QCon
全球软件开发大会

InfoQ 极客传媒

EchoMimic

多模态大模型驱动下的生成式数字人技术与应用

李宇明

蚂蚁集团–支付宝多模态应用实验室–研究员

目 录

AiCon
全球人工智能开发与应用大会

上海站 | 北京站

全览 AI 行业落地精华

加速技术与应用的融合

大模型架构创新与端侧实践

AI Agent 构建与应用

AI 出海策略

智能数据新基建

AI 产品创新设计

大模型赋能研发

多模态大模型实战

5 月上海站 | 2025 年 5 月 23-24 日 上海中优城市万豪酒店

6 月北京站 | 2025 年 6 月 27-28 日 北京国际会议中心

咨询购票>

AICon 上海站
专题详情>

01

传统数字人与生成式数字人技术背景

传统数字人技术介绍–2D数字人技术路线

方法

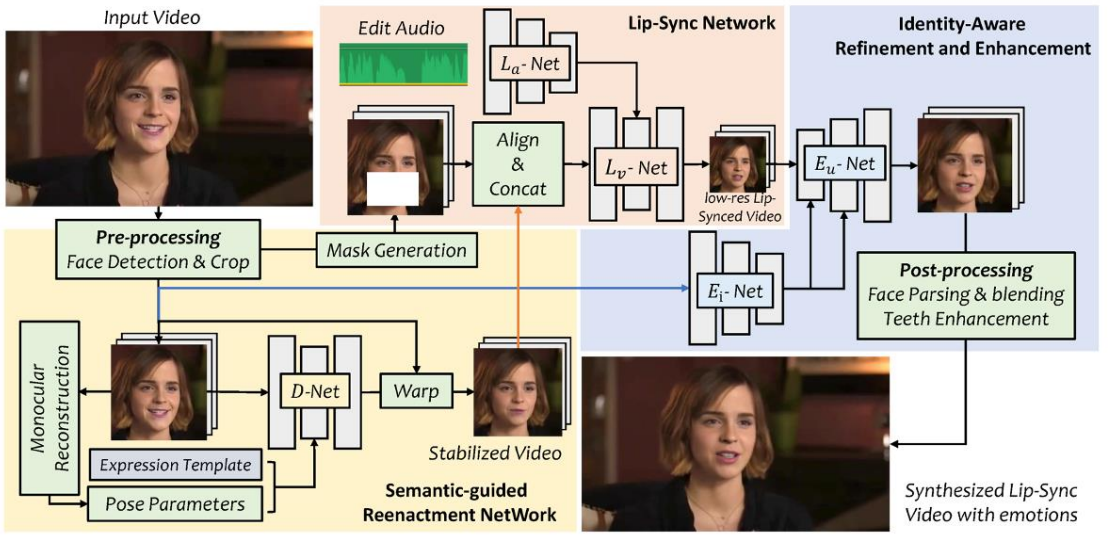
- 基于GAN的算法。通过对抗训练学习，对人物图像的嘴部进行精准编辑，确保嘴型与输入的语音同步，实现数字人语音播报。
- 基于NeRF的算法。通过构建神经辐射场对数字人进行个性化建模，提升嘴型生成的自然度、匹配度和语音播报个性化水平。

优势

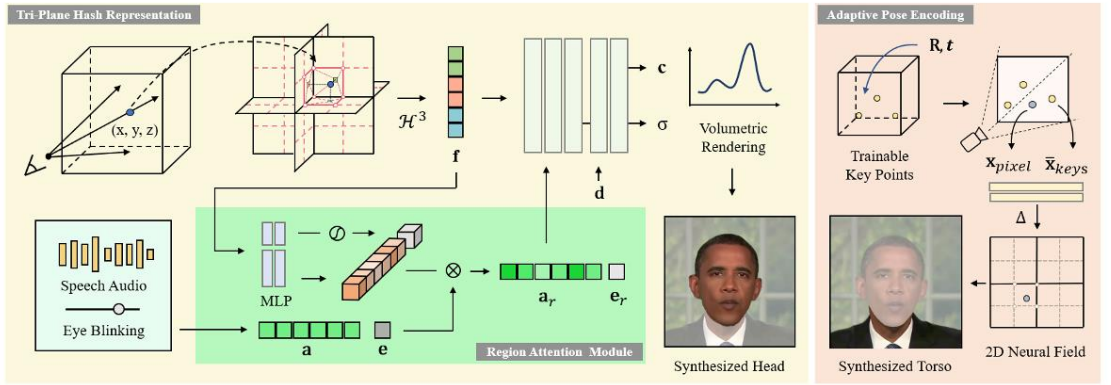
- 制作成本低，技术路线短平快。
- 在特定场景下能达到可接受的效果。

不足

- 优质的2D数字人应用效果依旧依赖于高水准的素材录制。
- 高质量躯体和手势动作视频生成仍面临挑战。
- 人物动作和嘴型生成的准确性、自然性和灵活性等方面仍有不足。



GAN类嘴形编辑算法

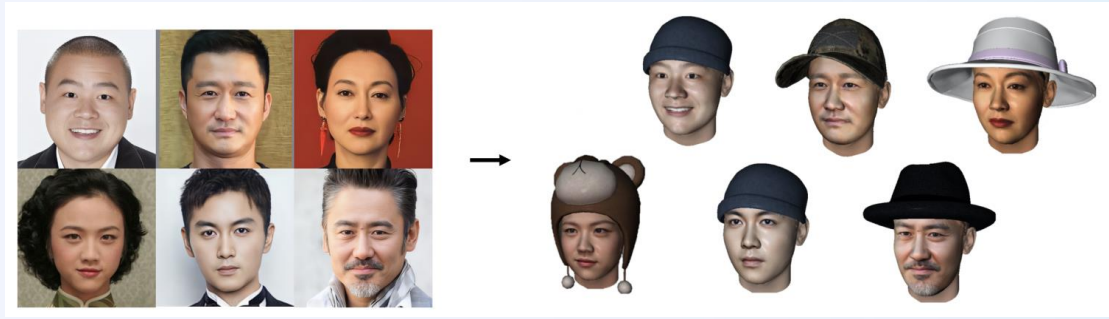


NeRF类数字人建模算法

传统数字人技术介绍–3D数字人技术路线

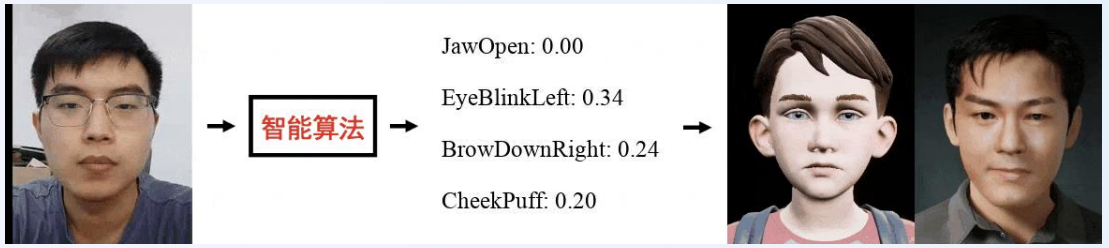
方法

- AI技术在3D数字人领域的应用主要集中在数字人智能建模和数字人智能驱动两个关键方向。
- 随着3DMM（三维人脸可形变模型）和可微分渲染技术的不断发展，现在可以以极低的成本实现3D数字人的建模和驱动。



