黄色的怪物，突然站起来，他的右手指向屏幕左手握拳，愤怒的说些什么				

强化学习调优



$$p(v^w \succ v^l|c) = \sigma(R(c, v^w) - R(c, v^l)),$$

$\mathcal{L}_{\text{DPO}}(\theta)$:

$$-\mathbb{E}_{(c,v^w,v^l)}\left[\log\sigma\left(\beta\log\frac{p_\theta(v^w \mid c)}{p_{\text{ref}}(v^w \mid c)} - \beta\log\frac{p_\theta(v^l \mid c)}{p_{\text{ref}}(v^l \mid c)}\right)\right],$$

$$\mathcal{L}_{\text{GAPO}}(\theta) = (G_w - G_l)\cdot\mathcal{L}_{\text{DPO}}(x, c, v^w, v^l),$$

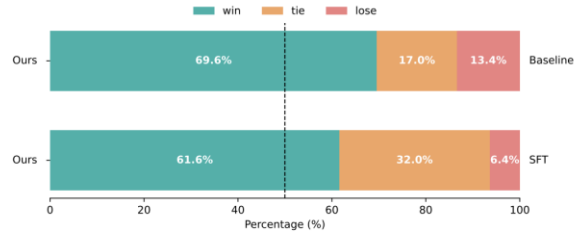
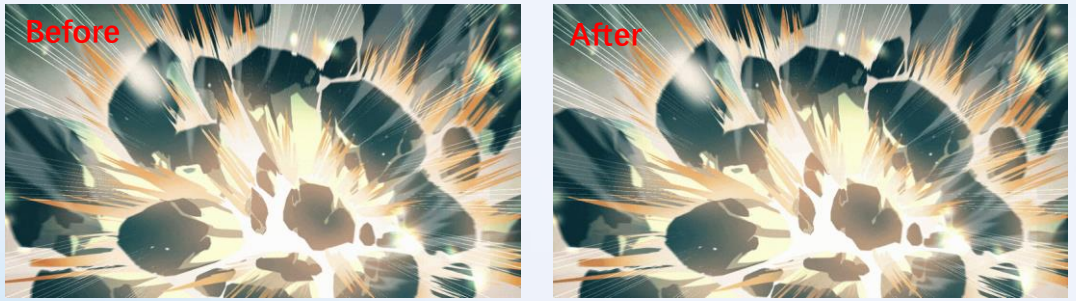


Figure 4. Human evaluation of generated anime videos on different models.

Method	AnimeReward						VideoScore				
	Visual Smooth	Visual Motion	Visual Appeal	Text Consist.	Image Consist.	Charac. Consist.	Visual Quality	Temporal Consist.	Dynamic Degree	Text Alignment	Factual Consist.
Baseline	52.63	65.13	47.68	71.50	74.92	88.85	3.361	3.080	**3.334**	2.945	3.114
SFT	52.62	**66.27**	48.11	72.81	75.25	90.38	3.358	3.075	3.329	**2.951**	3.100
Ours	**61.67**	59.86	**53.08**	**73.29**	**79.30**	**91.45**	**3.379**	**3.131**	3.290	2.942	**3.148**

Table 2. **Quantitive performance comparison on AnimeReward and VideoScore** [5], with best performance in **bold**. We achieve the best performance in both visual quality and motion smoothness, with a higher degree of distinction on AnimeReward.

