AIGC时代下的数字人技术革新与介绍

报告人：范肇心

个人介绍

范肇心

北航人工智能学院助理研究员

TeleAI双聘研究员

学习经历：

2019-2024中国人民大学 信息学院计算机博士

师从何军教授

2021-2022卡内基梅隆大学，获得全额奖学金资助

师从Min Xu教授

2022-2023香港科技大学，获得全额奖学金资助

师从陈凯教授（长江）

工作经验：

曾就职于格林深瞳、字节跳动&Xreal

2022年以来就职于深锶科技（千万美元Pre A融资）从事数字人技术研究

担任算法总监-首席青年科学家

提纲：

1. AIGC与2D 说话人驱动
2. 个性化与情绪化3D数字人
3. 人体重建技术探索与应用
4. 多么他驱动的人体动作生成

1.AIGC与2D 说话人驱动

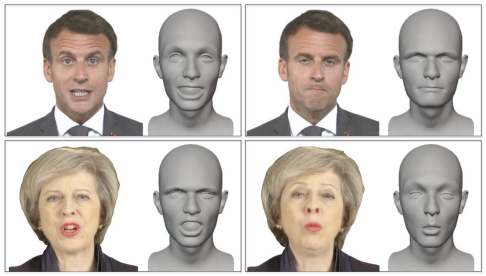
1.1 研究内容：2D精准说话人生成

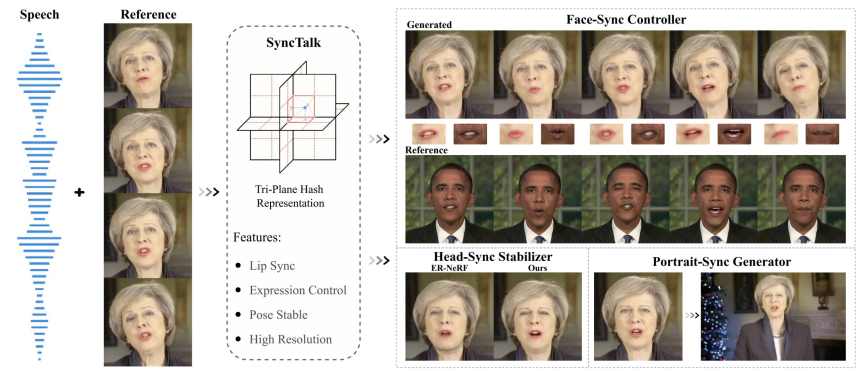
问题背景

现有方法生成说话人时人物特征难以保持，唇部同步性较差，表情无法控制，头部姿态抖动。

研究内容

- 说话人唇部与说话内容的同步响应
- 说话人表情控制
- 说话人头部姿态稳定性





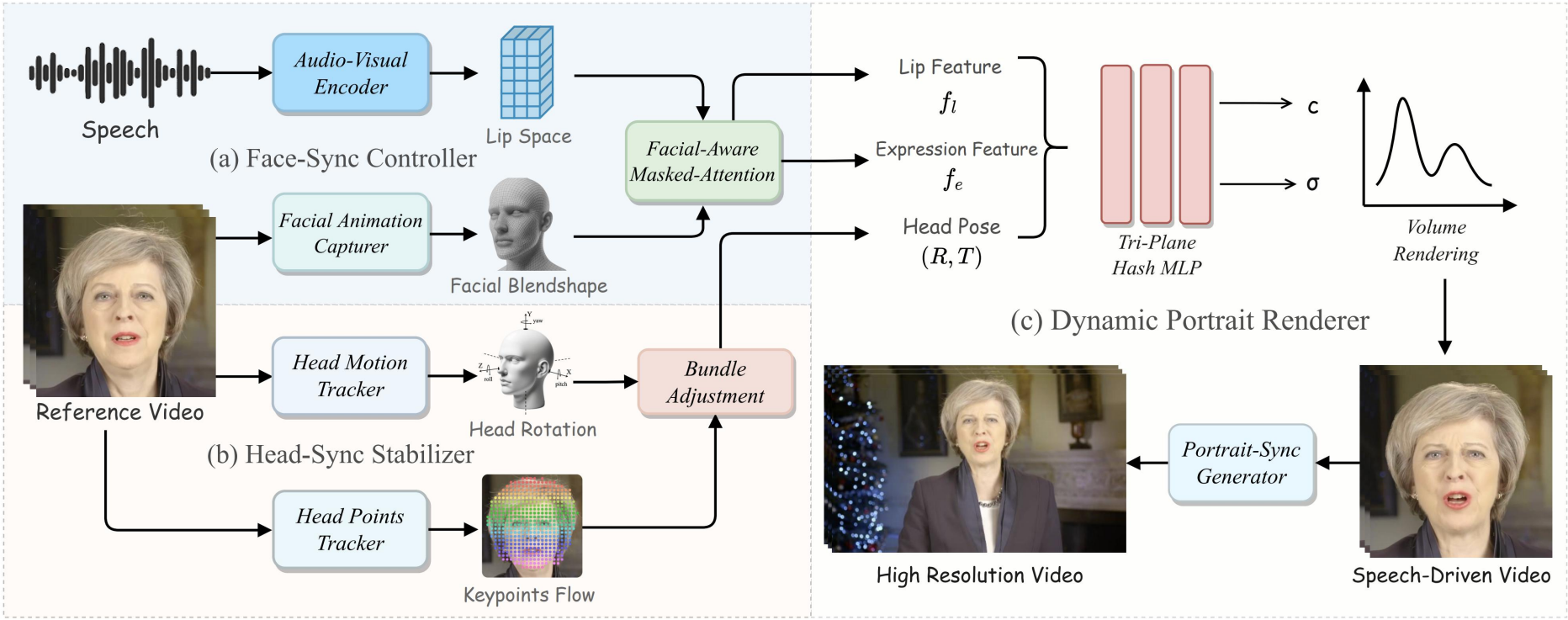
1.1 创新点：精确面部运动捕捉方法

提出SyncTalk，实现精确的面部运动捕捉

传统： 人物特征不稳定

瓶颈： 面部运动的准确捕捉

创新： 使用Blendshape进行精确的面部运动捕捉，并将其作为特征，对面部运动进行建模



1.1 创新点：精确面部运动捕捉方法

与其他方法进行对比

	Methods	PSNR ↑	LPIPS ↓	MS-SSIM ↑	FID ↓	NIQE ↓	BRISQUE ↓	LMD ↓	AUE ↓	LSE-C ↑
GAN	Wav2Lip (ACM MM 20 [35])	33.4385	0.0697	0.9781	16.0228	14.5367	44.2659	4.9630	2.9029	**9.2387**
GAN	VideoReTalking (SIGGRAPH Asia 22 [7])	31.7923	0.0488	0.9680	9.2063	14.2410	43.0465	5.8575	3.3308	7.9683
GAN	DINet (AAAI 23 [51])	31.6475	0.0443	0.9640	9.4300	14.6850	40.3650	4.3725	3.6875	6.5653
GAN	TalkLip (CVPR 23 [42])	32.5154	0.0782	0.9697	18.4997	14.6385	46.6717	5.8605	2.9579	5.9472
GAN	IP-LAP (CVPR 23 [54])	35.1525	0.0443	0.9803	8.2125	14.6400	42.0750	3.3350	2.8400	4.9541
NeRF	AD-NeRF (ICCV 21 [14])	26.7291	0.1536	0.9111	28.9862	14.9091	55.4667	2.9995	5.5481	4.4996
NeRF	RAD-NeRF (arXiv 22 [39])	31.7754	0.0778	0.9452	8.6570	13.4433	44.6892	2.9115	5.0958	5.5219
NeRF	GeneFace (ICLR 23 [45])	24.8165	0.1178	0.8753	21.7084	13.3353	46.5061	4.2859	5.4527	5.1950
NeRF	ER-NeRF (ICCV 23 [22])	32.5216	0.0334	0.9501	5.2936	13.7048	34.7361	2.8137	4.1873	5.7749
NeRF	SyncTalk (w/o Portrait)	35.3542	0.0235	0.9769	3.9247	**13.1333**	**33.2954**	2.5714	**2.5796**	8.1331
NeRF	SyncTalk (Portrait)	**37.4017**	**0.0113**	**0.9841**	**2.7070**	14.2165	37.3042	**2.5043**	3.2074	8.0263

	Wav2Lip [35]	DINet [51]	TalkLip [42]	IP-LAP [54]	AD-NeRF [14]	GeneFace [45]	ER-NeRF [22]	SyncTalk
Lip-sync Accuracy	3.839	3.696	2.893	3.161	2.696	2.982	3.189	**4.304**
Exp-sync Accuracy	3.536	3.482	2.607	3.411	2.250	3.036	2.946	**4.036**
Pose-sync Accuracy	3.571	3.571	2.875	3.696	2.232	2.929	2.607	**3.980**
Image Quality	2.500	2.696	2.054	3.571	2.464	3.482	3.036	**4.054**
Video Realness	2.929	2.429	2.429	3.161	2.036	2.732	2.518	**4.018**