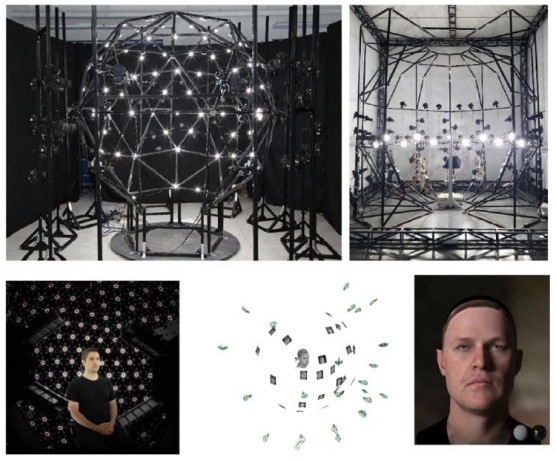
优势

- 3D数字人相比2D数字人有着更强的交互能力。
- 3D美术建模可以带来的更完美的数字人外貌与人设。



不足

- 技术链过长，人物建模、动作驱动、渲染展示等每个环节都有着复杂的技术栈。
- 智能化低成本的建模方式难以保障数字人建模质量，高质量的3D数字人建模依然依赖传统美工3D建模方式。
- 天然不适合需要超高写实人物形象的应用场景。



用于数字人建模的光场系统

专业动作捕捉设备

生成式数字人技术介绍

方法

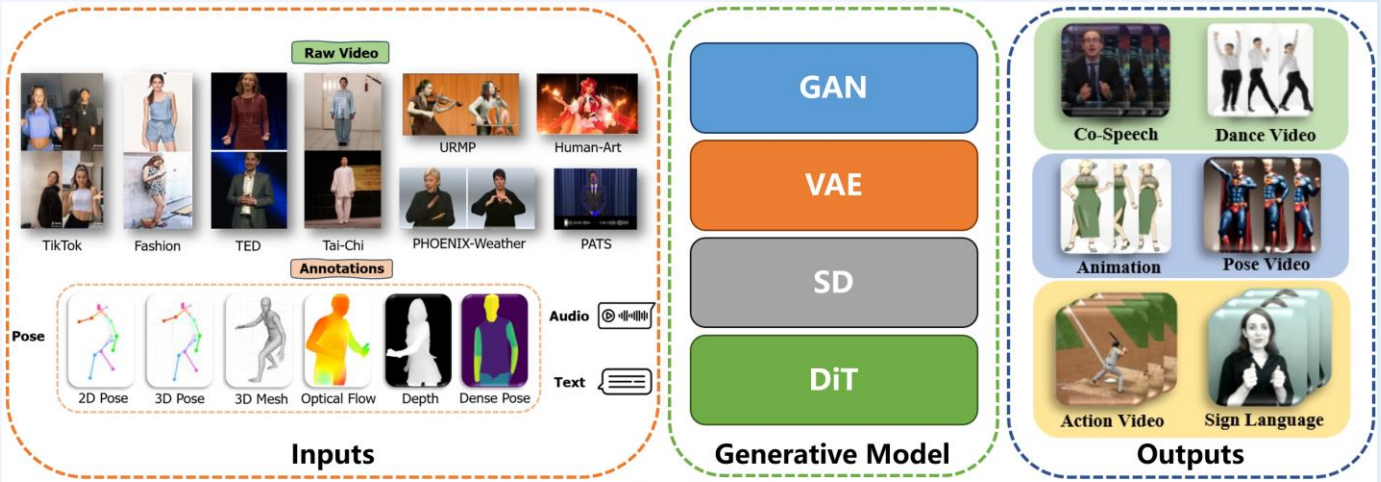
- 人工智能生成内容（AIGC）技术取得了突破性进展，AI绘画领域创新应用层出不穷。
- AIGC在视频生成方面也取得了显著成就，为生成式数字人领域带来了崭新的变化。

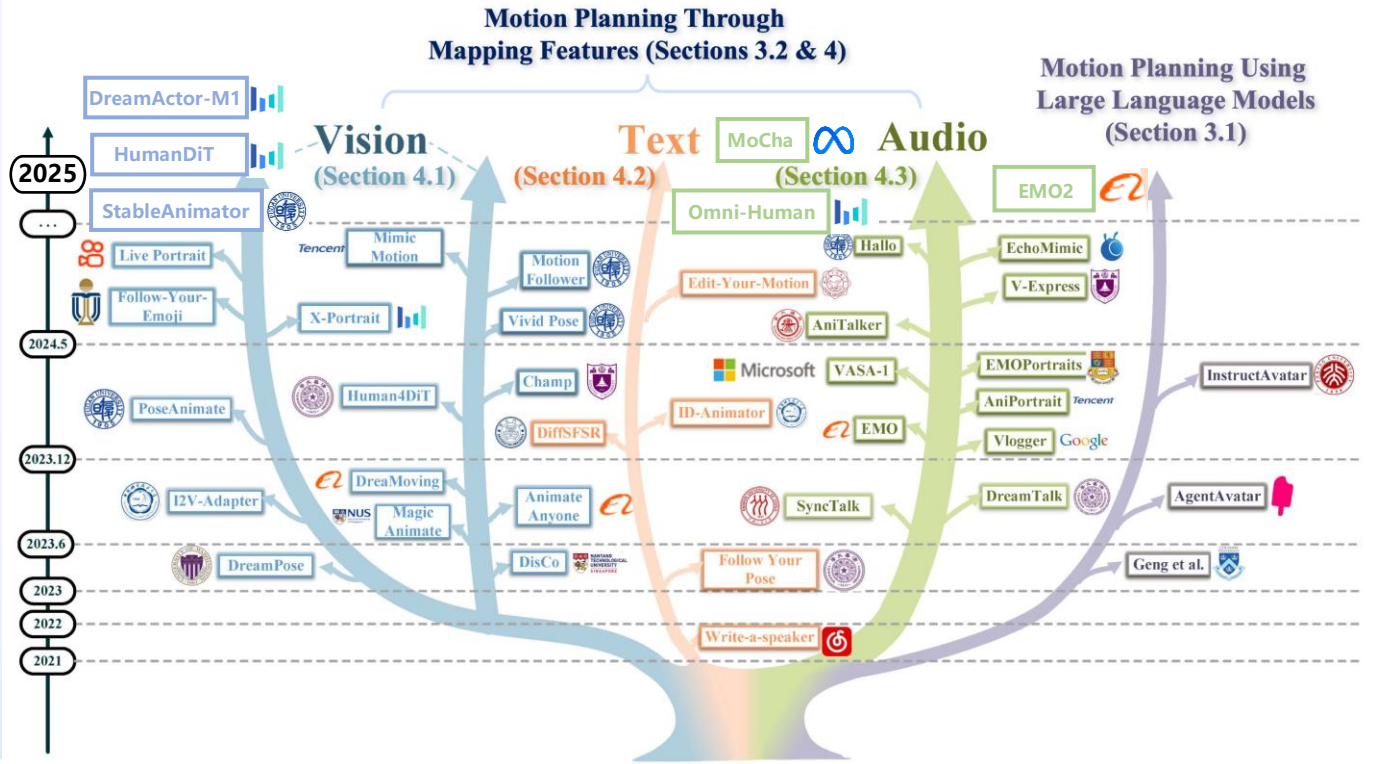
优势

- 在成本极低的情况下，可以创造出高品质的图像与视频内容。
- 数字人外貌与人设等展示素材均可以用AIGC生成。
- 可以利用语音、动作等对数字人进行相关控制。
- 算法效果天花板比较高。