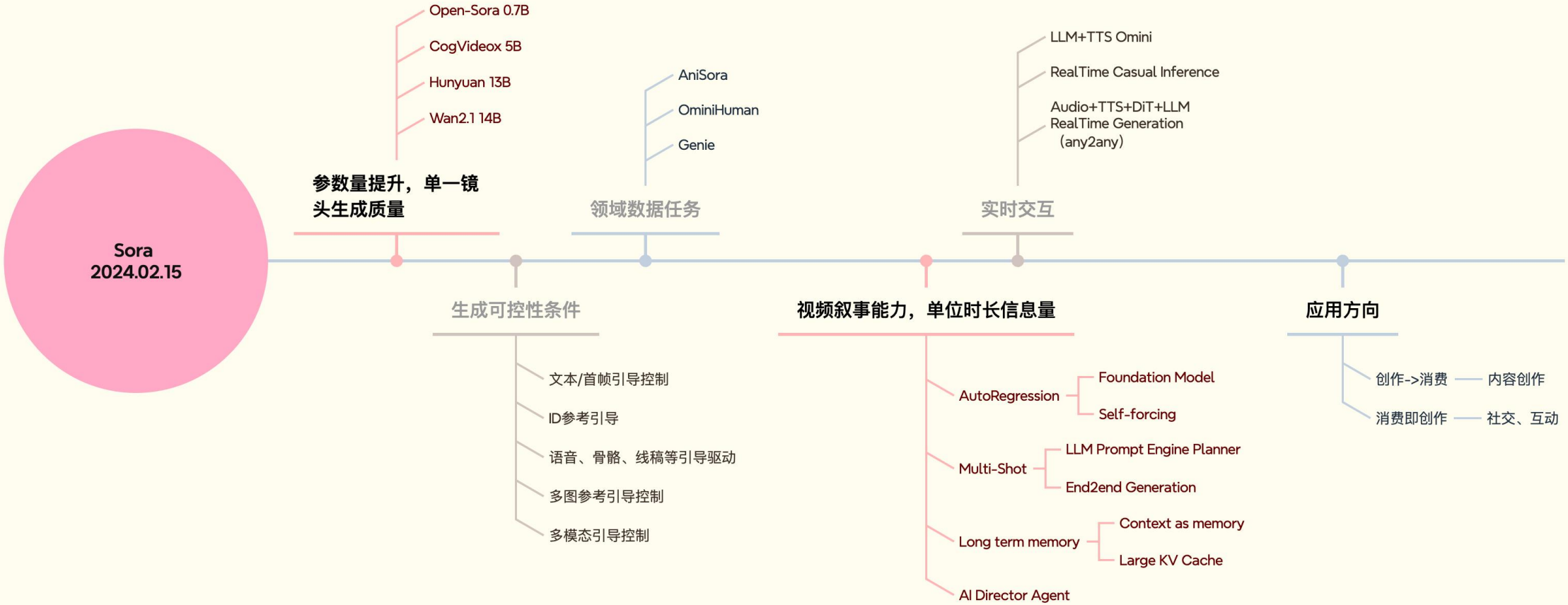


03

技术应用

技术落地的挑战和解决方案

技术到应用的演进


Sora
2024.02.15
参数量提升，单一镜头生成质量
Open-Sora 0.7B
CogVideox 5B
Hunyuan 13B
Wan2.1 14B
领域数据任务
AniSora
OminiHuman
Genie
实时交互
LLM+TTS Omini
RealTime Casual Inference
Audio+TTS+DiT+LLM RealTime Generation (any2any)
生成可控性条件
文本/首帧引导控制
ID参考引导
语音、骨骼、线稿等引导驱动
多图参考引导控制
多模态引导控制
视频叙事能力，单位时长信息量
AutoRegression
Foundation Model
Self-forcing
Multi-Shot
LLM Prompt Engine Planner
End2end Generation
Long term memory
Context as memory
Large KV Cache
AI Director Agent
应用方向
创作->消费
内容创作
消费即创作
社交、互动