通过定量评价和用户研究表明，我们的方法在多个任务上的视觉效果显著超越了先前的方法，并且得到了用户的认可。

相关成果以通讯作者发表在ICCV 2023上

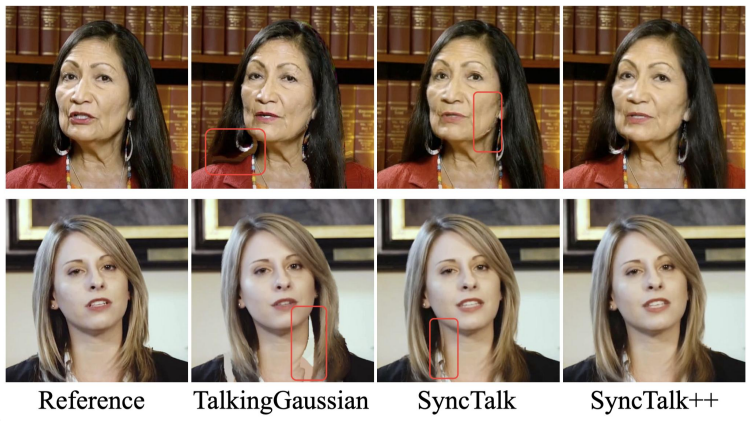
1.2 研究内容：2D高效说话人生成

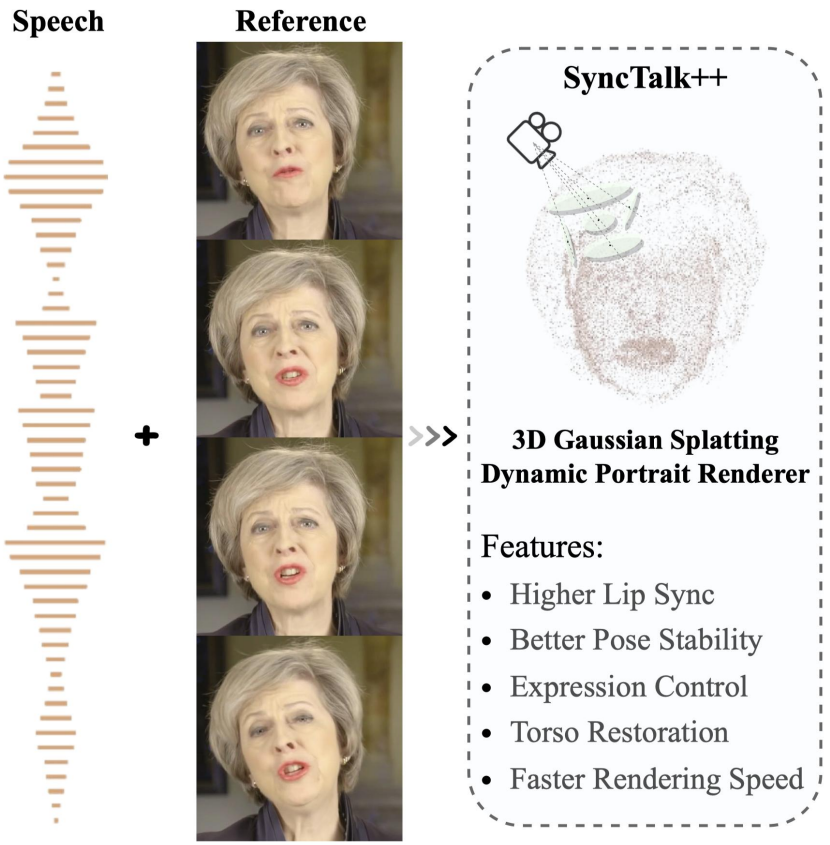
问题背景

现有方法使用NeRF渲染速度太慢，难以适应实时应用。
推理泛化性不高。

研究内容

- 找到问题根源并进行解决，提升当前数字人的实现效果
- 首要解决伪影问题
- 进一步提升面部运动稳定性
- 生成更加多样的表情





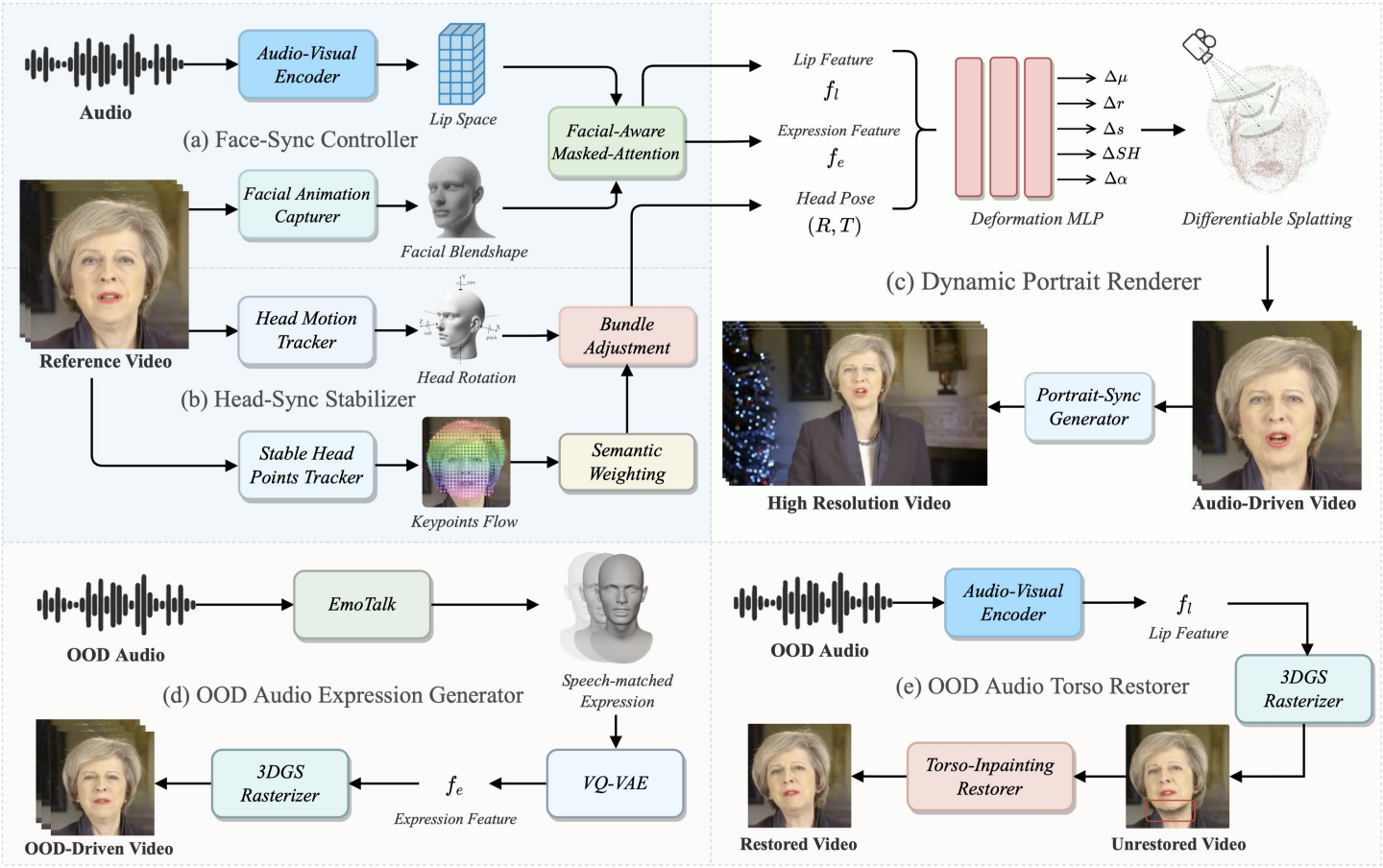
1.2 创新点：基于Gaussian Splatting的高效合成方法

提出了SyncTalk++，实现高效的说话人生成

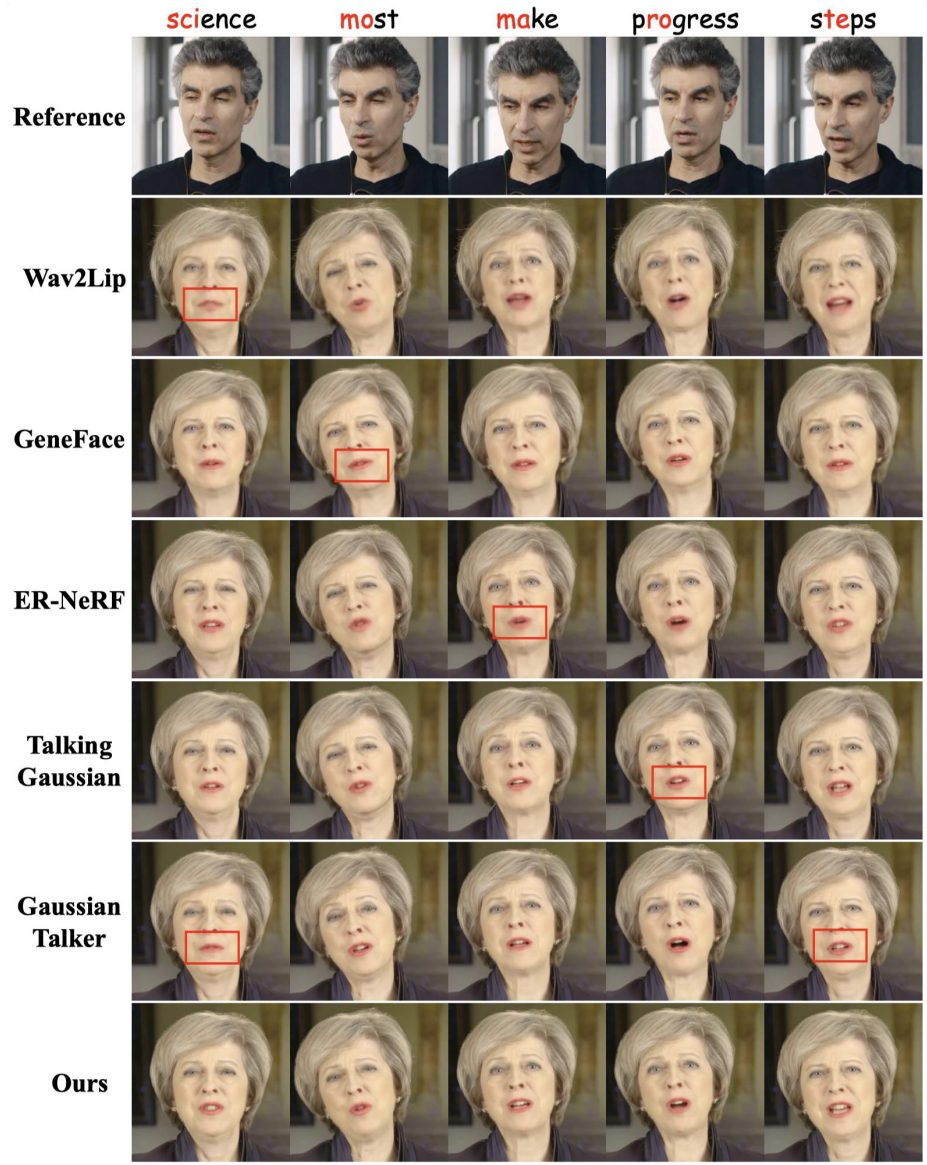
传统： 效率低下难以实时应用

瓶颈： 提升渲染速度

创新： Gaussian Splatting；面部追踪的语义加权模块，提升面部运动稳定性；表情生成器，生成更加多样的表情；躯干修复器，减少因为OOD音频推理带来的伪影问题



1.2 创新点：基于Gaussian Splatting的高效合成方法



	PSNR ↑	LPIPS ↓	MS-SSIM ↑	FID ↓
SyncTalk (w/o Portrait)	35.3542	0.0235	0.9768	3.9247
SyncTalk (Portrait)	37.4016	0.0113	0.9841	2.7070
SyncTalk++ (w/o Portrait)	36.3779	0.0201	0.9826	3.8771
SyncTalk++ (Portrait)	**39.5748**	**0.0097**	**0.9905**	**2.1958**

Methods	Audio A		Audio B	
	LSE-D ↓	LSE-C ↑	LSE-D ↓	LSE-C ↑
DINet (AAAI 23 [6])	8.5031	5.6956	8.2038	5.1134
TalkLip (CVPR 23 [7])	8.7615	5.7449	8.7019	5.5359
IP-LAP (CVPR 23 [9])	9.8037	3.8578	9.1102	4.389
GeneFace (ICLR 23 [21])	9.5451	4.2933	9.6675	3.7342
ER-NeRF (ICCV 23 [22])	11.813	2.4076	10.7338	3.0242
TalkingGaussian (ECCV 24 [28])	9.3027	4.8452	9.699	4.2032
GaussianTalker (ACM MM 24 [65])	10.1228	4.2625	10.0872	3.8152
SyncTalk++ (Ours)	**8.0808**	**6.4633**	**8.0217**	**6.0733**

相关成果以通讯作者发表在PATTERN ANALYSIS AND MACHINE INTELLIGENCE上

1.3 研究内容：2D高清说话人生成

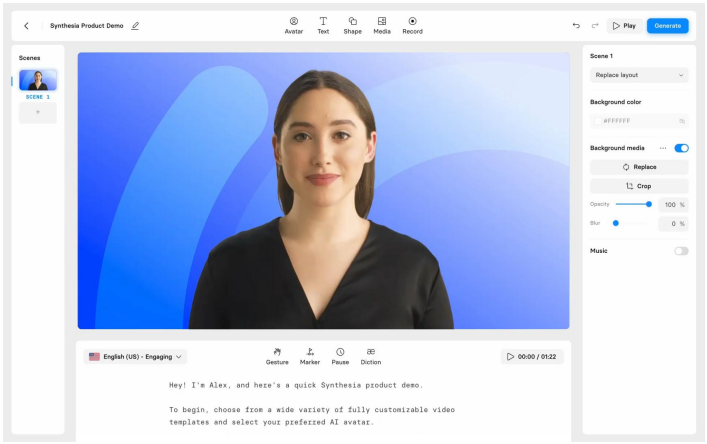
问题背景

现有的技术范式由于底层技术原因无法很好地拟合人物视频素材中的高频人脸细节，同时音频泛化性差。

嘴部细节模糊

研究内容

- 通过将高清的数字人视频生成问题转化为人脸修复问题
- 研究如何从有限的视频素材中表达人脸的高频细节同时进行时序一致地修复



音唇不匹配

1.3 创新点：基于李普希思连续性的潜空间去噪技术

建模现有模型的噪声，基于神经网络的李普希思连续性提出噪声增强损失

传统： 端到端驱动

瓶颈： 表达能力受限，生成内容缺失高频细节

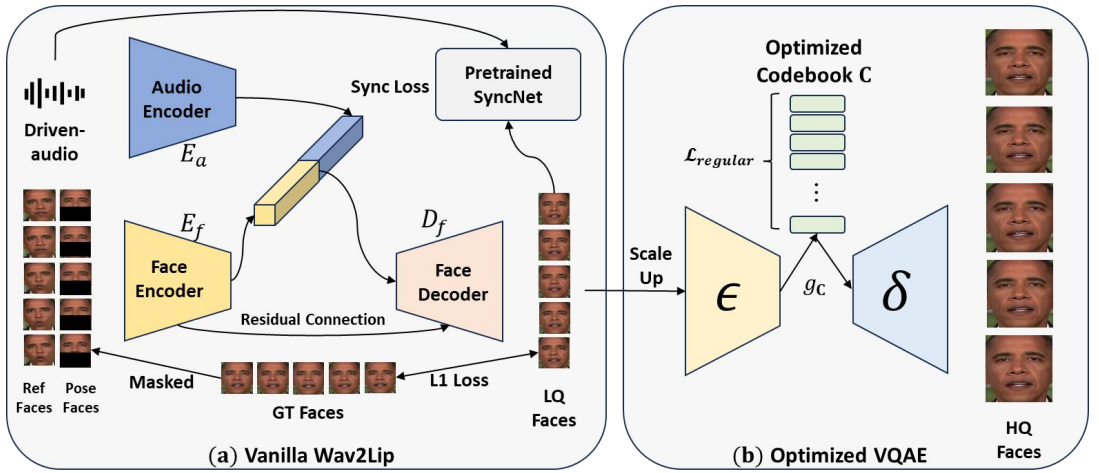
音唇对齐低清语义生成　　时序一致高清细节还原

创新

- 两阶段驱动方法
- 时序一致性去噪

优势

- 超高清内容生成
- 对音频泛化能力强
- 实时推理

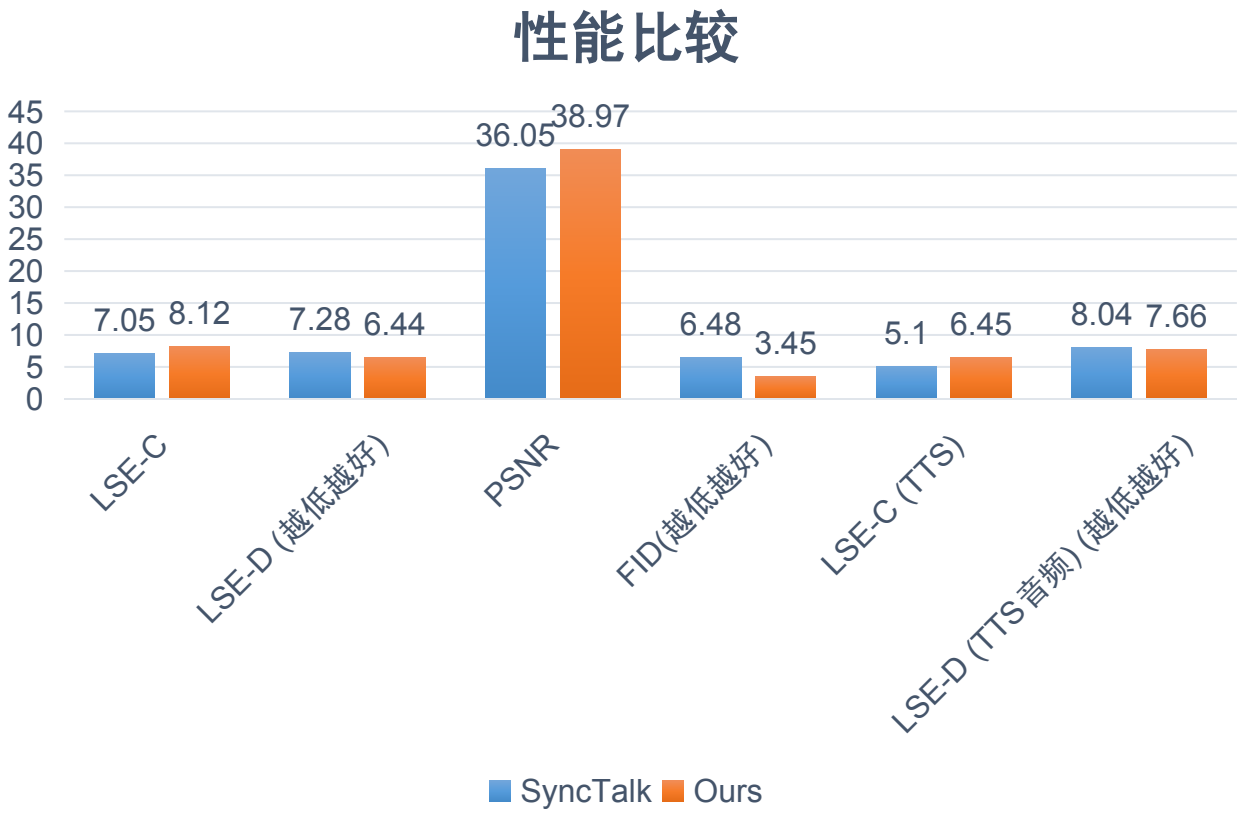


输入音频和参考视频　　潜空间语义对比学习　　音画解耦生成　　潜空间时序一致性去噪

1.3 创新点：基于李普希思连续性的潜空间去噪技术

音频驱动性能评估

相较于中国人民大学的此前最优方法SyncTalk，在原声驱动下，所有指标上均实现大明显领先；在TTS跨音频上的表现也实现明显领先。



相关成果以通讯作者投稿在IEEE VR 2025

2.个性化与情绪3D数字人

2.1 研究内容：3D说话人生成

问题背景

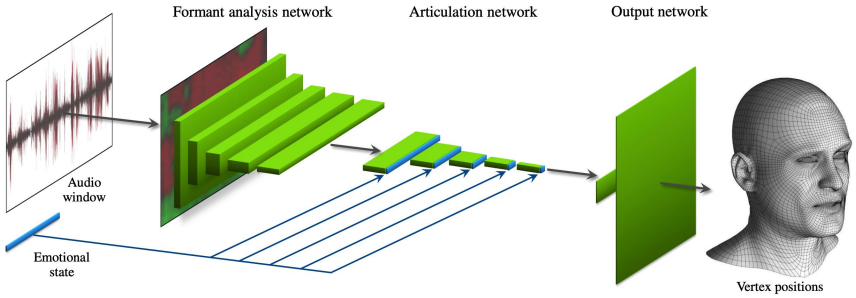
现有方法高度依赖数据体量，不同模态的映射很难精准，唇部的可理解性较差。领域训练数据稀缺。