不足

- 相关技术相对比较新、可参考的优秀工作不多。
- 算法推理成本和时间还比较高。
- 基于语音驱动的半身和全身数字人还没有成熟工作。

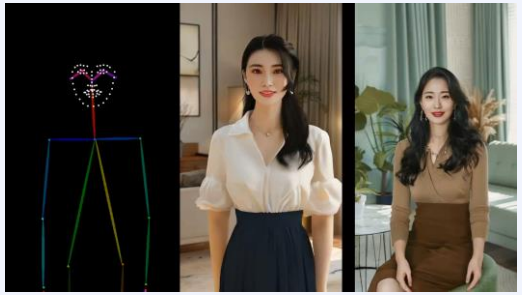




生成式数字人技术介绍

AnimateAnyone (Vision)
2023.11 阿里 未开源

EMO (Audio)
2024.02 阿里 未开源



MimicMotion (Vision)
2024.06 腾讯 开源

EchoMimicV1 (Audio+Vision)
2024.07 蚂蚁 开源

CyberHost (Audio+Vision)
2024.09 字节 未开源

EchoMimicV2 (Audio+Vision)
2024.11 蚂蚁 开源

EMO2 (Audio)
2025.01 阿里 未开源

HumanDiT (Vision)
2025.02 字节 未开源

OmniHuman (Audio)
2025.02 字节 未开源

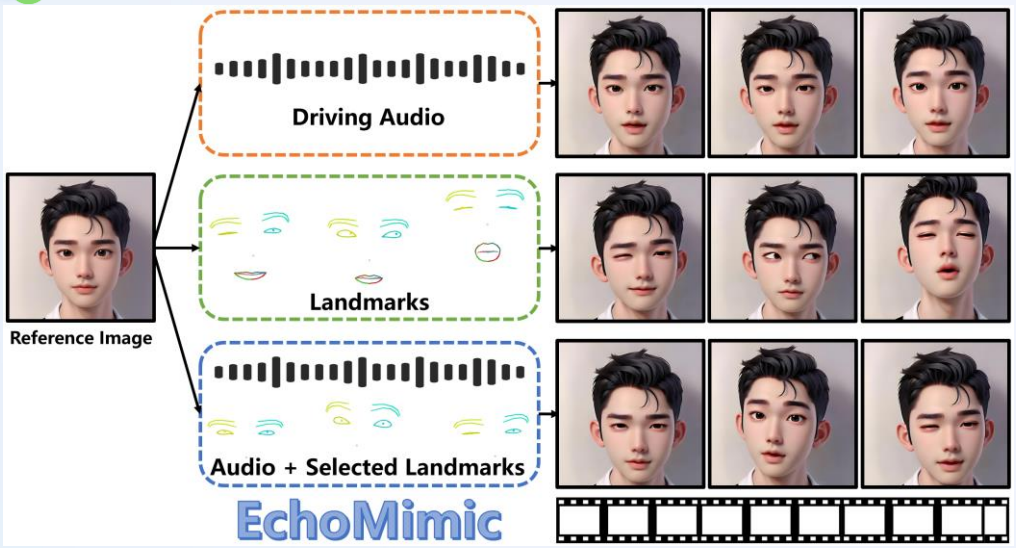
MoCha (Audio)
2025.03 Meta 未开源

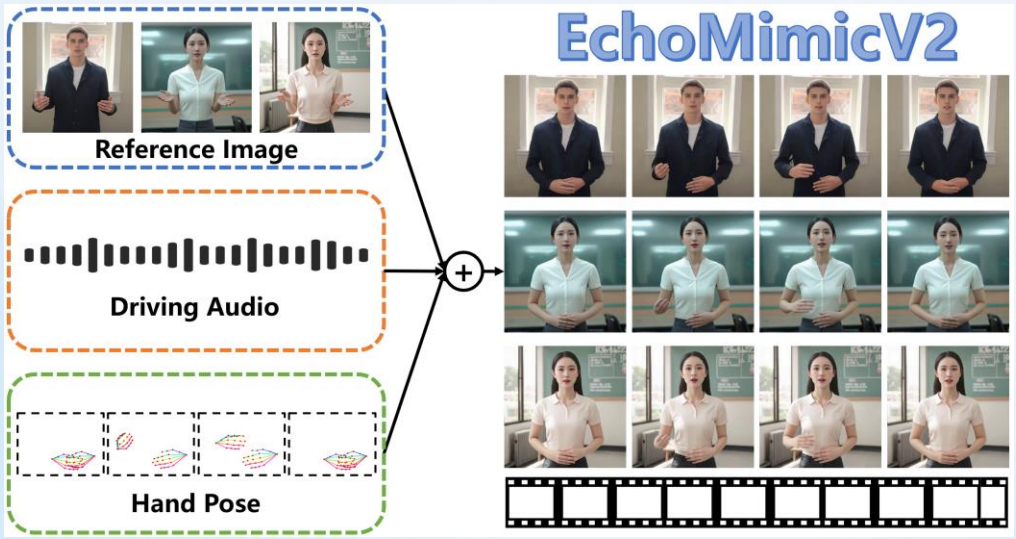
DreamActor–M1 (Vision)
2025.04 字节 未开源

02

EchoMimic
基于语音驱动的人像动画生成背后的技术

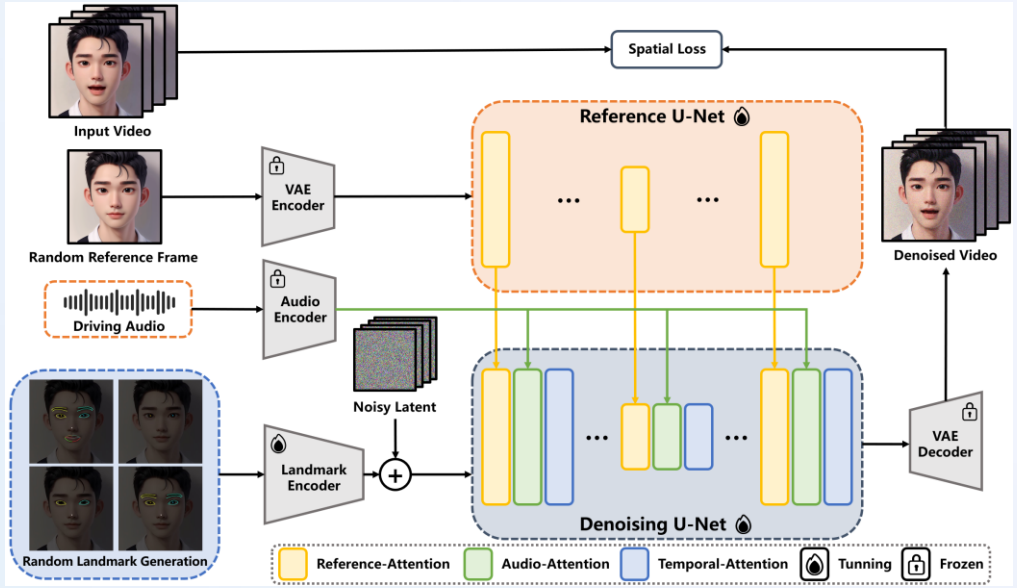
EchoMimic技术细节与亮点

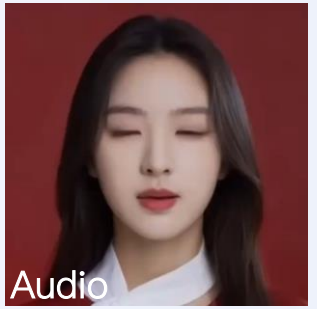


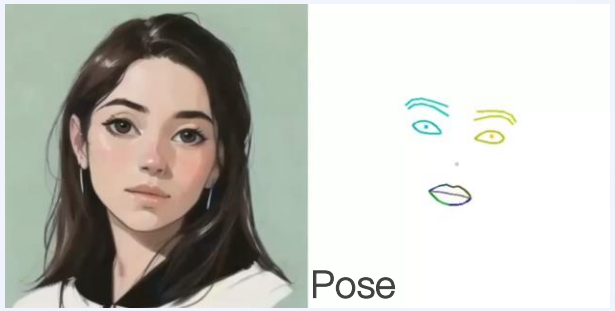


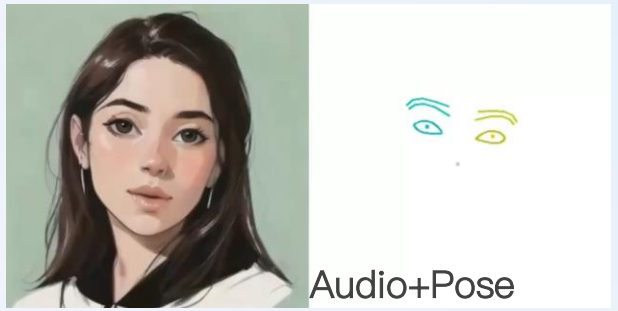
- EchoMimic是专注于增强2D数字人物驱动效能的算法，用户仅需上传一张数字人或真实人物的图片及一段语音或视频资料，即可生成与之匹配的说话场景视频。
- 该技术在表现效果上接近当前市场上的商业解决方案，且在驱动模式上展现出高度灵活性，支持语音、姿态或二者的组合驱动，为用户带来灵活的定制化体验。
- 项目开源地址：
  - V1版本：https://github.com/antgroup/echomimic
  - V2版本：https://github.com/antgroup/echomimic_v2