技术应用——长视频制作相关落地挑战

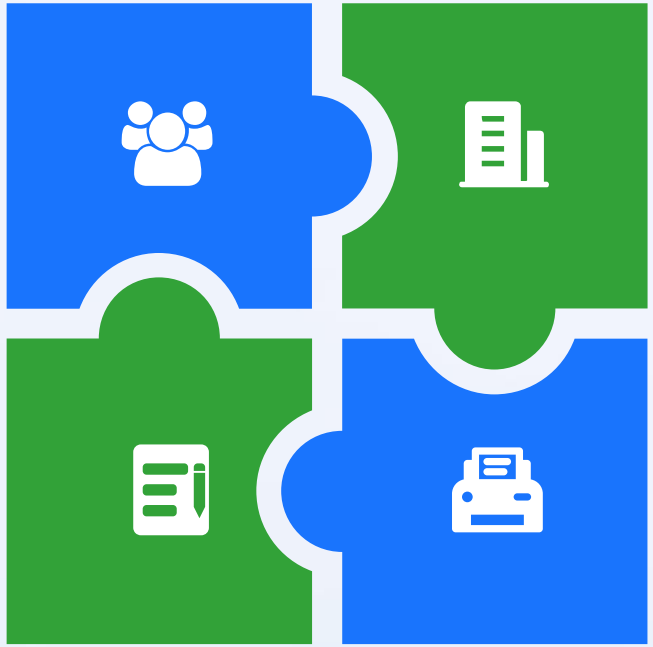
AI长视频制作的问题

人物、场景一致性问题

多主体的一致性保持问题
多视角下的人物全身一致性
跨镜头下的场景一致性

细节控制和易用性的权衡问题

用户友好的长视频生成控制条件
细粒度视频内容控制方法
多控制条件统一的控制方法



多模态交互问题

交互类型的支持方式
生成式交互的实时性
多模态统一建模方法

物理效果生成问题

“动画物理学”3D Vs 2D
数据驱动模式


AniSora V3 Preview

人物、场景一致性问题

多主体的一致性保持

存在挑战：

- 多主体关系的正确指代
- 主体的大小比例保持合理
- 参考主体数量多时，计算开销大、主体细节缺失问题

解决思路：

- 将视频历史帧作为视觉参考，通过视频基模统一视频和图像参考生成任务
- 通过context的方式提供视觉特征参考
- 向prompt中注入图像id，关联主体-提示词
- 参考生图模型生成关键帧+图生视频 / 参考主体生视频

人物、场景一致性问题

多视角下人物全身一致性

相同背景图在切镜后的几何比例保持

跨镜头下的场景一致性

存在挑战：

- 单一视角参考图的模式下，难以保证新视角的主体细节一致
- multiview参考模式下，如何正确预测主体的多视角图像
- 如何给定视频/3d模型，参考模型如何选择正确的视角进行参考

解决思路：

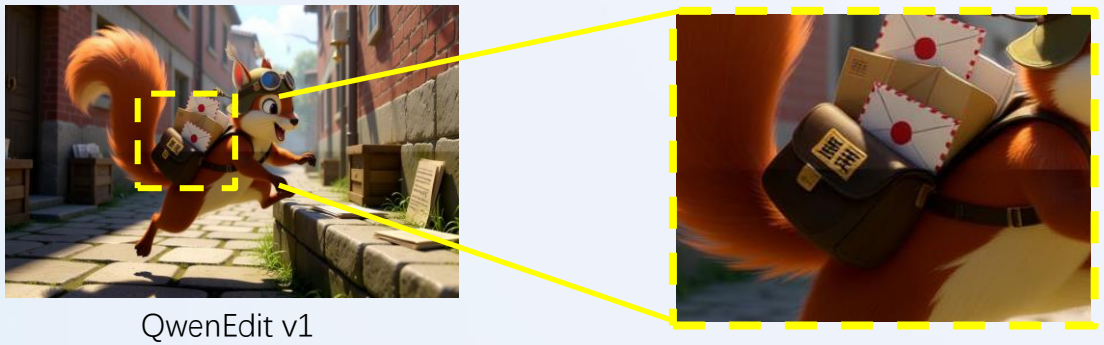
- 先进行图像的multiview预测，然后使用多视角参考图
- 使用图生视频/3D模型生成多角度视频，再用于多视角参考