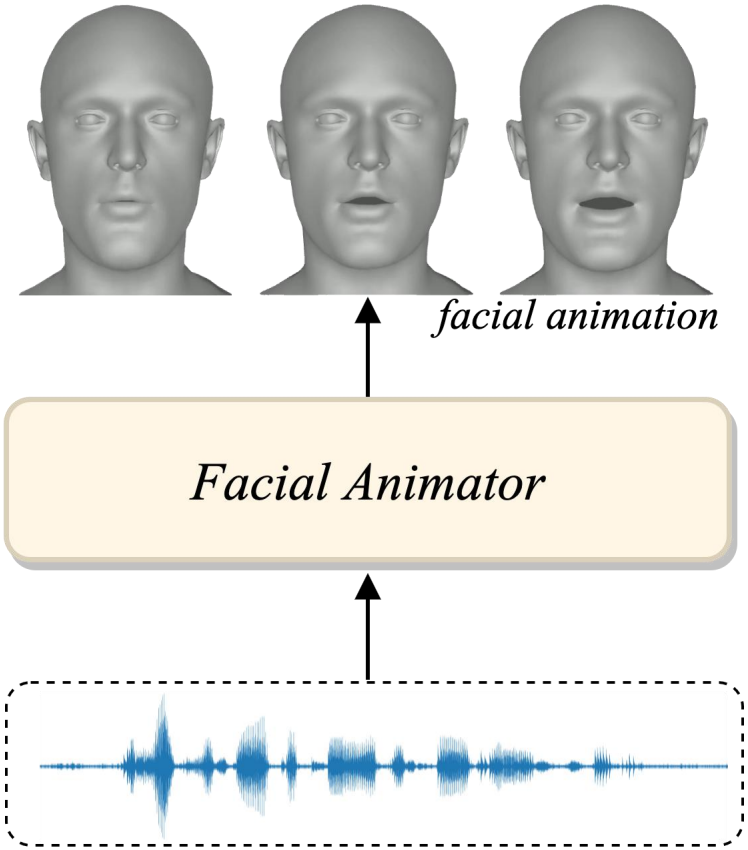
研究内容

构建一个交换训练图：

- 使模块能够在跨模态中学习隐特征之间的复杂联系
- 减少对标记数据的依赖

提升唇部运动的准确性





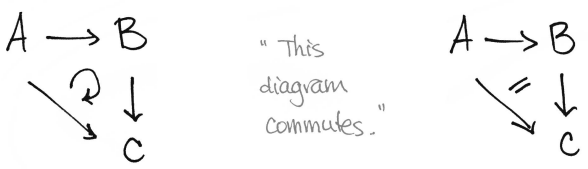
2.1 创新点：基于Commutative Diagram的交换训练图

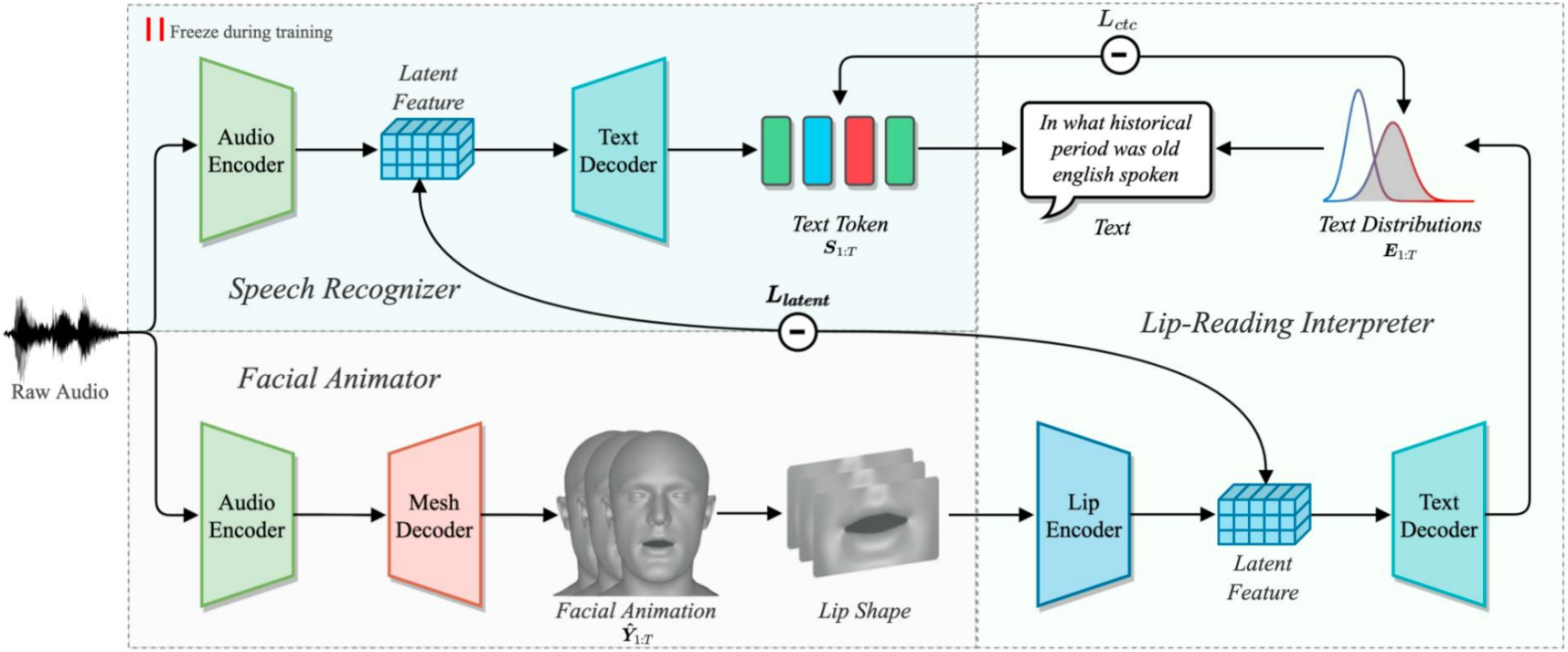
借鉴Commutative Diagram思想，构建交换训练图

传统： 过分依赖数据体量

瓶颈： 跨模态特征学习

创新： 采用类似于中文成语中的“殊途同归”的思想





2.1 创新点：基于Commutative Diagram的交换训练图

在多个数据集上的效果对比

Table 1: Quantitative evaluation results on VOCA-Test.

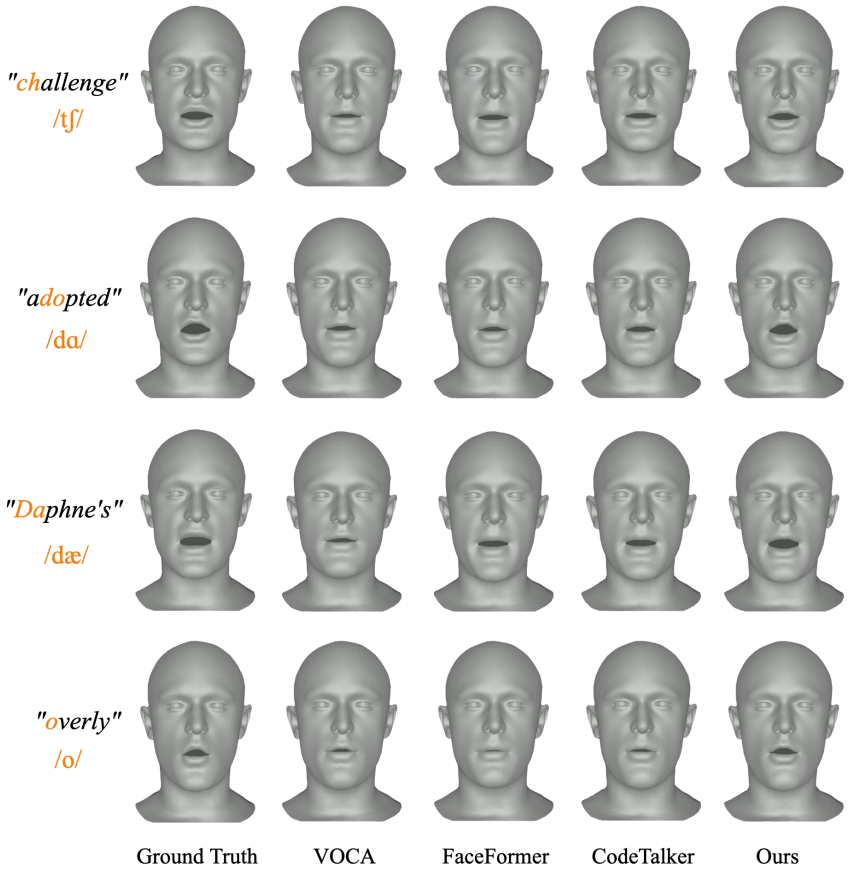
Methods	LVE↓ ($\times 10^{-5}$ mm)	FDD↓ ($\times 10^{-7}$ mm)	LRP↑
VOCA [8]	4.9245	4.8447	72.67%
MeshTalk [41]	4.5441	5.2062	79.64%
FaceFormer [13]	4.1090	4.6675	88.90%
CodeTalker [58]	3.9445	4.5422	86.30%
SelfTalk (Ours)	**3.2238**	**4.0912**	**91.37%**

Table 2: Quantitative evaluation results on BIWI-Test-A.

Methods	LVE↓ ($\times 10^{-4}$ mm)	FDD↓ ($\times 10^{-5}$ mm)	LRP↑
VOCA [8]	6.5563	8.1816	73.83%
MeshTalk [41]	5.9181	5.1025	80.97%
FaceFormer [13]	5.3077	4.6408	83.15%
CodeTalker [58]	4.7914	4.1170	84.62%
SelfTalk (Ours)	**4.2485**	**3.5761**	**88.31%**

SelfTalk在多个数据集上都取得了SOTA的结果

口型渲染对比



我们的方法在唇部运动方面具有更加准确和清晰的运动。

相关成果以通讯作者发表在ACM MM 2023上

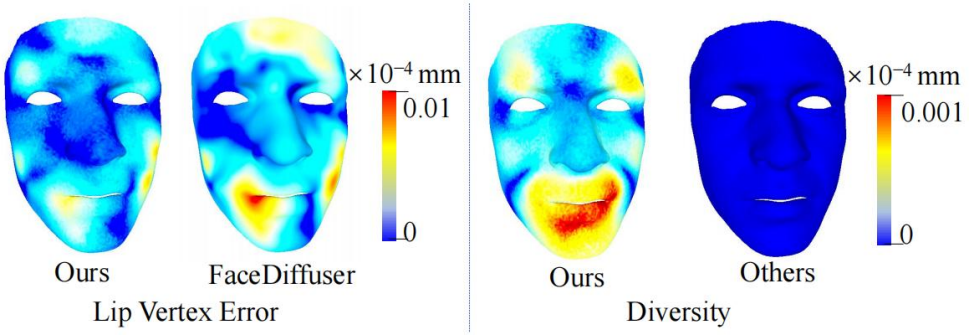
2.2 研究内容：精准3D说话人生成

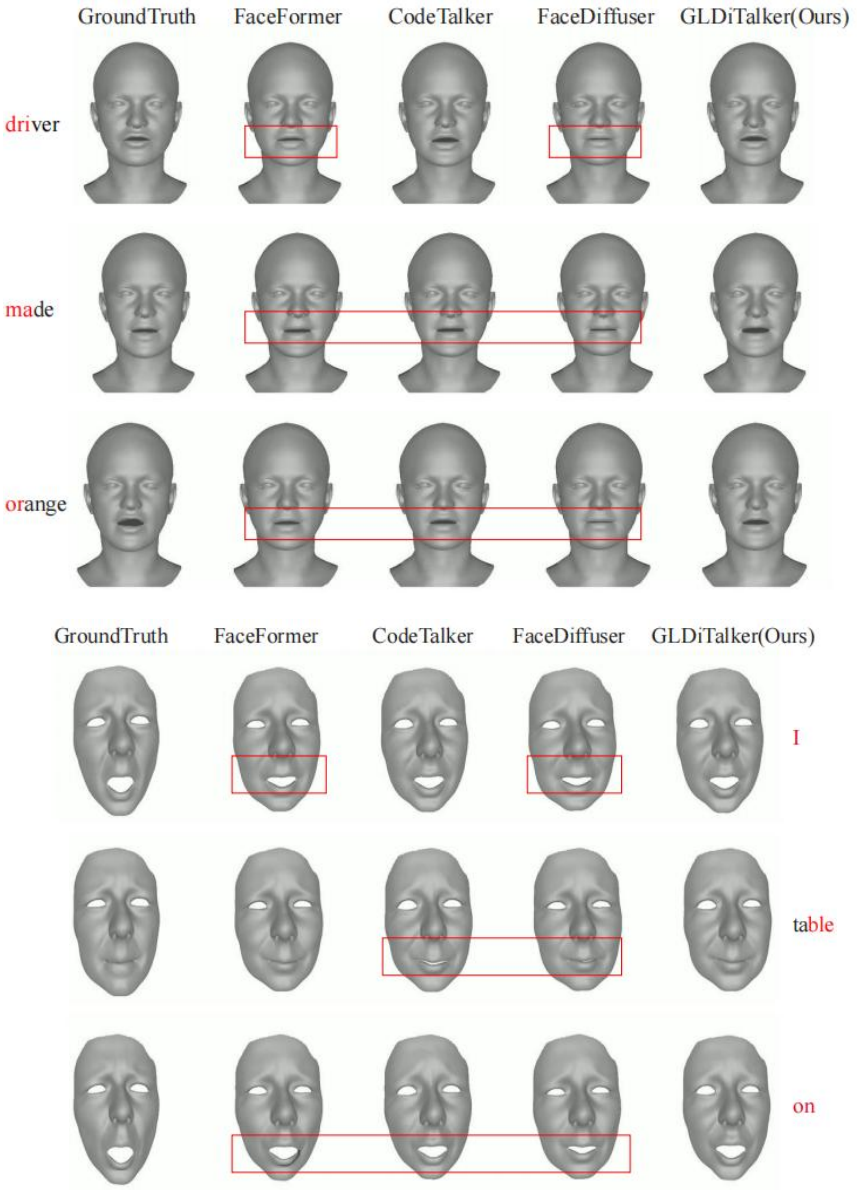
问题背景

现有大多数方法仍然存在模态不一致的问题，特别是音频与Mesh模态之间的不匹配，导致了运动多样性和音唇同步准确性方面的不一致。

研究内容

- 提出了首个量化时空Diffusion训练管线，用于语音驱动的3D面部动画，以减少音频与网格模态之间的不匹配，同时确保面部动作的多样性。
- 提出了图增强量化时空学习阶段，该阶段考虑了顶点之间的连接性，将面部动作转换到时空量化潜在空间，然后通过时空增强的潜在扩散阶段实现多样化的高质量跨模态生成。