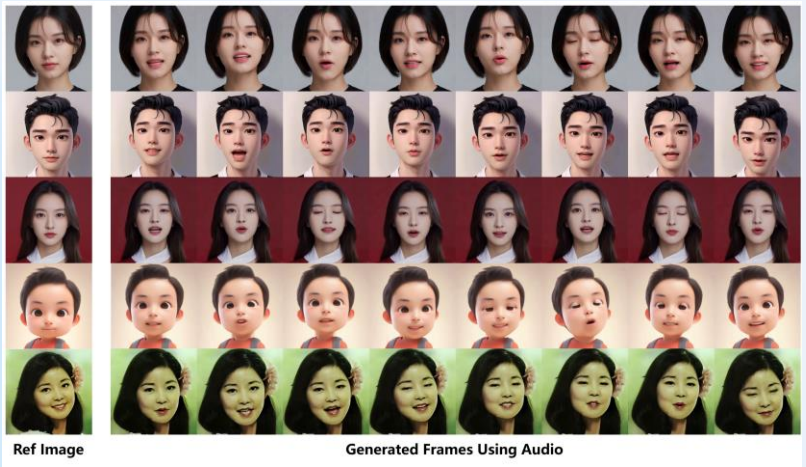
EchoMimic技术细节与亮点

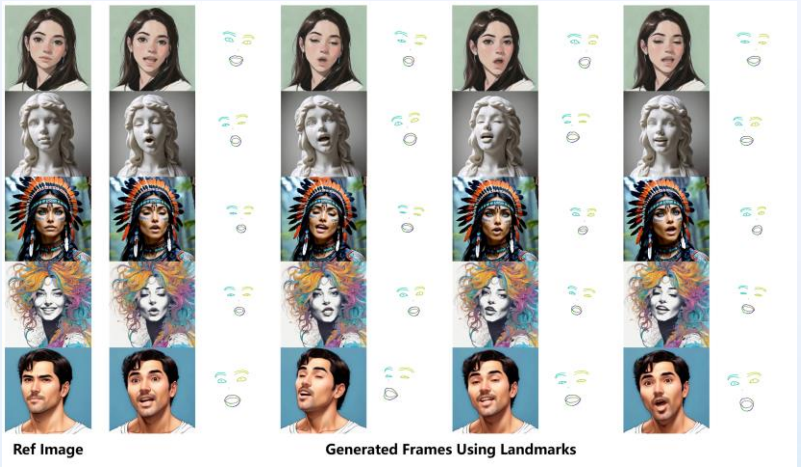


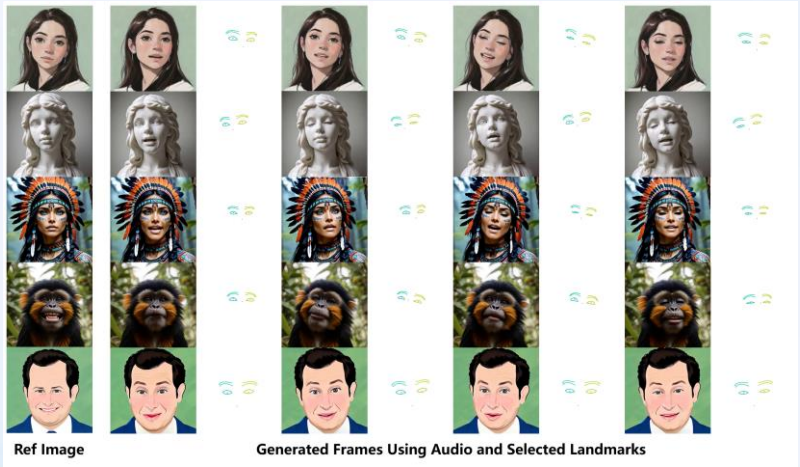




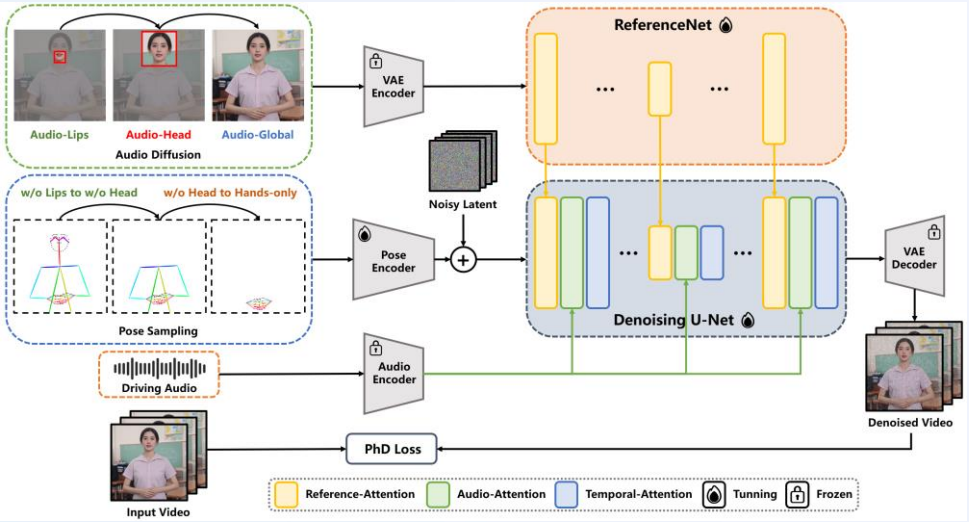


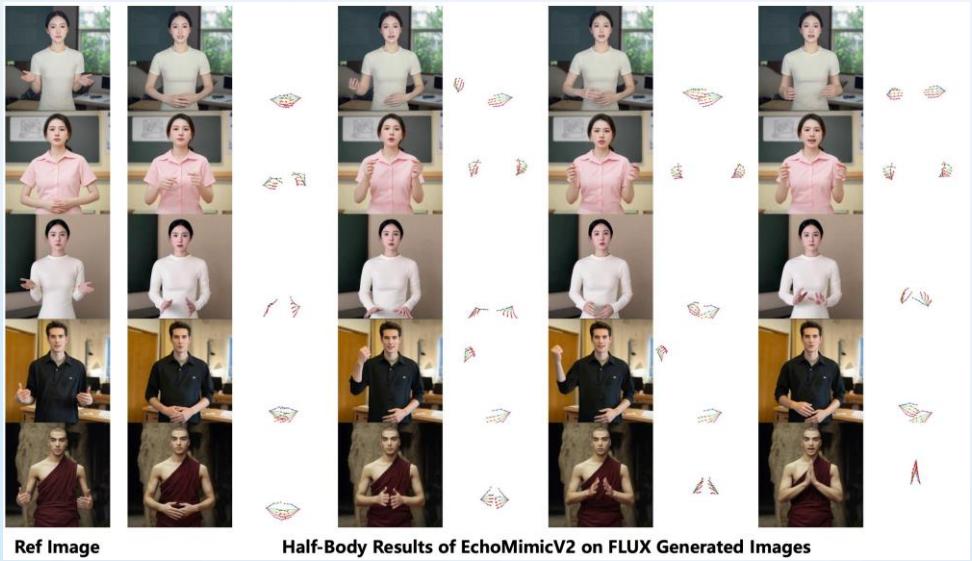


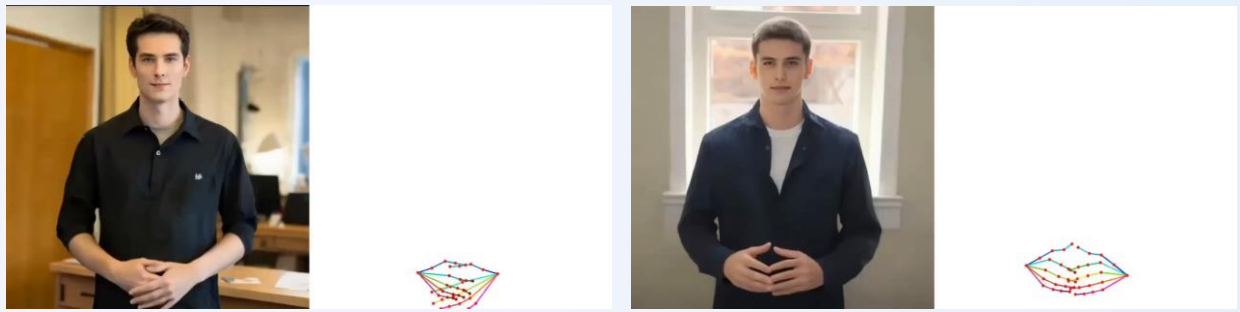




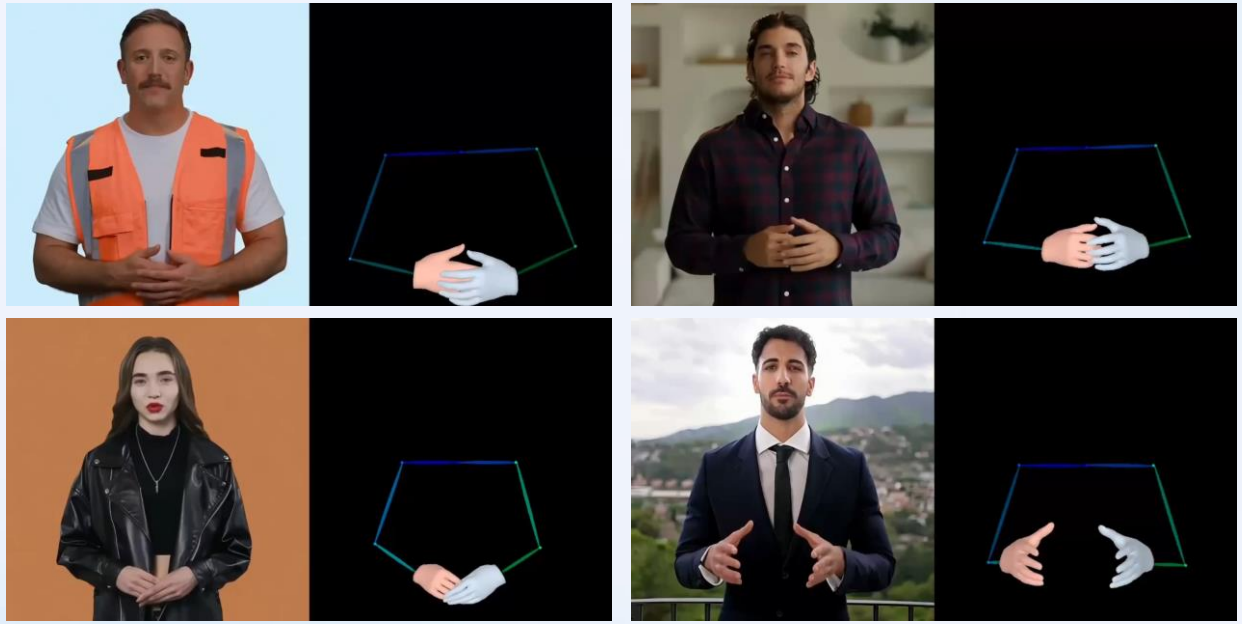
EchoMimic技术细节与亮点







开源版本（预定义手部Pose）



内部版本（Audio2Pose自动生成）

EchoMimic技术细节与亮点

- 利用步数蒸馏算法，对EchoMimic训练框架进行改造。
- 以原来的模型作为Teacher，将40步的Teacher模型分成4段，每一段由10小段组成。
- 训练中对每一段生成过程，将10步Teacher的推理用MSC蒸馏为一步，作为 Student 的拟合目标。
- 加速后，视频生成耗时在A100 GPU提速约9倍。