视频生成构建角色/场景多视角参考

通过Token Selection实现高效一致性参考

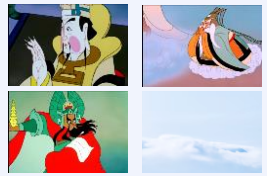
仿照视频记忆参考的思路，选择适合的参考Token，减少计算消耗同时增强参考细节

输入参考图

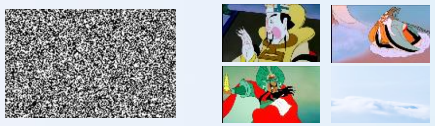


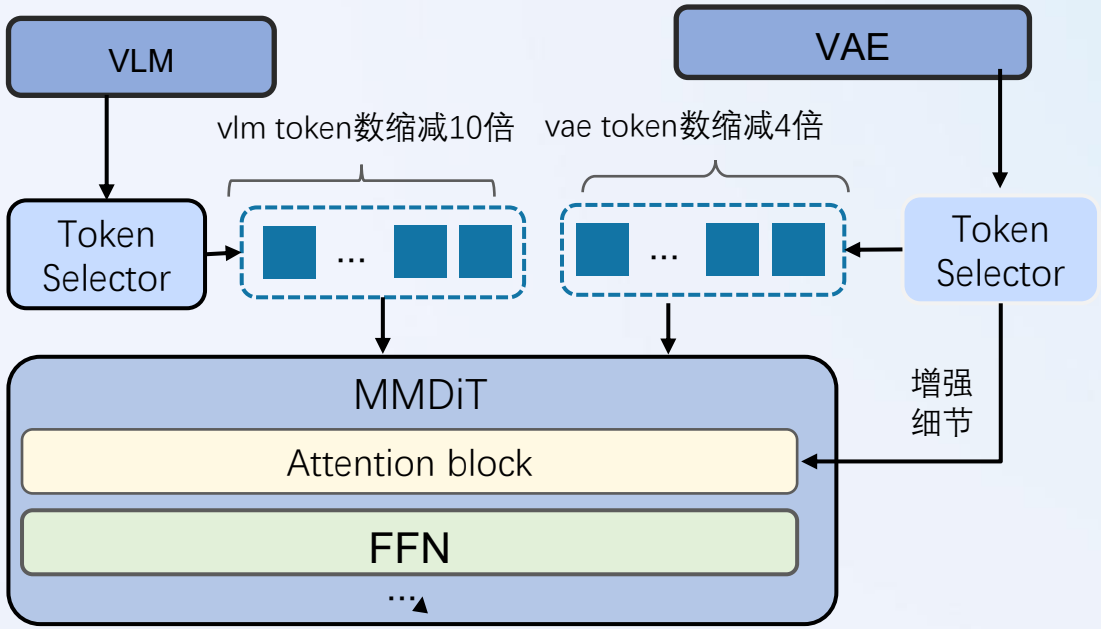
QwenEdit v1

TokenSelect参考



authority sits on a throne, exuding authority, while two other characters ..