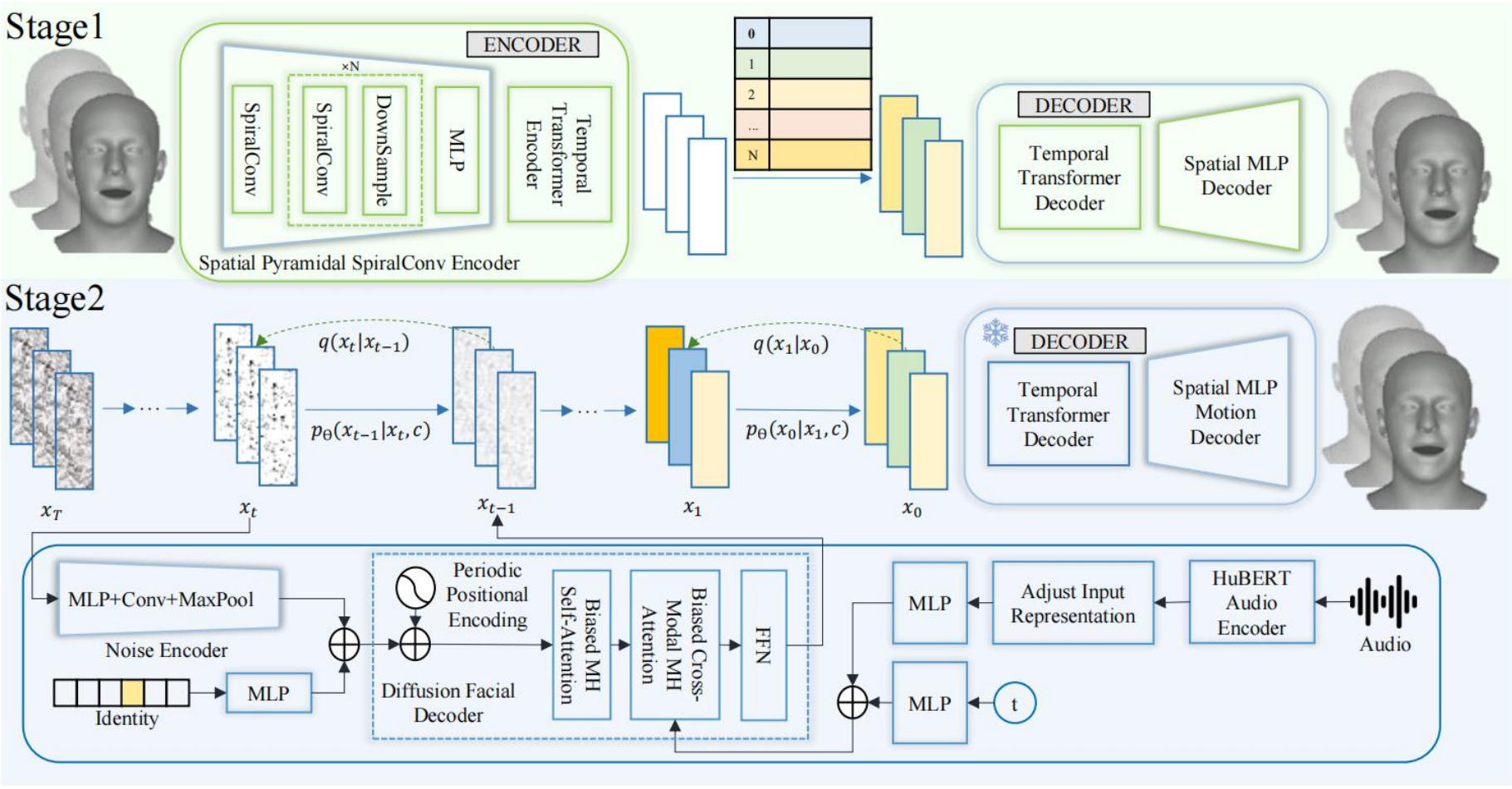
2.2 创新点：GLDiTalker

提出GLDiTalker，实现更精准的音唇同步和更多样的面部运动

传统： 模态对齐存在困难

瓶颈： 音唇同步和面部表情僵硬

创新： 设计了量化化时空Diffusion训练管线，在获得更好的音唇同步准确性的同时生成更加多样的面部运动。



2.2 创新点：GLDiTalker

在BIWI数据集上和其他方法对比：

Methods	LVE ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)	Diversity ↑ ($\times 10^{-4}$mm))
VOCA	6.5563	8.1816	0
MeshTalk	5.9181	5.1025	0
FaceFormer	5.3077	4.6408	0
CodeTalker	4.7914	4.1170	0
FaceDiffuser	4.7823	3.9225	5.6421×10^{-5}
GLDiTalker	**4.6440**	**3.8474**	**8.2176**

Table 1: Quantitative evaluations on BIWI-Test-A.

各个模块消融实验对比：

Methods	LVE ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)	Diversity ↑ ($\times 10^{-4}$mm)
w/o Graph Enhanced Quantized Space Learning Stage	4.7823	3.9225	5.6421×10^{-5}
Spatial MLP Encoder	4.8909	4.3384	8.4251
Spatial Pyramidal SpiralConv Encoder	**4.6440**	**3.8474**	**8.2176**

Table 3: Ablation study for Graph Enhanced Quantized Space Learning Stage on BIWI-Test-A.

主观指标-MOS打分和其他方法对比：

Methods	VOCASET-Test		BIWI-Test-B	
	Realism↑	LipSync↑	Realism↑	LipSync↑
FaceFormer	3.02	3.08	3.38	3.50
CodeTalker	3.46	3.37	3.56	3.45
FaceDiffuser	2.55	2.47	3.04	2.97
GLDiTalker	**3.94**	**4.05**	**3.85**	**4.00**

Table 2: User study results.

各方法面部偏移量标准差热图对比：

1.GLDiTalker比现有方法，具备**更好的多样性**。
2.LVE和FDD指标在BIWI数据集上得到**明显提升**。
3.主观MOS打分上，GLDiTalker相比其他方法在**真实性和音唇同步表现俱佳**。

2.3 研究内容：3D说话人情感构建

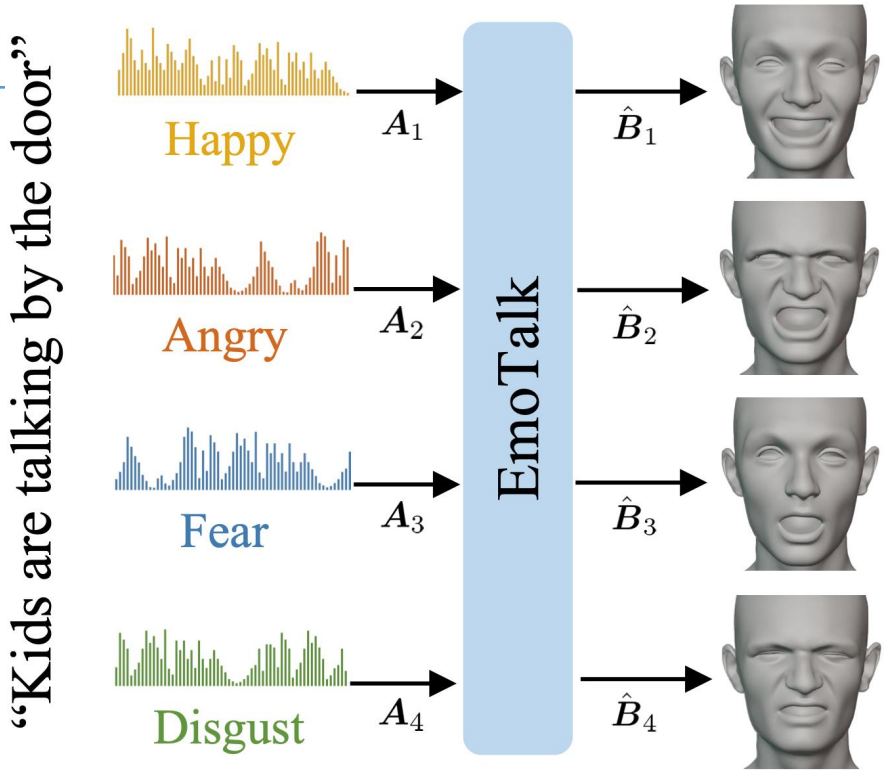
问题背景

现有的语音生成模型不能对语音中的情感变化做出很好的反映。

研究内容

情感构建：

- 如何解耦提取文本内容和情绪内容
- 如何获取更具有情感的数据，构建包含情感的高质量3D数据



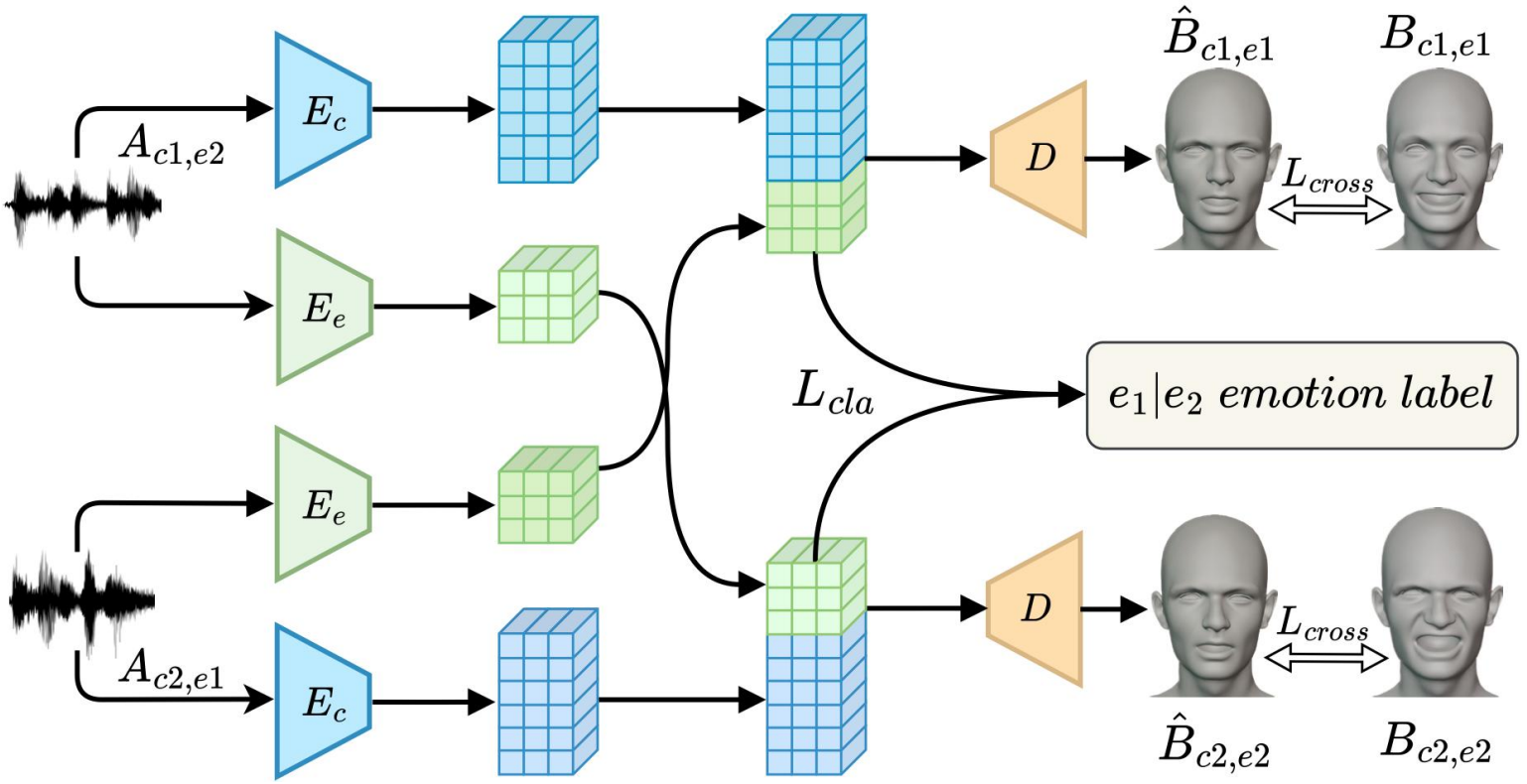
2.3 创新点：数据构造与情感学习

基于数据集的构造和情绪编码与预测，实现情绪生成

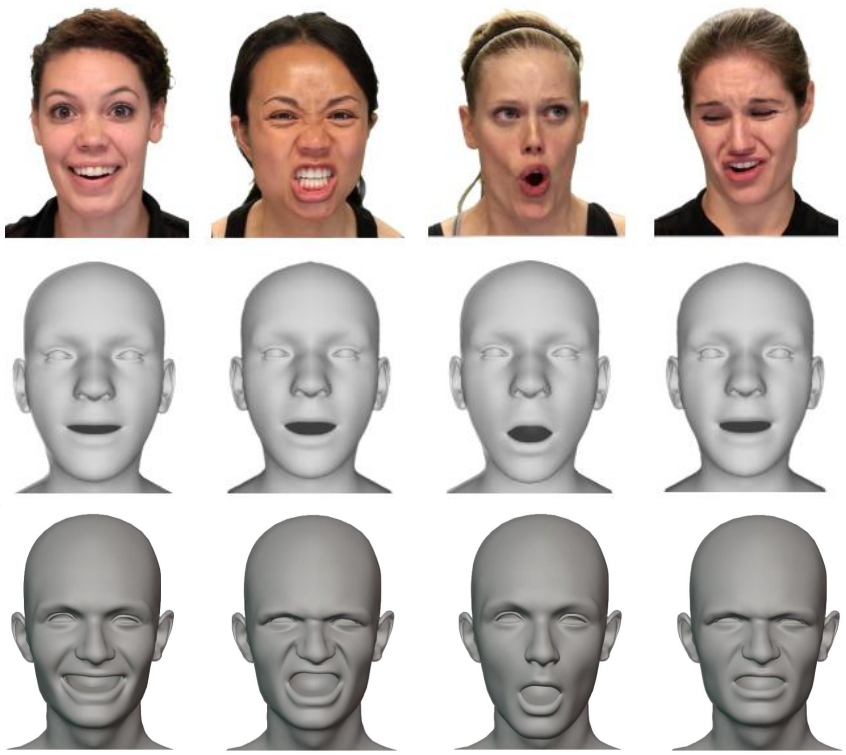
传统： 未对情感变化进行构建

瓶颈： 数据集构建和情绪编码

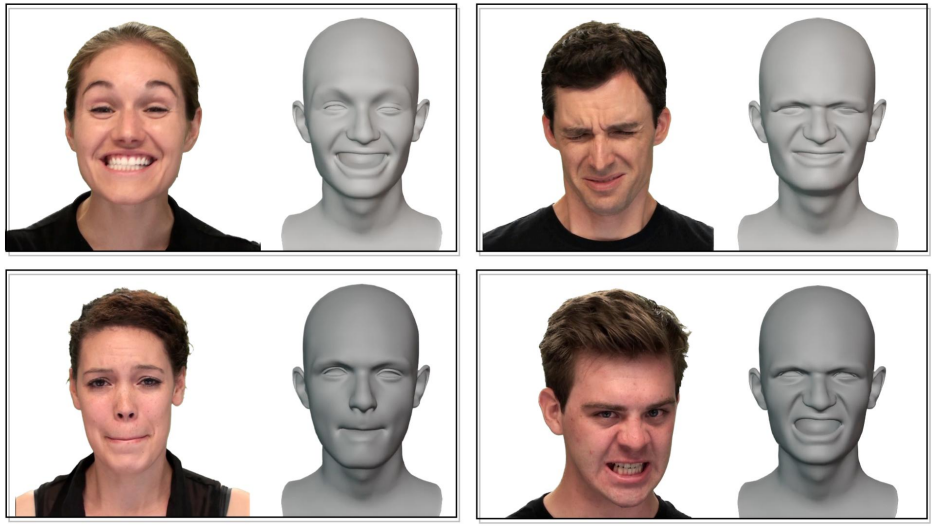
创新： 从语音中解耦情绪特征和内容特征。



2.3 创新点：数据构造与情感学习



通过自建数据集，我们的方法能够实现
富有情绪的面部渲染输出



	RAVDESS (emotion)		HDTF (no emotion)	
Method	LVE(mm)↓	EVE(mm)↓	LVE(mm)↓	EVE(mm)↓
VOCA	5.091	4.188	4.447	3.286
MeshTalk	3.459	3.386	3.886	3.124
FaceFormer	3.247	3.757	3.374	3.142
Ours	**2.762**	**2.493**	**2.892**	**2.364**

Method	LVE(mm)↓	Train on VOCASET
VOCA	4.704	✓
MeshTalk	4.513	✓
FaceFormer	4.418	✓
Ours	**4.134**	✗