原始模型　加速模型　原始模型　加速模型

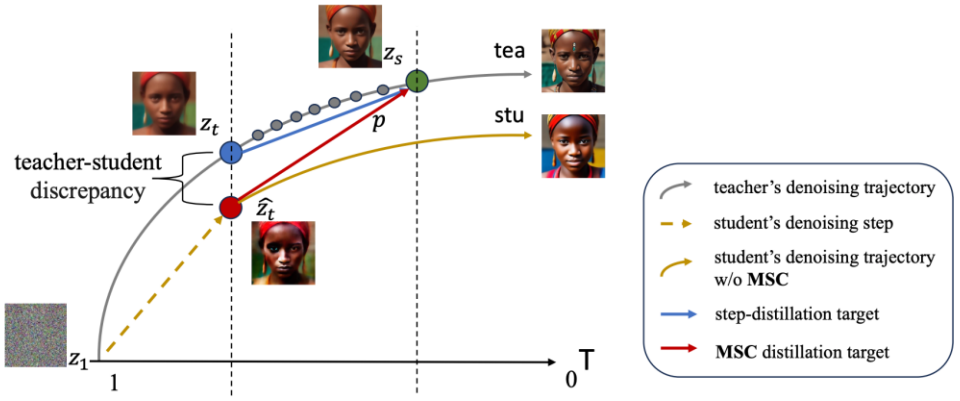
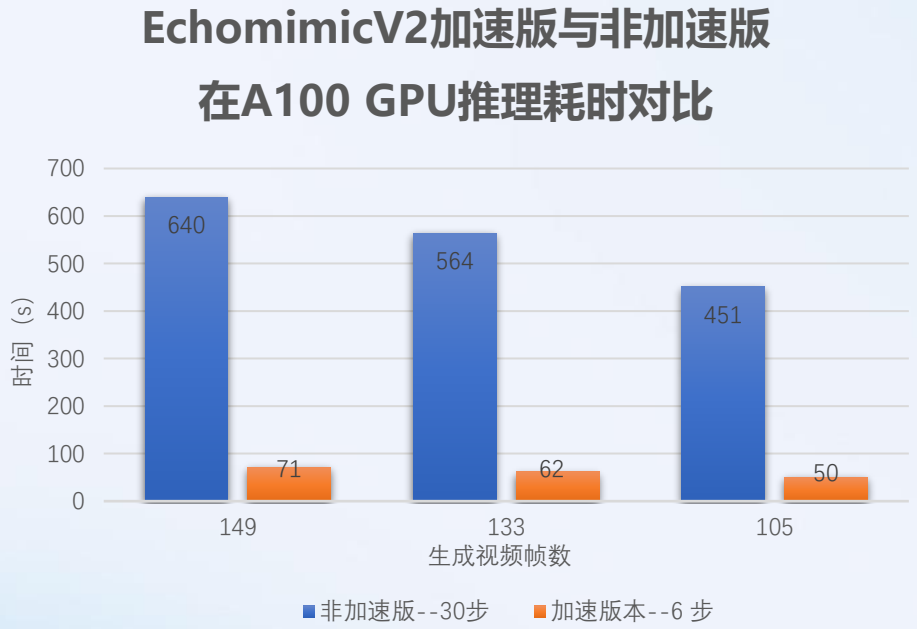


Fig. 3: An illustraion of Muiti-step Consistency (MSC). When distilling a faster student model, teacher-student discrepancy exists and gradually accumulates, causing the content generated by the student to be inconsistent with the teacher (from the same noise). Based on the step distillation method, MSC is used to train the student to approach the teacher's trajectory even when error occurs, thus ensuring consistency in muiti-step samplings.