采样输出

细节控制和易用性的权衡问题

骨架图序列控制动作

目标框控制位置

用户友好的视频生成控制条件

存在挑战：

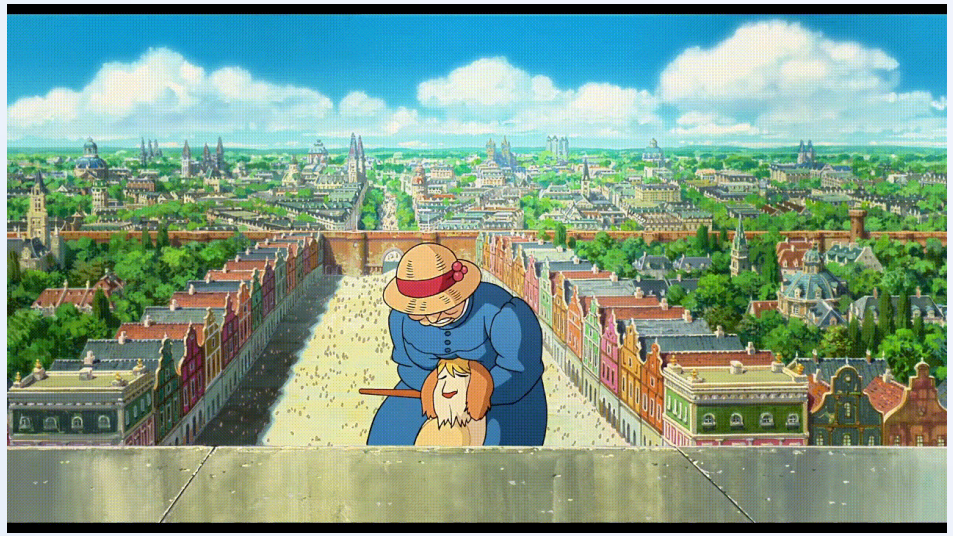
- 进行逐帧控制时，控制信号的构造成本高，高难度动作难以准确生成
- 面部/演员替换方法适用于二创，难以用于完全原创视频
- 难以对视频内容进行像素级精准控制
- 多种控制条件同时启用的兼容问题

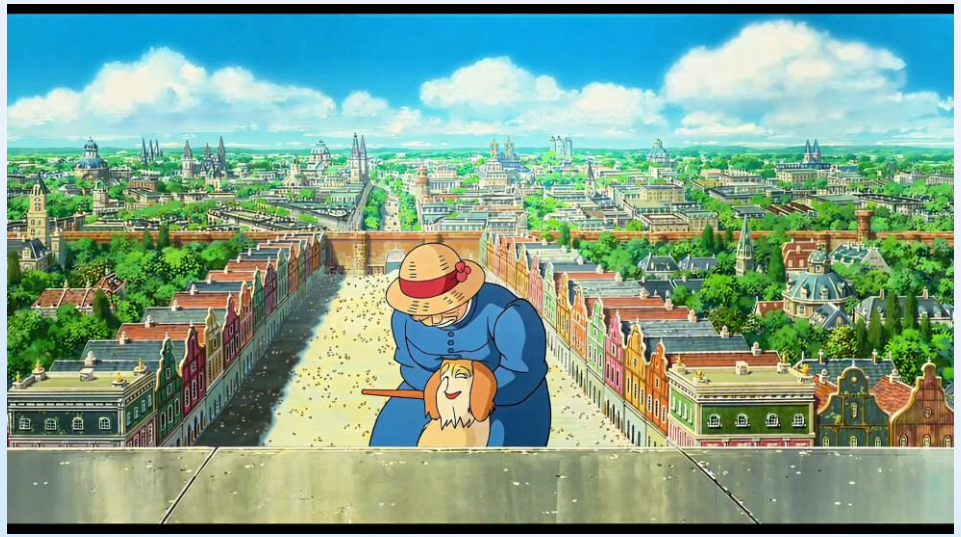
解决思路：

- 相机轨迹及位姿作为运镜控制
- 与帧对齐的布局/骨架图/参考视频控制主体运动
- 多模态条件生成模型
- 建立从粗到细的条件控制数据金字塔，迭代式训练控制条件，提供多种控制条件并重的可控生成模型