相关成果以通讯作者发表在ICCV 2023上

2.4 研究内容：基于小样本的3D说话人生成

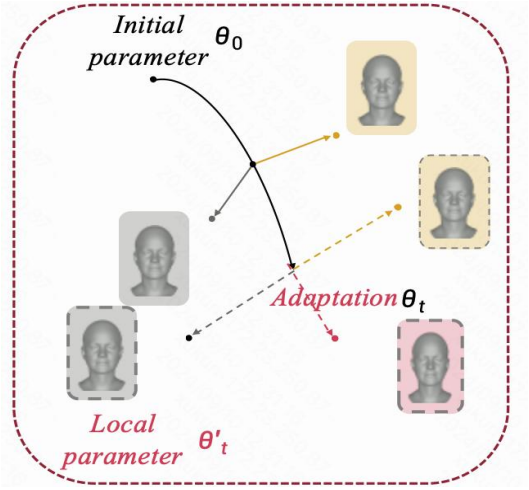
问题背景

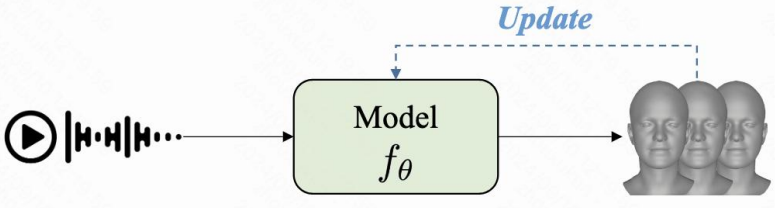
数据采集难度大；数据质量要求高。
难以针对于每一个人都采集足量人脸数据。

研究内容

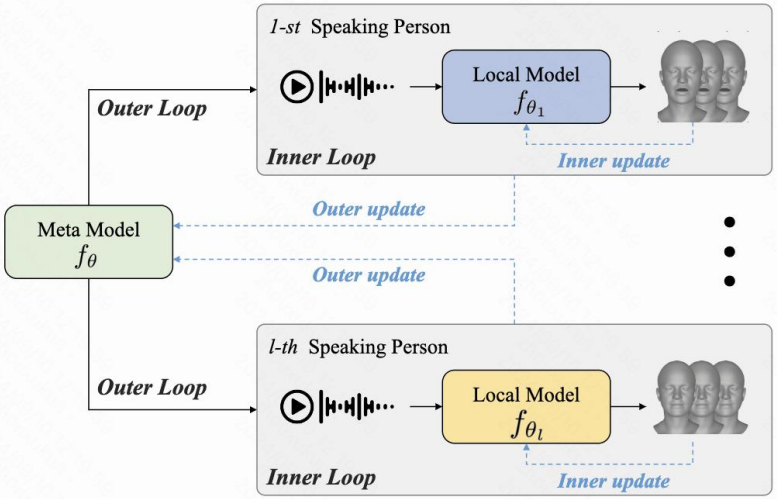
设计小样本学习方法：

- 从大量人脸中学习一个对于适应最优的人脸
- 对于小样本进行快速适应优化
- 利用神经过程构建样本以及对于未见过样本的预期





(a) Speaking style adaptation mode of existing methods.



(b) Speaking style adaptation mode of MetaFace.

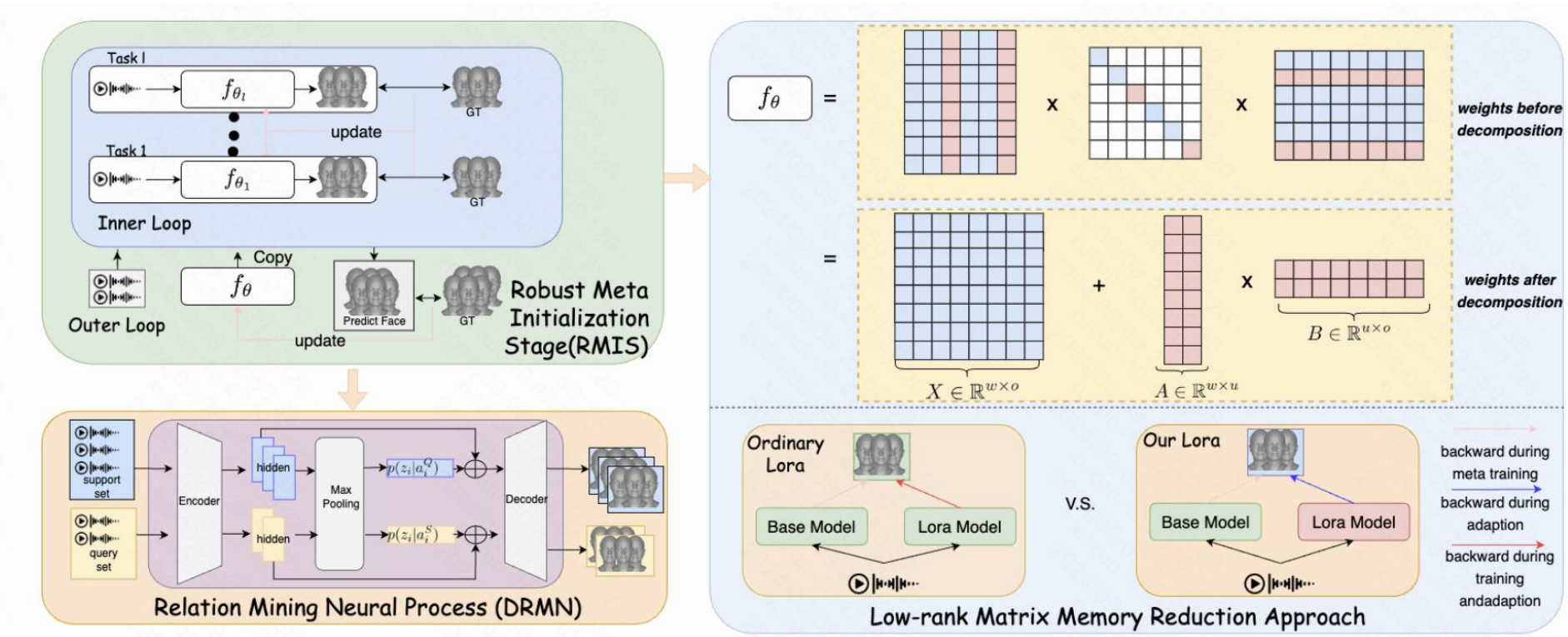
2.4 创新点：利用小样本学习快速适应

提出了小样本学习方法，克服了对数据的过度依赖

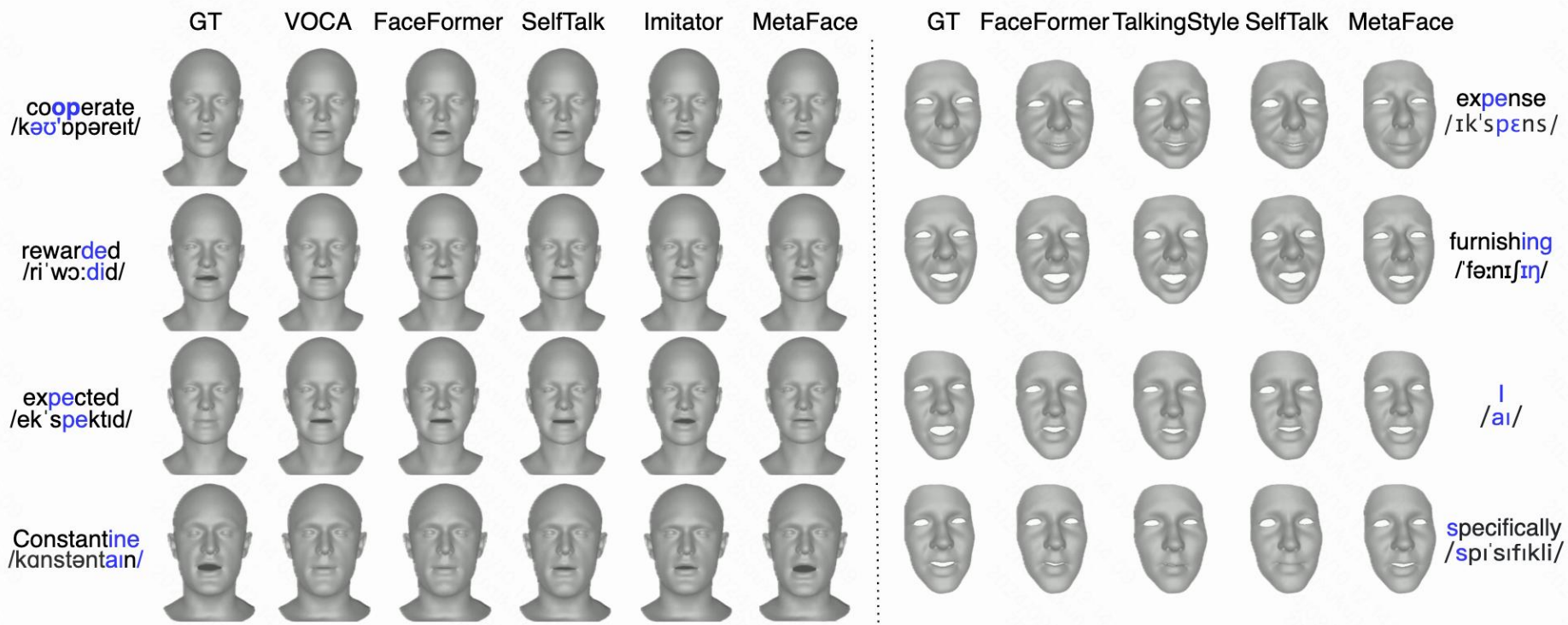
传统： 数据采集难度大

瓶颈： 构建样本预期

创新： 设计小样本学习方法，利用神经过程构建样本以及对于未见过样本的预期。



2.4 创新点：利用小样本学习快速适应



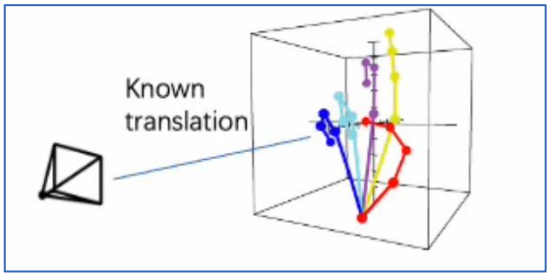
Method	VOCASet				BIWI			
	$L2_{Face}\downarrow$	$L2_{lip}\downarrow$	$L2_{max}\downarrow$	$lip_{sync}\downarrow$	$L2_{face}\downarrow$	$L2_{lip}\downarrow$	$L2_{max}\downarrow$	$lip_{sync}\downarrow$
VOCA(Cudeiro et al. 2019)	7.02	7.84	10.26	7.23	29.20	30.28	77.28	30.36
FaceFormer(Fan et al. 2022)	1.08	5.18	9.96	4.00	11.92	12.71	35.44	12.43
FaceFormer(Fan et al. 2022)†	0.85	3.24	6.62	2.95	11.56	12.33	33.94	12.05
SelfTalk(Peng et al. 2023a)	1.07	2.74	7.06	2.54	11.65	12.52	33.53	12.23
SelfTalk(Peng et al. 2023a)†	0.82	2.61	6.02	2.53	10.52	11.06	31.52	11.30
StyleTalk(Song et al. 2024)	0.95	4.22	8.37	3.04	13.19	13.66	37.57	13.64
StyleTalk(Song et al. 2024)†	0.89	3.71	7.66	2.65	12.42	12.54	36.18	12.54
Imitator(Thambiraja et al. 2023)†	0.90	2.09	5.28	1.72	-	-	-	-
Ours†	**0.62**	**1.86**	**4.43**	**1.56**	**9.15**	**9.81**	**26.33**	**9.49**

3.人体重建技术探索与应用

3.1 研究内容：精准实时多目人手重建

问题背景

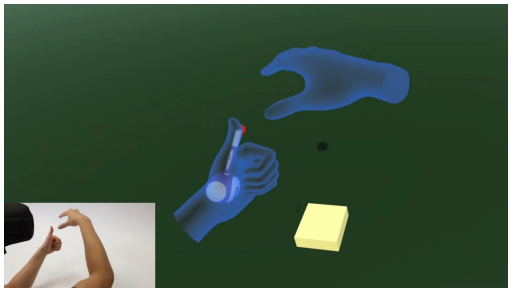
单目视觉场景下的人手动捕由于深度模糊性、人手自遮挡、手物遮挡等问题，导致精度低，无法支撑真实感的人机交互体验。