实验结果分析–EchoMimicV1

对比实验结果

第三方评测结果

Table 1: The quantitative comparisons with the existed portrait image animation approaches on the HDTF.

Method	FID↓	FVD↓	SSIM↑	E-FID↓
SadTalker	41.535	1138.056	0.790	2.248
AniPortrait	53.143	1038.239	0.751	1.939
V-Express	58.230	1184.203	0.724	1.807
Hallo	37.659	501.074	0.781	1.525
EchoMimic	**29.136**	**492.784**	**0.812**	**1.112**

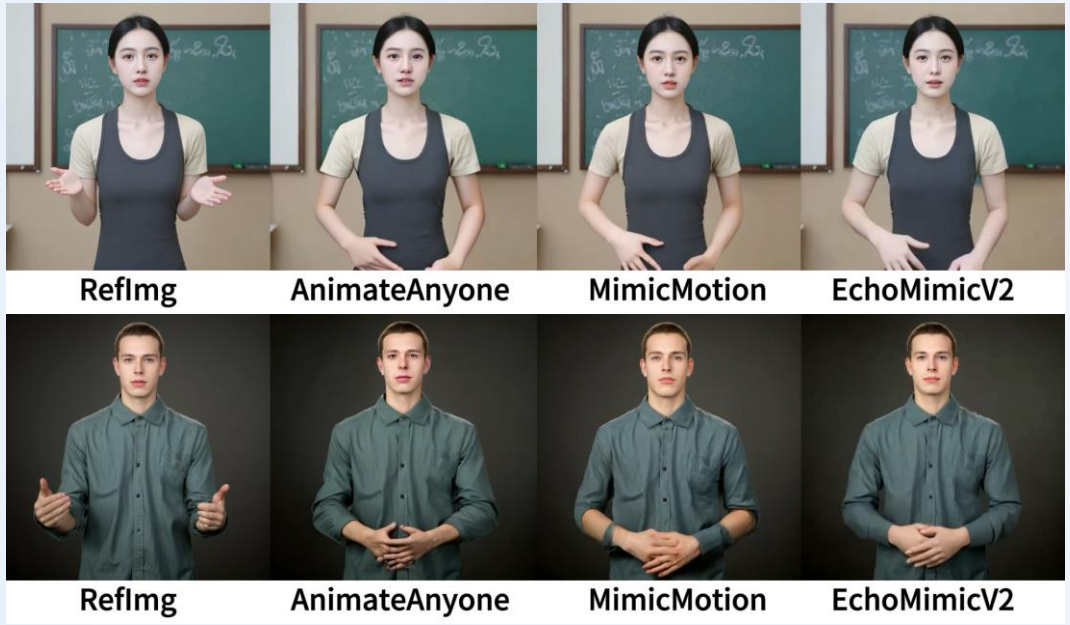
Table 2: The quantitative comparisons with the existed portrait image animation approaches on the CelebV-HQ dataset.

Method	FID↓	FVD↓	SSIM↑	E-FID↓
SadTalker	93.883	1454.328	0.641	3.971
AniPortrait	92.003	1297.805	0.609	3.916
V-Express	95.483	2126.248	0.524	4.720
Hallo	70.420	**1073.718**	**0.644**	2.851
EchoMimic	**63.258**	1115.857	0.633	**2.723**

Table 3: The quantitative comparisons with the existed portrait image animation approaches on the our collected dataset.

Method	FID↓	FVD↓	SSIM↑	E-FID↓
SadTalker	64.633	1681.836	**0.699**	2.150
AniPortrait	66.884	2054.527	0.665	2.312
V-Express	62.721	2103.213	0.658	1.689
Hallo	50.474	1405.215	0.690	1.452
EchoMimic	**43.272**	**988.144**	0.691	**1.421**

实验结果分析–EchoMimicV2



与Pose驱动算法对比实验结果

Methods	FID↓	FVD↓	SSIM↑	PSNR↑	E-FID↓	Sync-D↓	Sync-C↑	HKC↑	HKV↑	CSIM↑
AnimateAnyone[14]	58.98	1016.47	0.729	20.579	3.784	13.887	0.987	0.809	23.87	0.387
MimicMotion[38]	53.47	622.62	0.702	19.278	2.628	7.958	1.495	0.907	24.82	0.526
EchoMimicV2	**49.33**	**598.45**	**0.738**	**21.986**	2.218	7.021	7.219	0.923	25.28	0.558
w/o Initial Pose	49.99	602.08	0.730	21.708	2.235	6.976	7.019	0.873	23.97	0.550
w/o Iterative Pose Sampl.	50.01	605.29	0.727	21.276	2.276	6.987	7.005	0.917	25.24	0.527
w/o Spatial Pose Sampl.	49.39	593.98	0.740	21.994	2.208	7.023	7.220	0.922	25.25	0.532
w/o Headshot Data Aug.	51.29	610.87	0.711	20.965	2.961	6.792	6.394	0.921	25.27	0.527
w/o Audio-Lips Sync.	49.37	599.82	0.722	21.988	2.629	6.803	6.463	0.919	25.24	0.512
w/o Audio-Face Sync.	51.11	620.75	0.719	21.837	2.983	6.807	6.286	0.922	25.26	0.518
w/o Audio-Body Corr.	50.36	603.84	0.733	21.980	2.217	7.025	7.226	0.906	24.98	0.541
w/o L_{pose}	51.30	620.27	0.695	20.783	2.766	6.954	7.063	0.874	22.83	0.549
w/o L_{detail}	51.61	623.56	0.689	20.038	2.904	6.807	6.985	0.896	24.37	0.541
w/o L_{low}	50.86	602.31	0.675	19.786	2.226	7.028	7.223	0.913	25.20	0.547
Straightforward Baseline	51.53	624.98	0.664	19.269	2.779	7.208	6.706	0.825	22.57	0.508

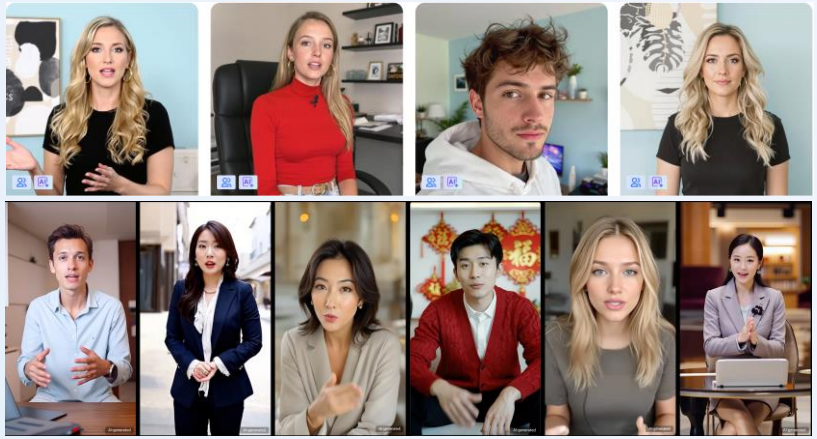
Table 2. Quantitative comparison and ablation study of our proposed EchoMimicV2 and other SOTA methods.