视频质量提升

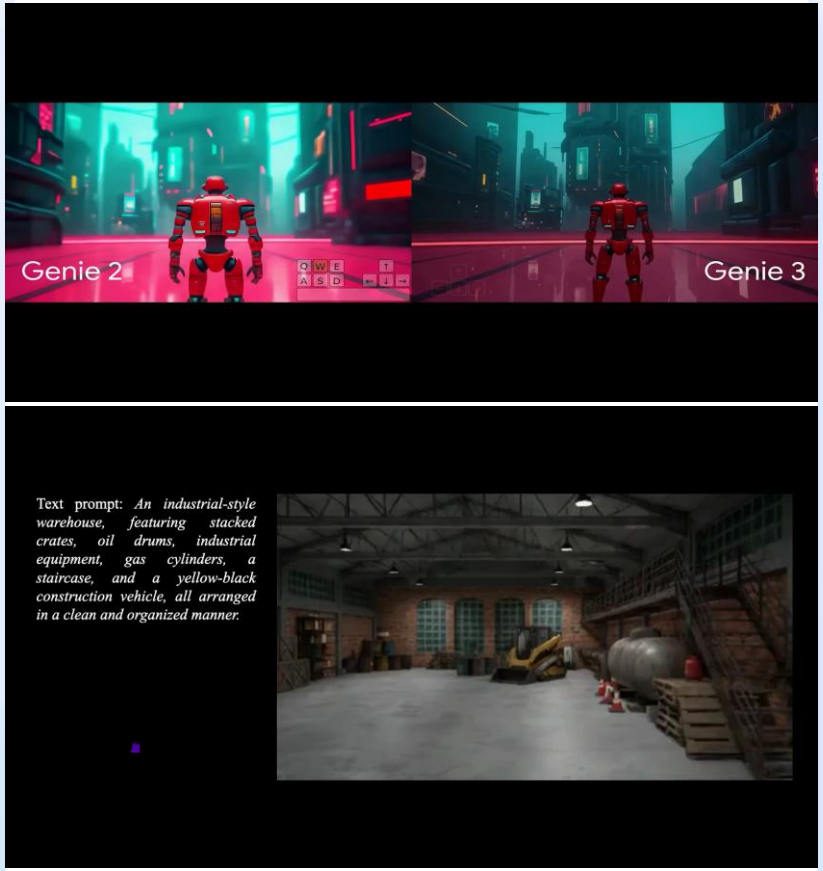
90P -> 1080P





https://github.com/bilibili/Index-anisora （Coming Soon）

多模态交互问题



镜头控制交互的视频生成

多模态交互问题

存在挑战：

- 高分辨率的**视频**仍然**难**以达到**实时**生成的速度
- 真**实**世界**视频**的缺少交互信号，**导**致写**实风**格的交互模型**难**以**训练**
- 交互模型**对**多模**态**信号的理解+生成**统**一

解决思路：

- 将**视频**模型改造**为**自回**归**生成模式，**缩**短生成**时长**，降低响**应**延**迟**
- 通**过**游**戏**引擎采集鼠**标**、**键盘**操作数据，**训练**可控**视频**生成模型
- 将**镜头**作**为驱动**条件，**训练视**角可**动**的交互式**视频**模型

素材来源：Gene3,Context As memory

物理效果生成问题分析和解决思路

动态/静态世界规律学习

存在挑战：

- 动画中的运动规律（特别是2D动画）不同于真实世界，但彼此之间又有很多的联系，我们叫它“动画物理学”
- 数据驱动方法容易生成“视觉合理”但不满足物理规律的结果
- 在训练分布外的物理环境中（新材质、新约束条件）表现较差
- 高质量物理交互数据昂贵、难以规模化采集，缺乏真实世界多模态数据（力、速度、材质属性等）