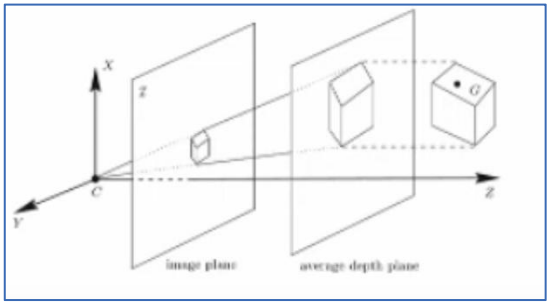


深度模糊性

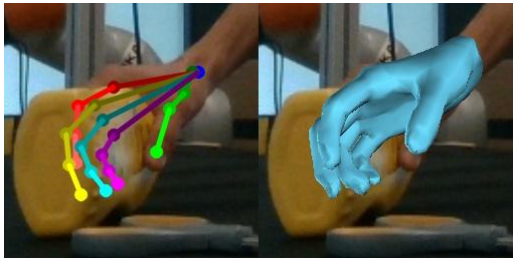
研究内容

- 通过有效融合多视图信息，解决遮挡和深度模糊问题提高重建精度
- 通过对人手几何的高效建模，在保证精度的同时实现实时推理，满足实际需求



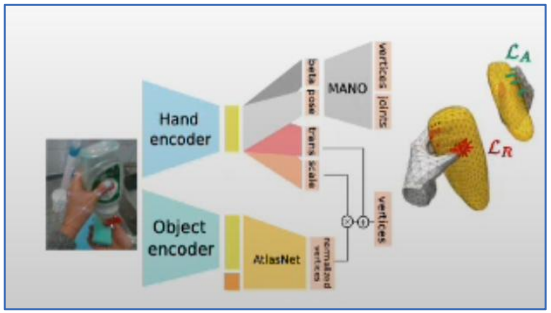


弱透视模型



遮挡视角

非遮挡视角



手-物遮挡

3.1 创新点：人手铰接几何建模与多视图信息融合

提出了几何解耦的人手几何模型，实现精准实时多目人手重建

传统： 单目几何重建　复杂特征交互

瓶颈： 立体信息缺失　推理性能低效

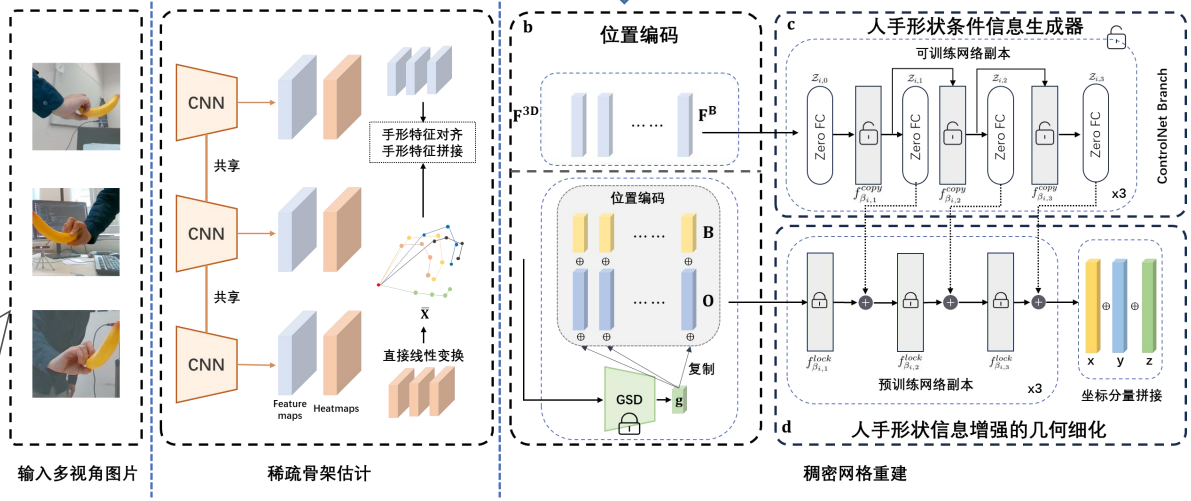
多视几何　铰接建模　多视信息增强

创新

- 多视几何信息融合
- 人手几何铰接建模

优势

- 高效实时推理
- 跨视角信息互补
- 铰接几何精准建模

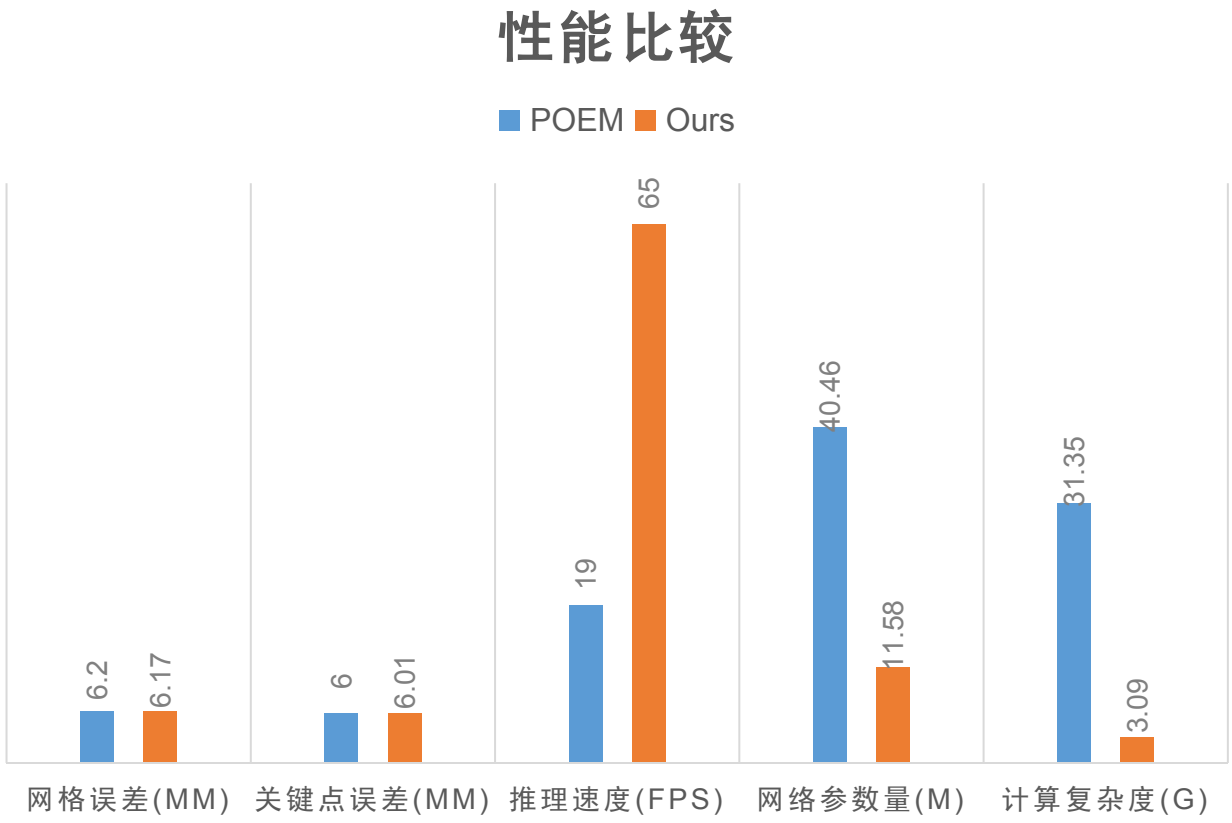


输入多视图　多视几何估计　铰接几何建模　多视几何信息融合

3.1 创新点：人手铰接几何建模与多视图信息融合

多视图条件下人手几何重建性能

在多种多目相机设置下，相较于上海交通大学的此前最优方法POEM，相对精度相近，但是推理速度提升2倍，计算复杂度减少90%，参数量减少75%。



相关成果以通讯作者发表在ECCV 2024

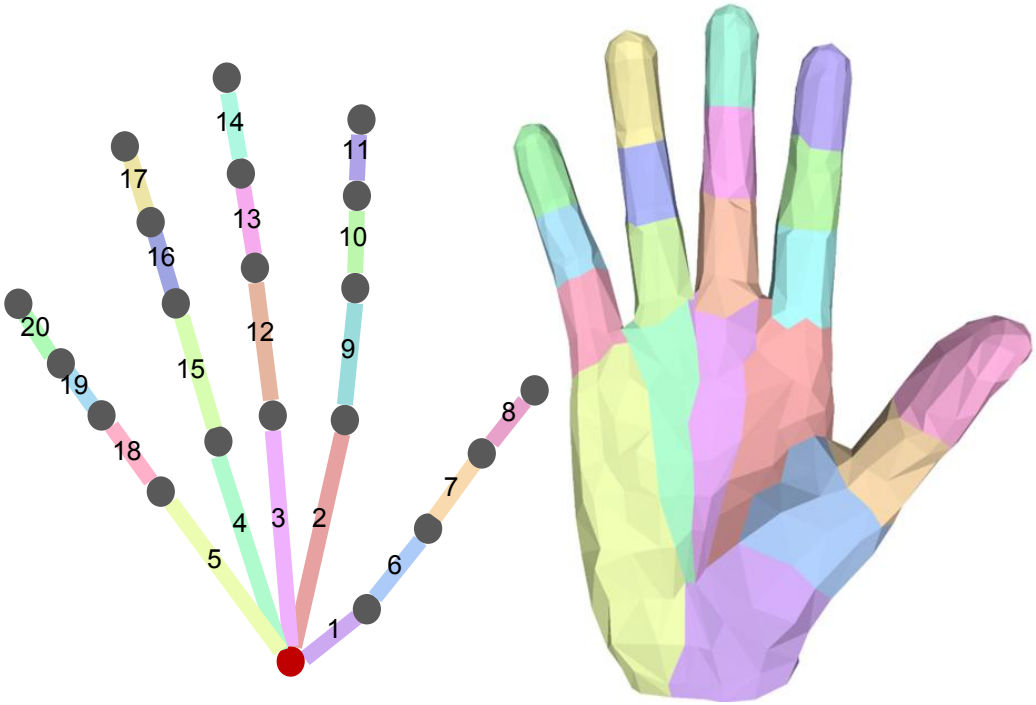
3.2 研究内容：高效实时多目人手重建

问题背景

现有多视图3D手部网格重建方法存在模型复杂，重建速度慢的问题。

研究内容

平衡模型复杂度和准确度，构建高效的多视图3D手部网格重建方法，实现实时重建。



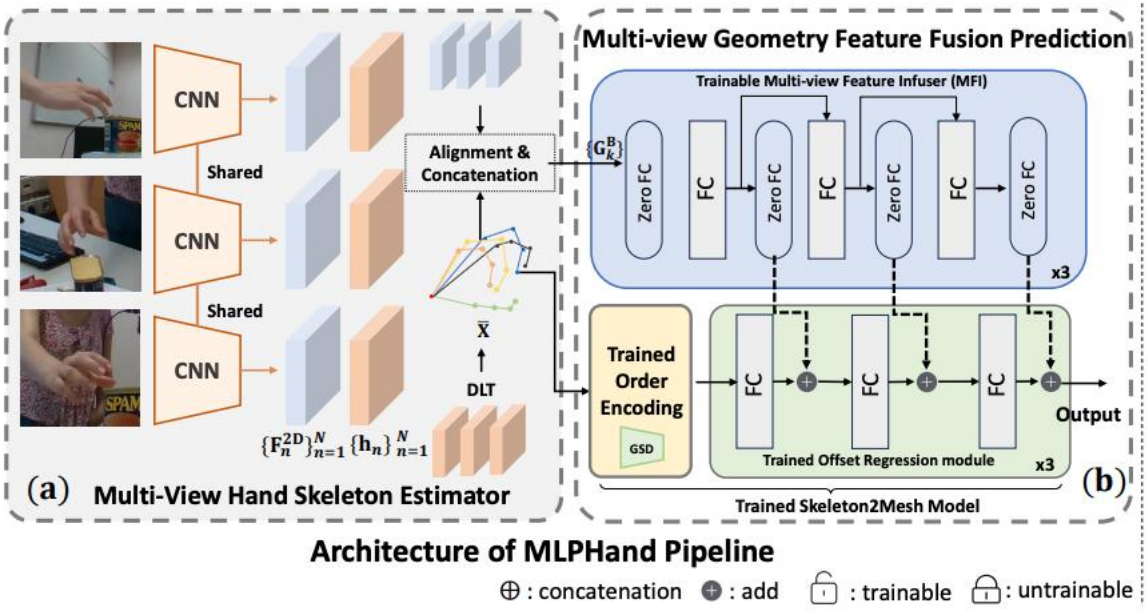
3.2 创新点：轻量化MLP-based Skeleton2Mesh方法

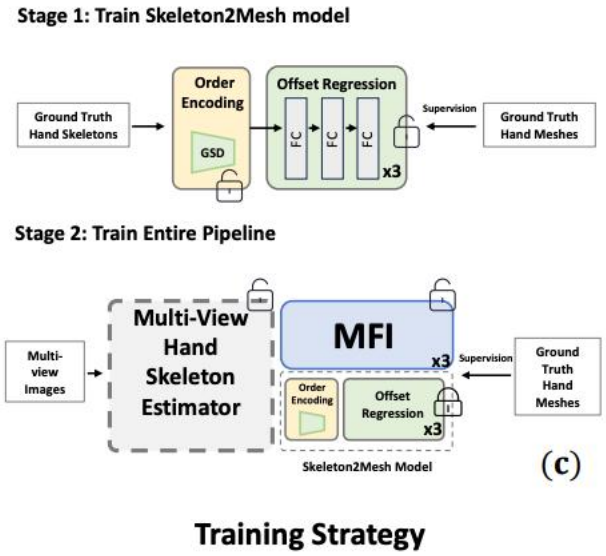
提出了轻量化方法，实现高效多视图3D手部网格的实时重建

传统： 复杂模型重建速度慢

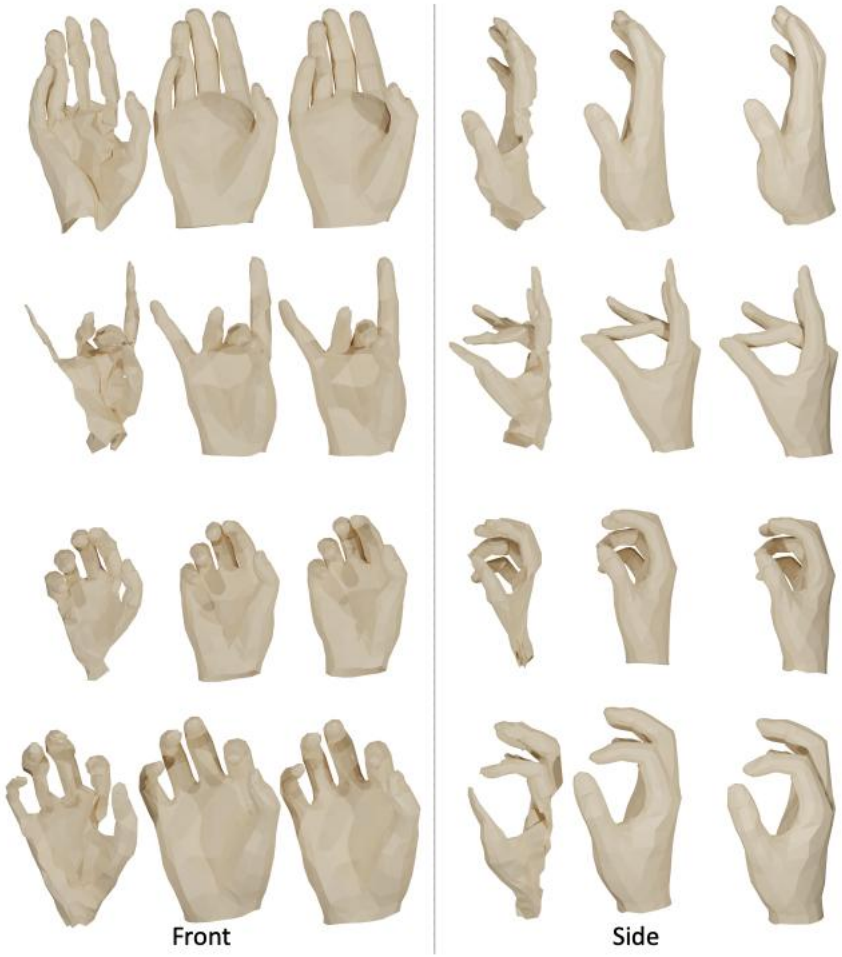
瓶颈： 重建速度与准确度的平衡

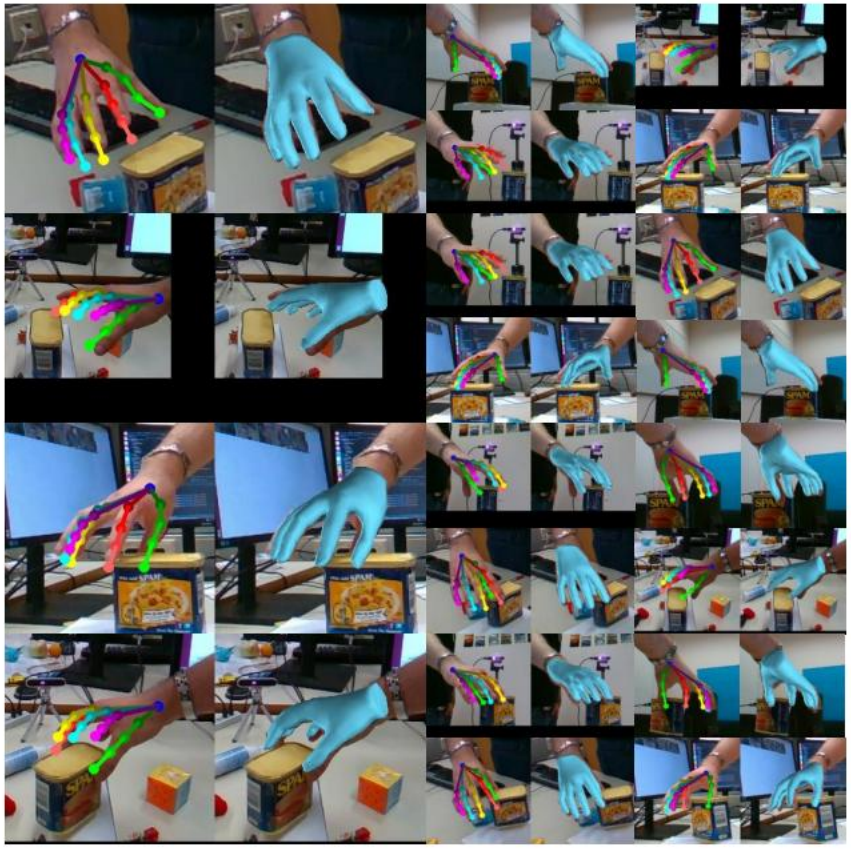
创新： 构建轻量化的MLP-based Skeleton2Mesh方法，有效加速重建速度