与Audio驱动算法对比实验结果

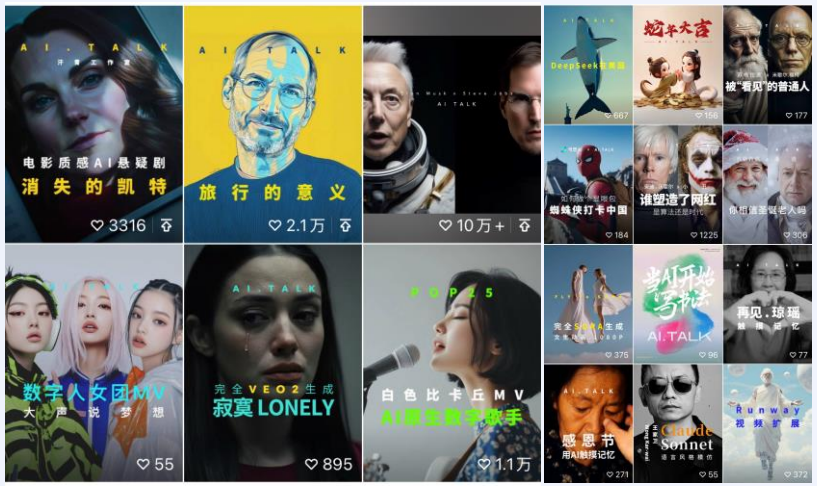
03

应用场景探索

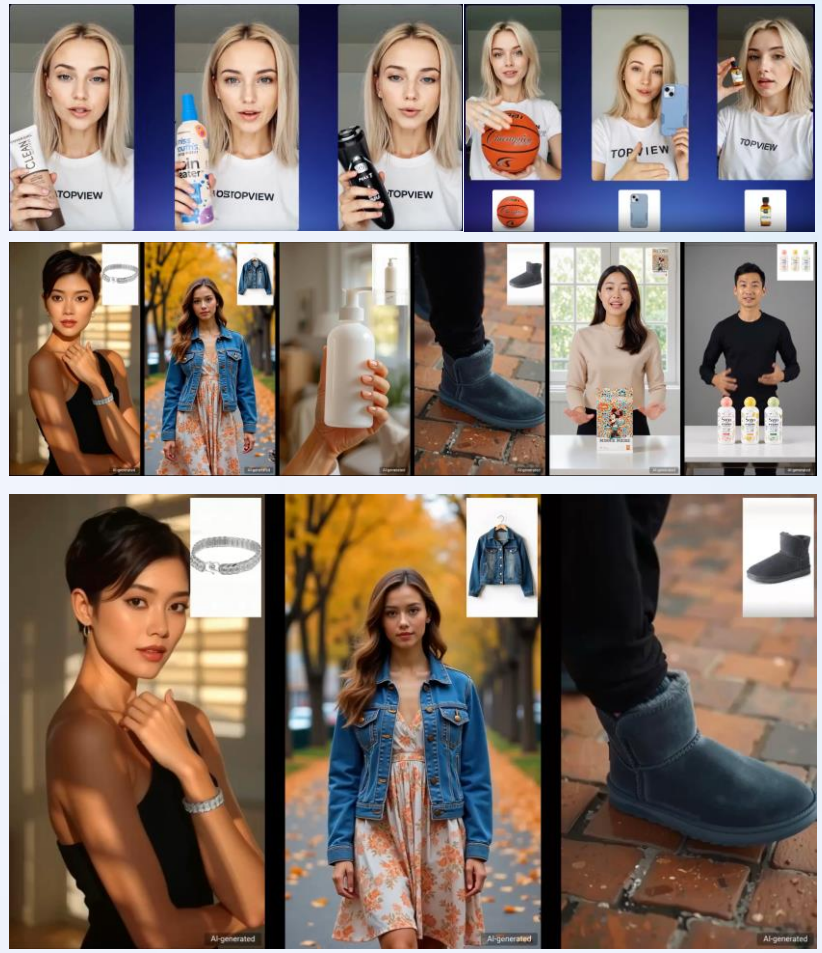
应用场景探索



实时交互
生成式数字人+多模态大模型



AI创作
生成式数字人+音乐生成大模型



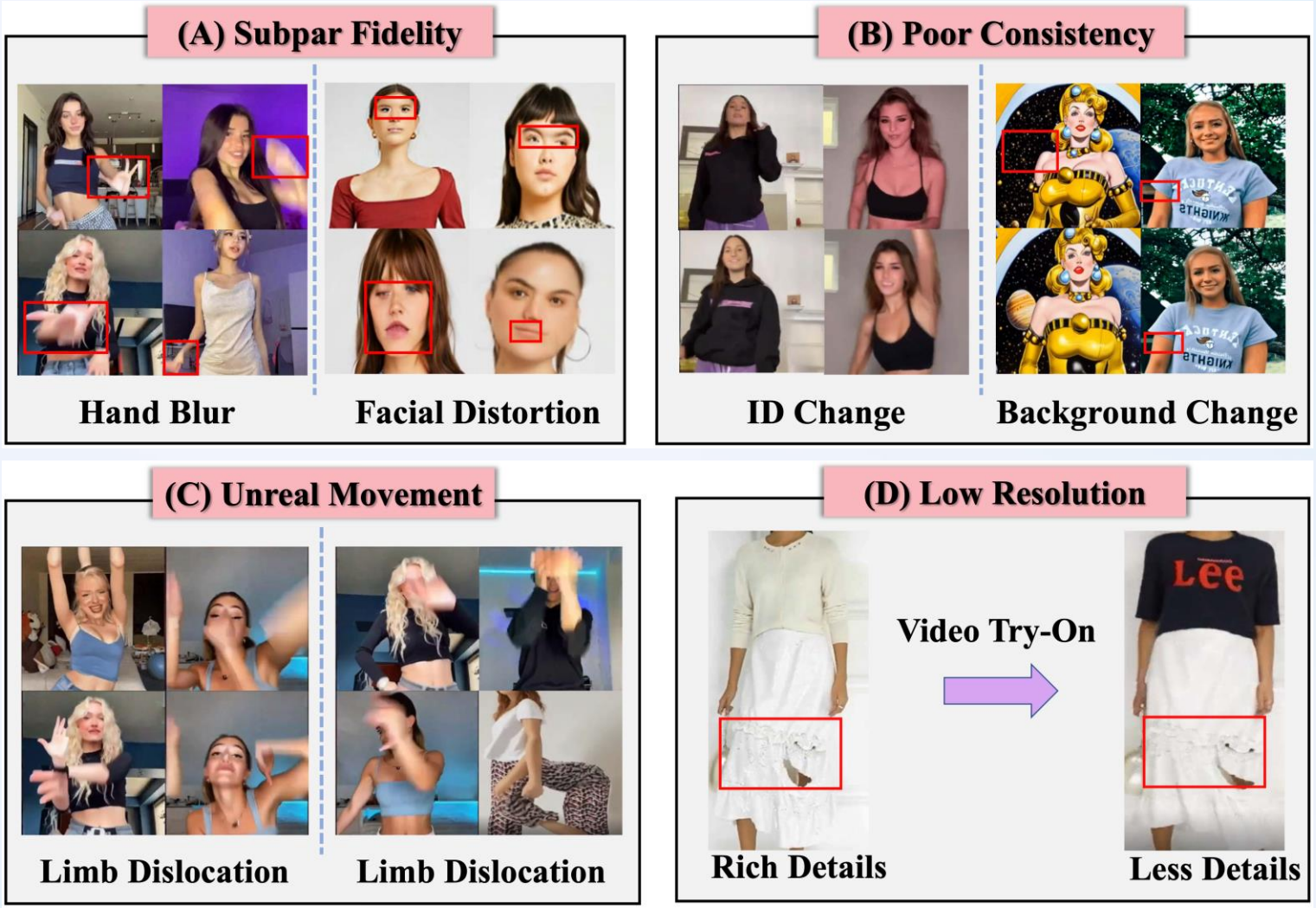
视频广告
商品交互数字人垂类模型

04

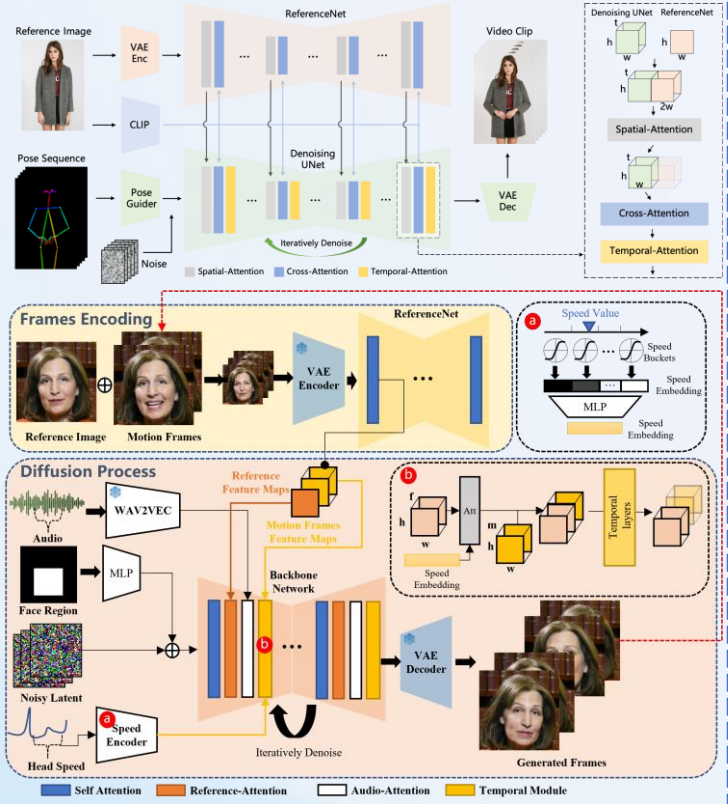
总结与展望

生成式数字人存在的问题和挑战

- **保真度差**。手部、牙齿、面部细节。
- **一致性差**。ID变化、背景不协调、衣物细节变化。