解决思路：

- Data driven
- 几何先验引入（如3D预训练模型VGGT）到视频生成模型中
- 基于AR模式（如下一帧预测）的训练范式从时序上建模运动模式

素材来源：VideoREPA

04

动画视频智能创作的未来

长视频创作的核心问题

视频质量 * 信息密度 * 视频长度

空镜头 = **视频质量** * 信息密度 * **视频长度**

Sora2 = 视频质量 * **信息密度** * 视频长度

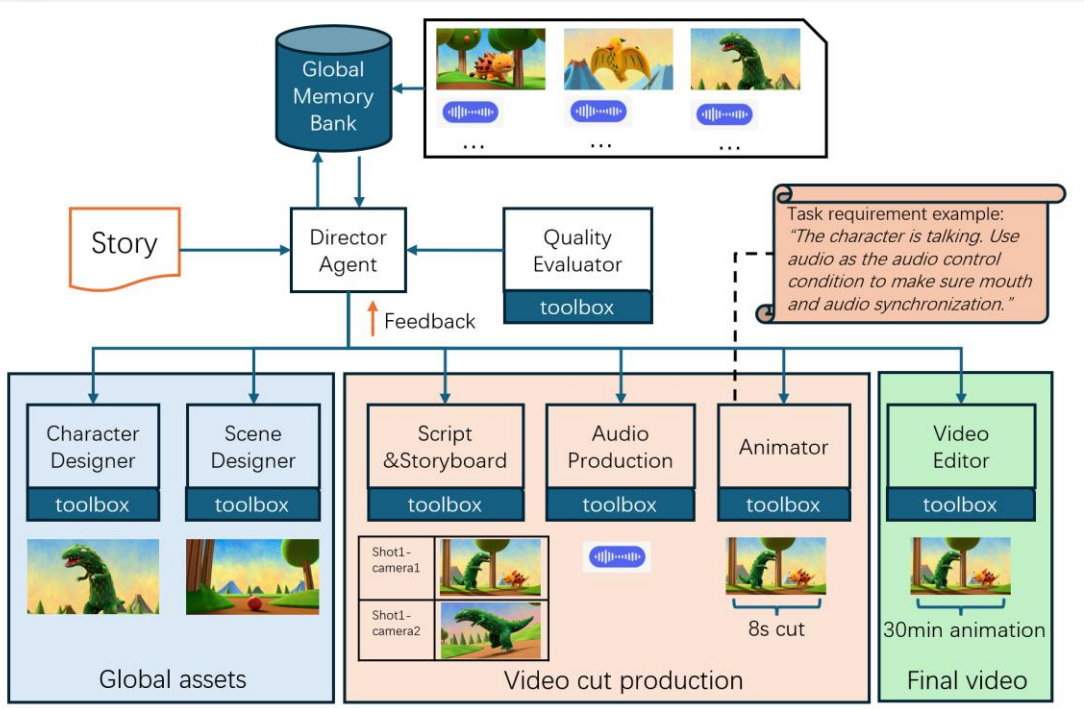
信息密度（叙事能力）：多镜头 * 多模态

https://github.com/RVC-Boss/GPT-SoVITS 51.6K Star

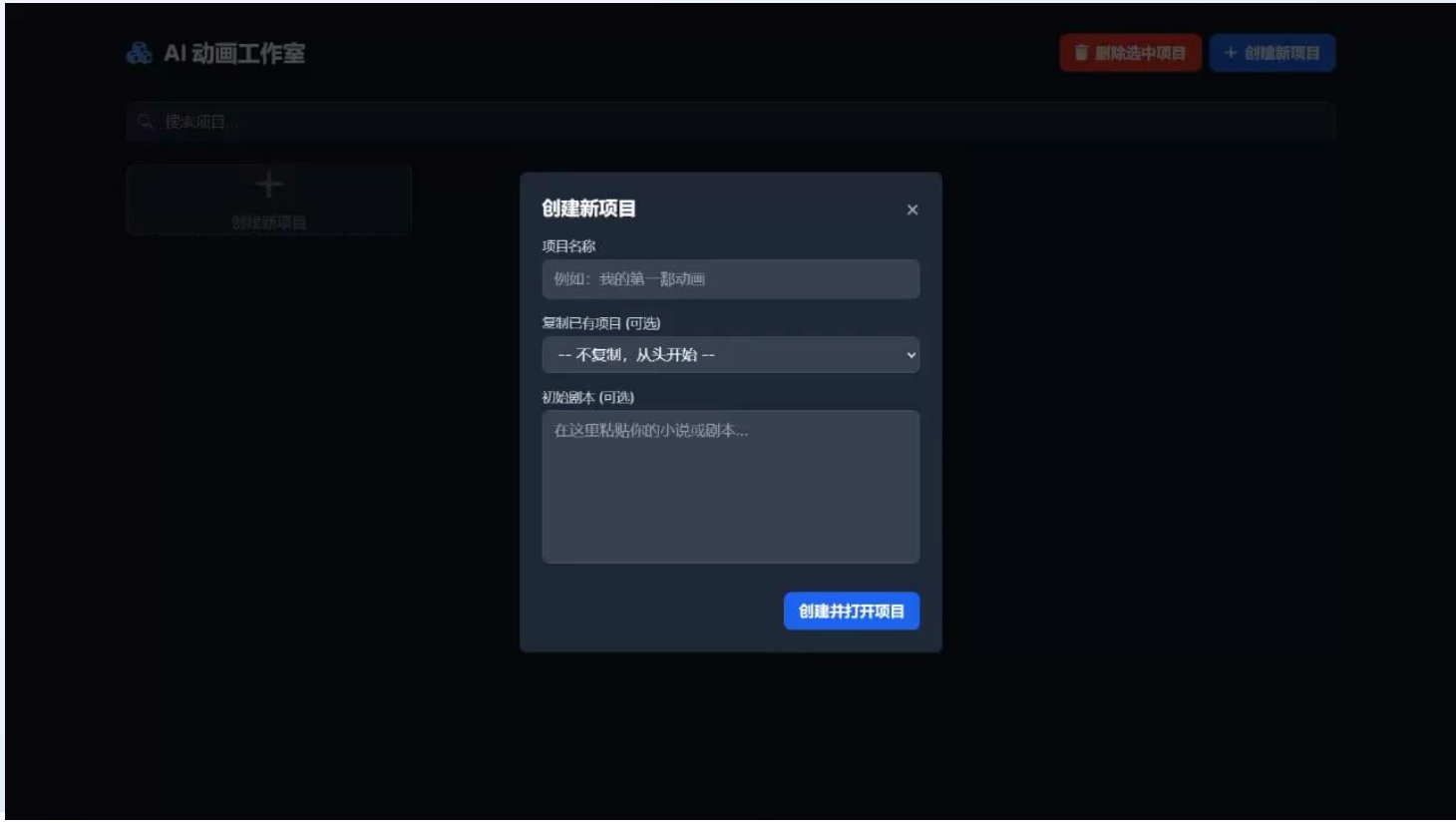
https://github.com/index-tts/index-tts 13.5K Star

https://github.com/bilibili/Index-anisora 2.1K Star

Multi-agent 辅助创作



AniME: Adaptive Multi-Agent Planning for Long Animation Generation. Lisai Zhang, Baohan Xu, et. al ***Accepted by Siggraph Aisa25***



动画视频创作现状和发展趋势

工具能力现状

- AI能够制作出对标传统动画行业1w-5w/分钟的动画视频，制作成本1-3K/分钟。
- 非动画专业创作者一天可以制作1分钟的动画

Agent辅助长视频创作的问题

- 帮助节约40%-60%的人工成本
- 由于创作内容质量反馈的主观性和稀疏性，系统难以得到不断的增强
- 单纯通过LLM planner组织workflow不能带来系统能力的增强，数据和反馈难以在系统的多个agent之间流转

技术发展方向

- 高信息密度的长视频生成，多模态、多镜头输出
- 实时的交互式创作： 对长视频生成过程中的问题及时纠正，创作工具与创意的充分融合，解决对ai工具使用经验依赖，降低创作门槛和难度
- 统一多模态整合（Any2Any），更加深度的视觉语义理解，深刻的理解用户创作意图

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

QCon
全球软件开发大会

InfoQ 极客传媒

THANKS

Join Us！Star AniSora！

https://github.com/bilibili/Index-anisora/tree/main