3.2 创新点：轻量化MLP-based Skeleton2Mesh方法



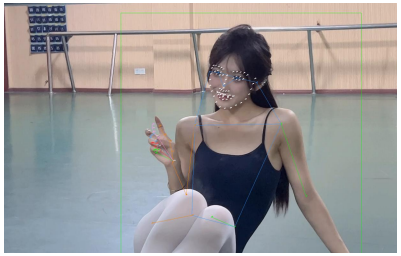


相关成果以通讯作者发表在ECCV 2024上

4.1 研究内容：单视角单目人体重建

问题背景

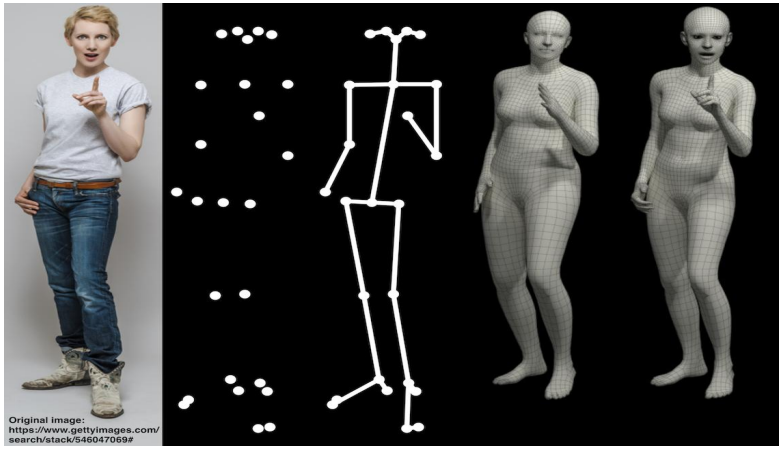
现有方法存在数据集不足、训练场景不足、训练动作多样性不足的问题。

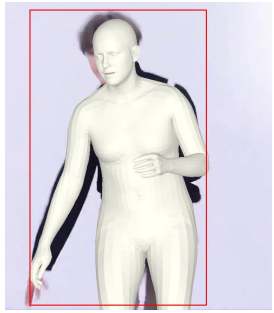


研究内容

人体动作估计：

- 有效融合多个模态的伪标签实现高效、准确标注
- 高效驱动3d人体动作
- 减轻对于制作动画、虚拟人过程中的数据采集、穿戴设备的工作量





4.1 创新点：构建了大规模标注数据方法

提出了伪标签标注算法，并且构建了大规模数据集

传统： 缺少大规模标注数据

瓶颈： 标注效率低下

创新： 提出了伪标签标注算法，减轻对于制作动画、虚拟人过程中的数据采集、穿戴设备的工作量。构建了大规模丰富数据集。

4.1 创新点：构建了大规模标注数据方法

Type	Dataset	Annotation Part	Annotation Format	Frame Count	Scenes	Back-ground	clothing type	Video	Scene
Rendered	Agora [PHT*21]	Body&Hand &Face	SMPL-X	17*K*	Daily	-	-	N	Virtual
	GTA-Human [CZR*21]	Body	SMPL	1.4*M*	Daily	-	-	N	Virtual
Marker/Sensor-based MoCap	Human36M [IPOS13]	Body	SMPL	3.6*M*	Daily	1	1	Y	Lab
	3DPW [VMHB*18]	Body	SMPL-X	$>51K$	Daily	-	-	Y	Diverse
Marker-less Multi-view MoCap	MPI-INF-3DHP [APGS14]	Body	SMPL-X	$>1.3M$	Daily	1	1	N	Diverse
	EHF [PCG*19]	Body&Hand &Face	SMPL-X	0.1K	Daily	1	1	Y	Lab
	ZJU-MoCap [FSD*21]	Body	SMPL-X	$\geq 237K$	Daily	1	1	Y	Fitting Room
Pseudo-3D Labels	PennAction [ZZD13]	Body	SMPL-X	77*K*	Fitness	-	-	Y	Diverse
	MSCOCO [LMB*15]	Body	SMPL-X	200*K*	Daily	-	-	N	Diverse
	COCO-Wholebody [JXX*20b]	Body&Hand &Face	2D KPT	200*K*	Daily	-	-	N	Diverse
	MPII [MRC*17]	Body	SMPLX	40*K*	Daily	-	-	N	Diverse
	MTP [MOT*21]	Body	SMPLX	3.8*K*	Daily	-	-	N	Diverse
	FBA [RF20a]	Body	SMPLX	13*K*	Vlog& Cook&Daily	-	-	N	Diverse
	Multi-shot-AVA [PMK22]	Body	SMPL	350*K*	Movie	-	-	N	Diverse
	UBody [LZW*23]	Body&Hand &Face	SMPL-X& 2D Kpt	$>1050K$	Daily	15	-	Y	Diverse
	D-Body(Ours)	Body&Hand &Face	SMPL-X& 2D Kpt	$>8M$	Dance & Teaching	986	9	Y	Diverse

4.2 研究内容：穿衣人体3D重建

问题背景

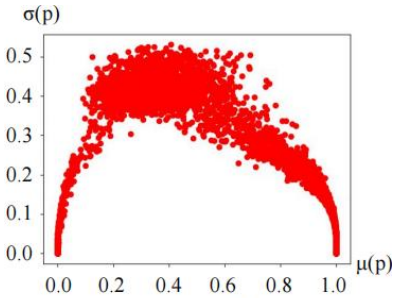
现有方法没有考虑衣服的随机性。而是直接预测每个点的sdf值造成肢体末端和衣服细节差的问题。

研究内容

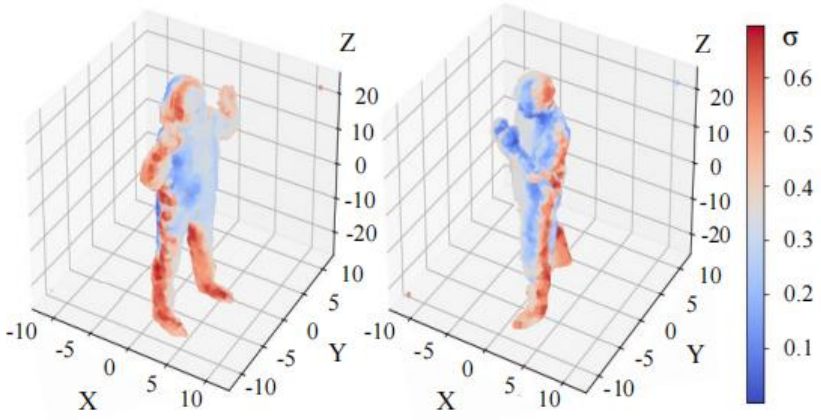
统计得到空间中不同点的随机性：

- 距离人体表面距离不同
- 在人体表面的位置不同

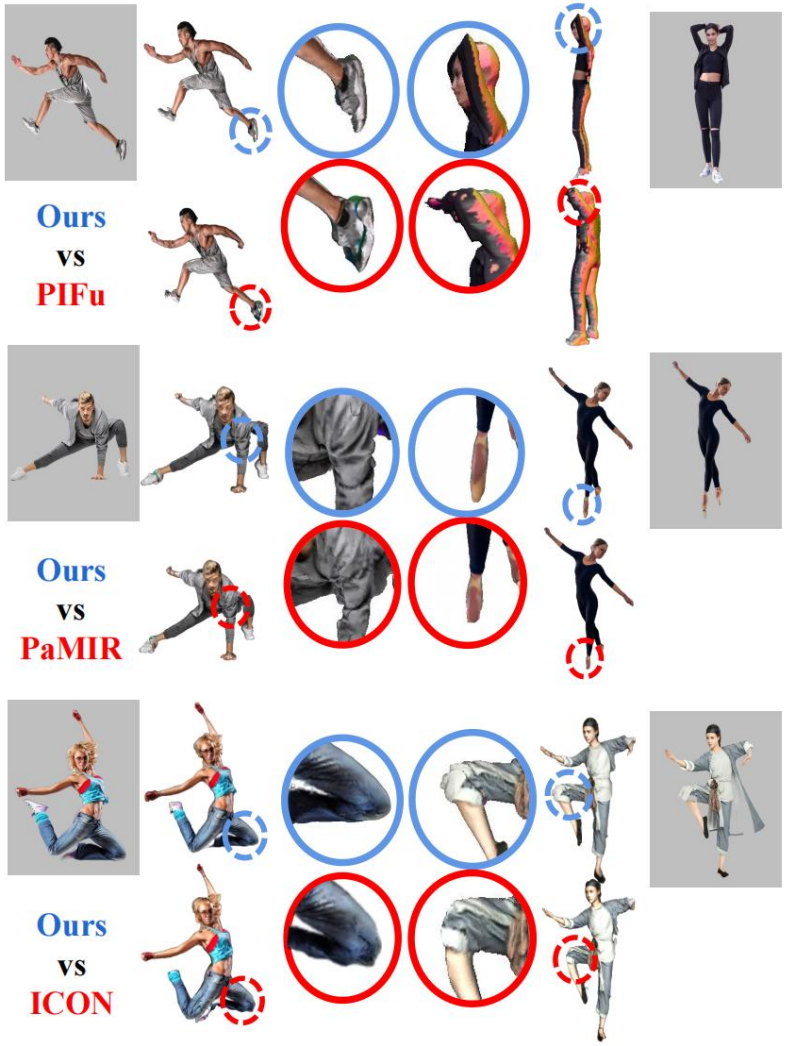
提出分布隐式场 （D-IF），
用于更精细得重建带衣服的3D人体模型



(a) points around the mesh



(b) points on the mesh



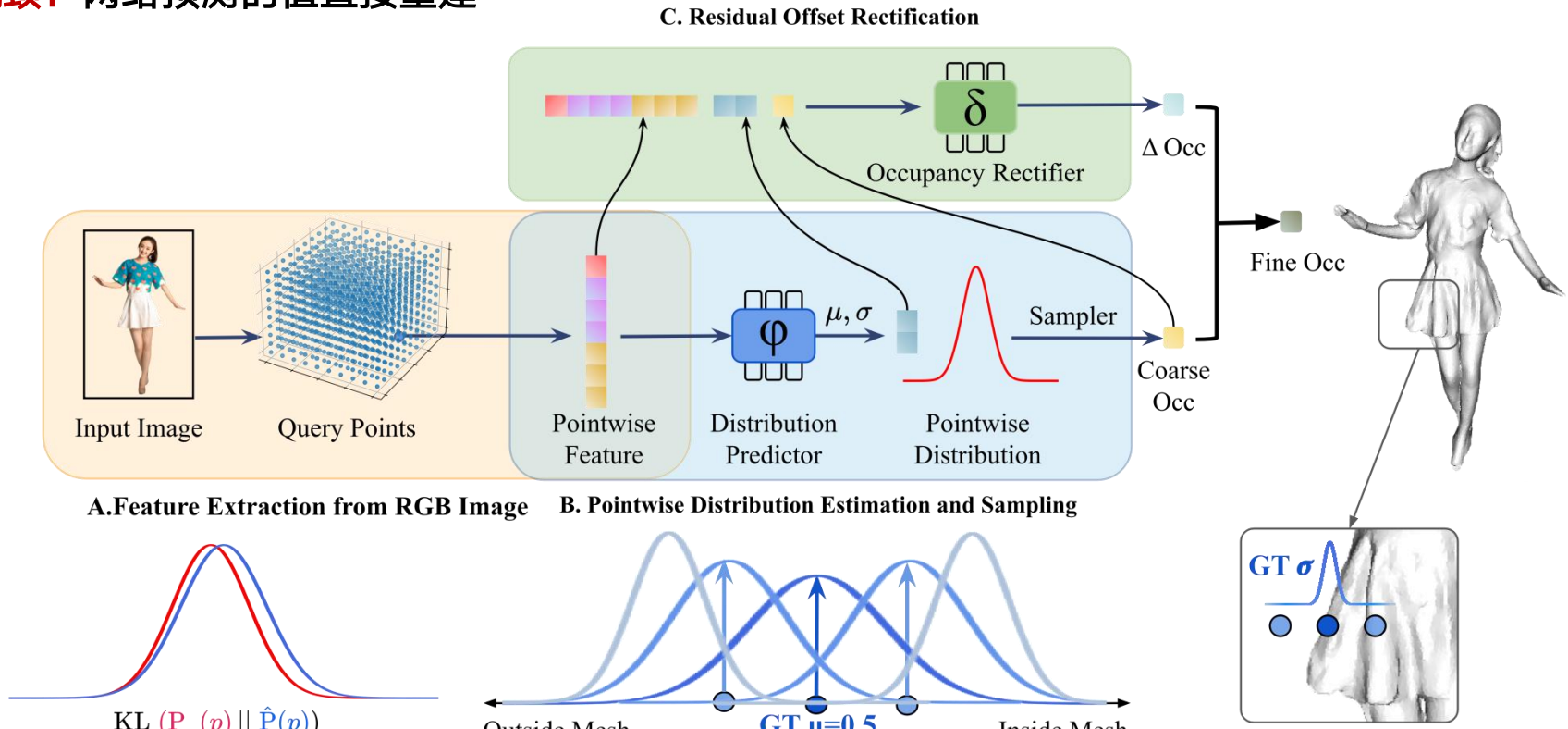
4.2 创新点：分布隐式场重建精细的带衣服人体模型

提出了分布隐式网络（D-IF），实现精细的人体和衣服重建

传统： 未考虑衣服的不确定性

瓶颈： 网络预测的值直接重建

创新： 分布隐式场除了输出点的sdf，还预测了sdf的随机性。进而在预测的分布中采样得到最终结果。



最终根据统计分析的不确定性分布结果对预测的分布进行监督。

4.2 创新点：分布隐式场重建精细的带衣服人体模型

在CAPE数据上和其他方法对比：

Methods	Smooth Occupancy	Uncertainty Dist. Loss	CAPE-FP			CAPE-NFP			CAPE		
			Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓
Ours	✓	✓	0.684	0.677	0.048	0.838	0.821	0.055	0.785	0.771	0.050
PIFu* [37]	×	×	2.525	1.905	0.155	4.143	2.773	0.202	3.603	2.484	0.186
PaMIR* [47]	×	×	1.517	1.331	0.098	1.768	1.450	0.102	1.684	1.410	0.101
ICON [43]	×	×	0.775	0.715	0.054	1.004	0.930	0.063	0.928	0.859	0.060
ECON [42]	×	×	0.912	0.907	0.037	0.926	0.917	0.037	0.921	0.914	0.037

在In-the-Wild数据结果：

我们提出的分布隐式场方法与马普所提出的ICON方法相比，

Chamfer指标在CAPE-NFP的复杂穿衣人体数据集上提升了16.5%。

在全部CAPE数据集上，P2S和Normals指标也分别提升了10.2%和16.6%。

相关成果以通讯作者发表在ICCV 2023上

4.多模态驱动的人体动作生成

4.3 研究内容：多模融合学习态势感知

问题背景

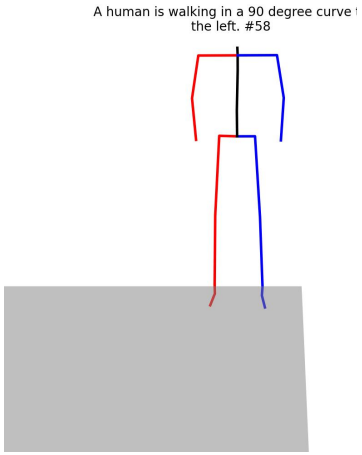
轻量化多模态动作生成框架面临动作生成自然度与流畅度欠缺的问题，依赖单模态模型设计训练模式。