- **动作不自然**。动作与语音、图像细节不一致。
- **分辨率低**。高清生成、快速生成仍有困难。

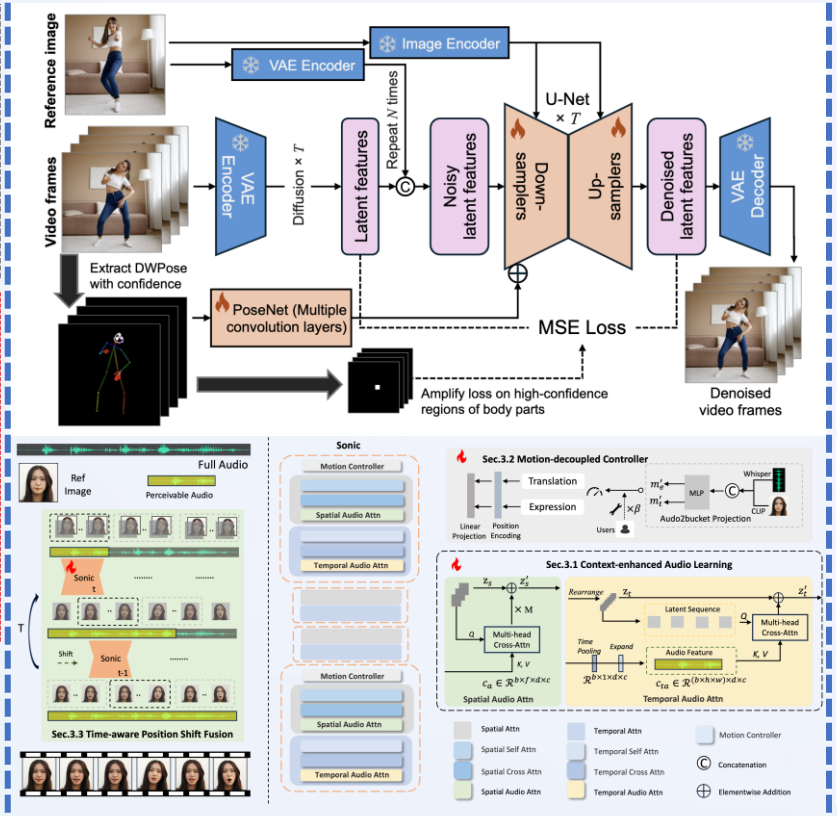


生成式数字人开发新范式



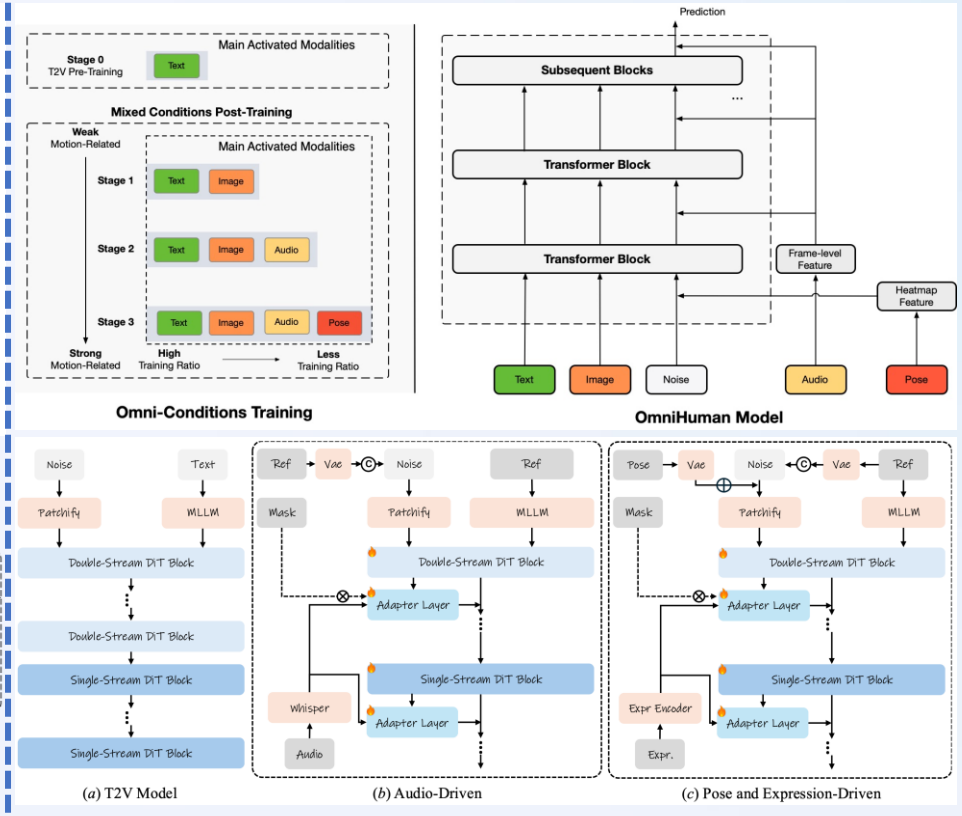
SD双塔架构

AnimateAnyone, EMOV1/V2, EchoMimicV1/V2



SVD单塔架构

MimicMotion, Sonic, StableAnimator



视频生成I2V基模+组件

Omni-Human (Seaweed+), Hunyuan+, Wan2.1+, Goku+

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