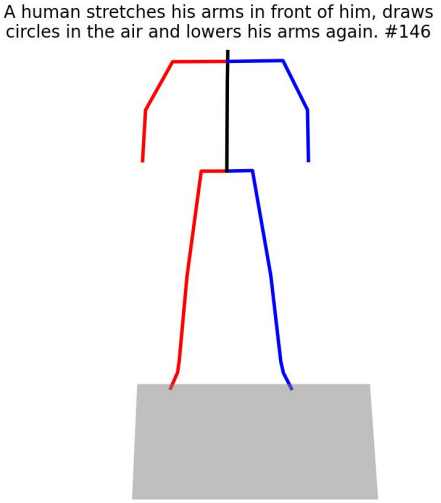
研究内容

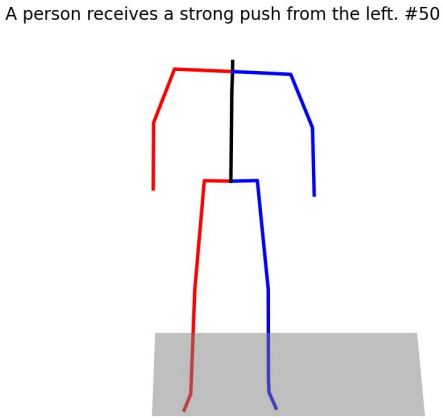
单模型轻量化高质量的人体动作生成：

- 融合单模态预训练的特征提取器
- 基于Transformer和Perceiver的隐空间特征mapping与处理模块
- 大规模预训练的motion prior

达到高质量的生成效果





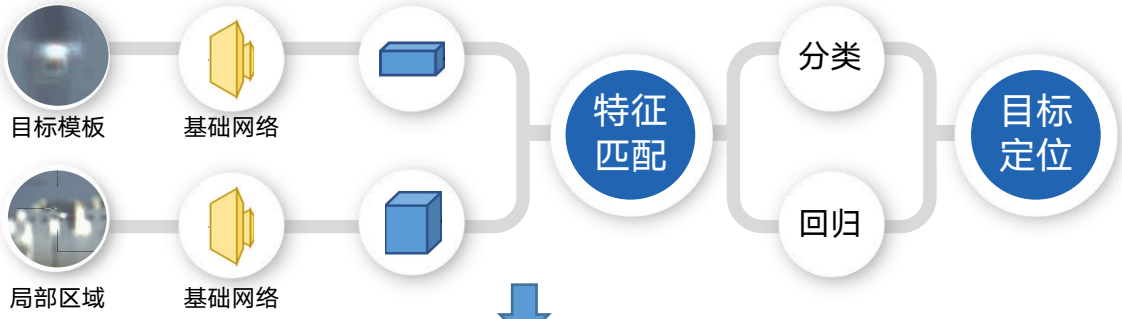


4.3 创新点：隐空间多模态融合架构与Prior加强模块

提出了双流迁移多模融合建模，实现精准鲁棒实例级目标跟踪

传统： 单一生成　固化应用

瓶颈： 统一困难　运用繁琐



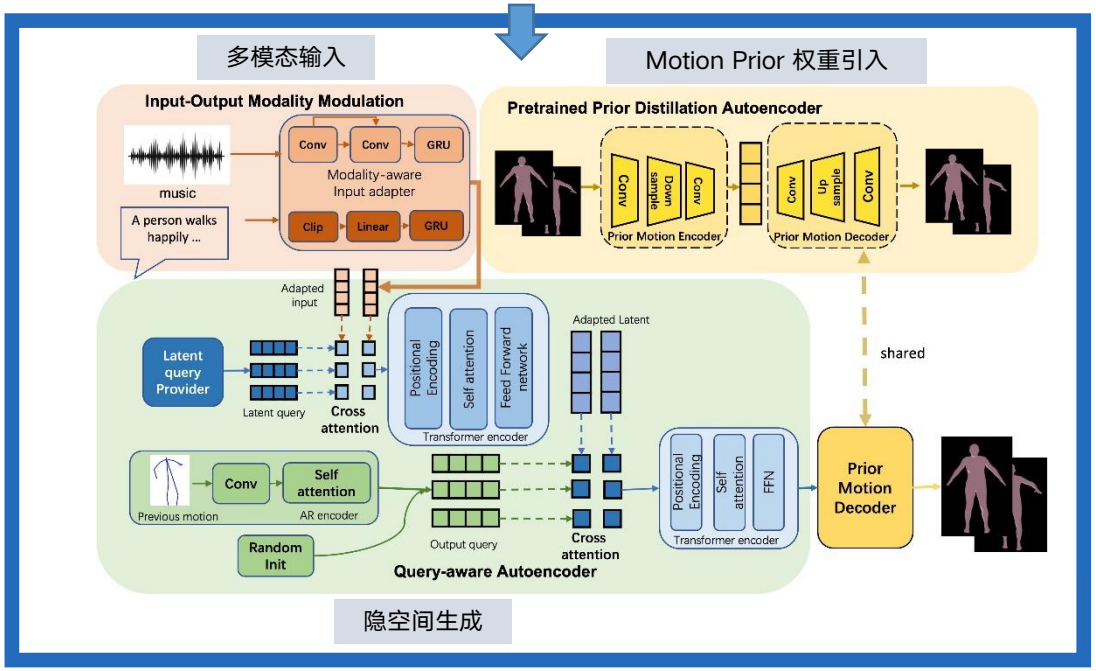
模态统一映射　隐空间生成　Prior知识增强

创新

- 统一模态-联合生成
- 预训练模型能力迁移

优势

- 多阶信息融合互补
- 时空线索联动建模
- 层级知识级联迁移



4.3 创新点：隐空间多模态融合架构与Prior加强模块

KITML基于文字描述的动作数据集

为文字描述引导的多样人体动作生成提供高质量的基准数据集，相较此前南洋理工大学S-lab提出的最优模型MotionDiffuse，FID生成质量指标超幅45.6% 相较此前类似backbone的模型速度提升接近7倍

Method	FID ↓	R precision (Top 1) ↑	R precision (Top 2) ↑	R precision (Top 3) ↑	Diversity ↑
Ground Truth	0.031	0.424	0.649	0.779	11.08
Test2Gesture	12.12	0.156	0.255	0.338	9.334
Language2Pose	6.545	0.221	0.373	0.483	9.037
MoCoGAN	94.41	0.037	0.072	0.106	0.462
Guo et.al	2.770	0.370	0.569	0.693	10.91
Motiondiffuse	1.954	0.417	0.621	0.739	11.10
E2M (Ours)	1.060	0.385	0.574	0.685	11.15

AIST++ 音乐舞蹈基准数据集

在新颖并动作更复杂的舞蹈数据集下，多模态音乐分支相较于谷歌提出的此前最优benchmark方法AI Choreographer (FACT)，在取得类似FID的质量基础上 多样性与节奏匹配度提升超过8% 速度提升3倍

Method	FID ↓	Diversity ↑	Beat Align ↑
Ground Truth	10.60	7.45	0.237
Li et.al	43.46	3.32	0.160
Dancenet	25.49	2.85	0.143
DanceRevolution	25.92	4.87	0.195
FACT	22.11	6.18	0.221
E2M (Ours)	38.23	6.68	0.248

Method	Task	Time (s) ↓
Guo et al.	Text2Motion	0.6220
E2M	Text2Motion	**0.0904**
FACT	Music2Dance	0.1082
E2M	Music2Dance	**0.0384**

相关成果以通讯作者发表在T-PAMI 2023、T-MM 2021和IJCAI 2024、ICIP 2023上

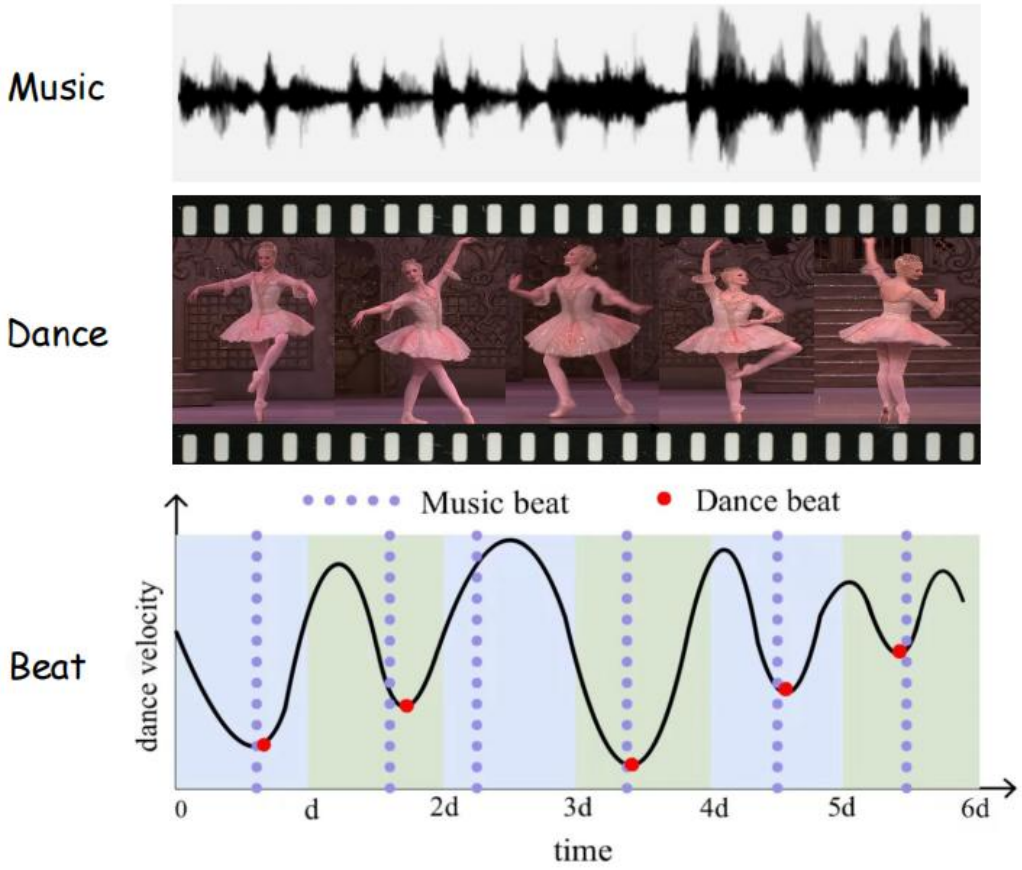
4.4 研究内容：Music-Dance Retrieval

问题背景

舞蹈和音乐是密切相关的表达形式，舞蹈视频与音乐之间的相互检索可广泛应用于教育、艺术和体育等多个领域。现有方法未能充分探索音乐与舞蹈之间的相关性以实现准确的相互检索。

研究内容

- 如何构建有效的数据集用于Music-Dance Retrieval任务
- 如何实现Music-Dance语义实时对齐



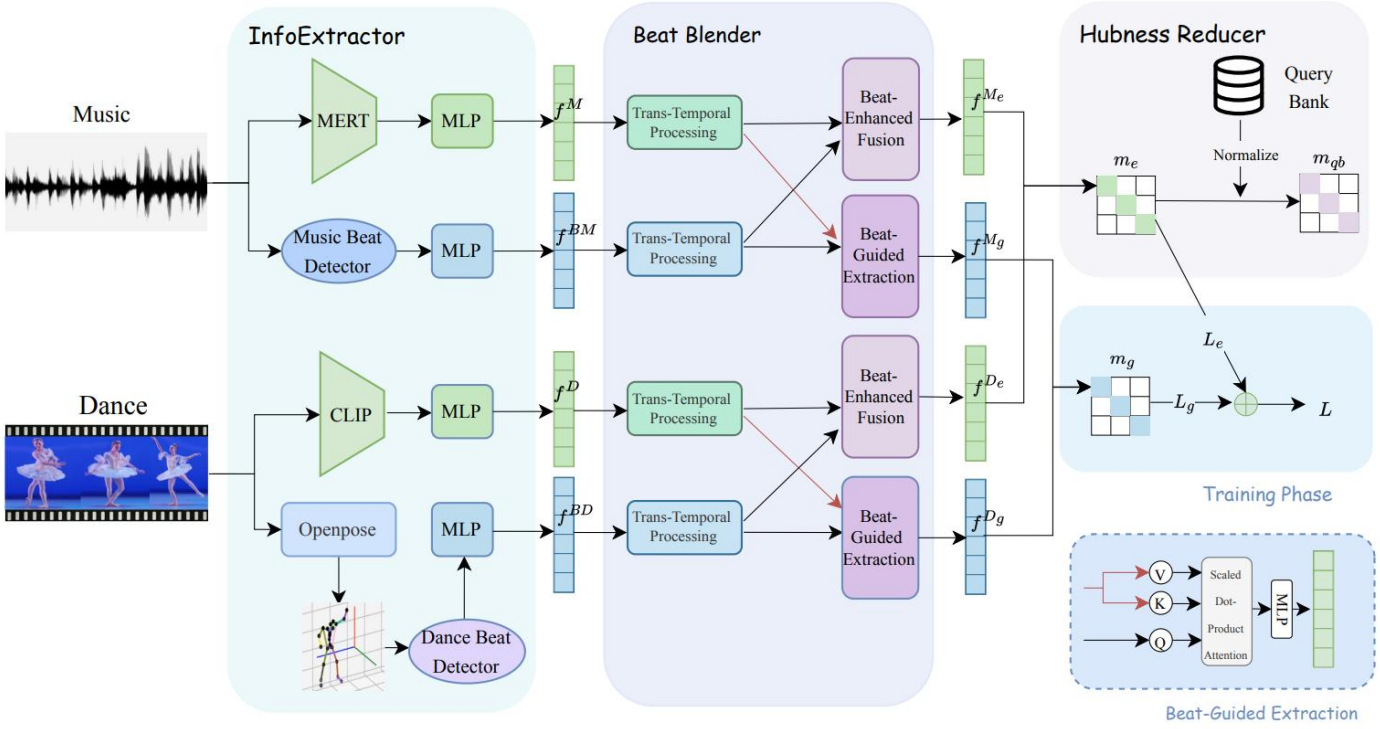
4.4 创新点：Music-Dance之间Beat相关性构建

构建BeatDance，建模Music-Dance之间Beat相关性

传统： 未充分探索音乐与舞蹈间相关性

瓶颈： 相关性构建

创新： 提出领域第一个开源大型数据集；依托Music-Dance之间Beat相关性提高了Music-Dance之间特征的相关性建模



4.4 创新点：Music-Dance之间Beat相关性构建

Table 1: Comparisons with state-of-the-art results on M-D dataset for music-to-dance retrieval and dance-to-music retrieval.

Method	Music ⟹ Dance		Dance ⟹ Music	
	Recall@1/10/50/100↑	MeanR/MedianR↓	Recall@1/10/50/100↑	MeanR/MedianR↓
CBVMR[12]	0.83/6.35/20.71/30.61	245.5/333.91	1.24/6.11/20.79/31.02	236.5/333.64
SCFEM[27]	0.99/7.76/23.10/35.81	196.0/306.05	0.91/8.25/23.27/35.31	192.0/305.65
MQVR[40]	1.65/8.91/26.90/39.60	152.5/263.80	1.24/9.49/26.90/39.11	152.0/265.36
MVPt[34]	1.57/8.25/26.24/38.78	162.5/258.15	1.23/9.46/27.81/39.42	166.0/254.81
XPool[10]	1.57/9.41/27.72/41.50	148.0/248.79	1.49/8.83/28.55/41.58	148.0/253.80
BeatDance	**2.48/13.12/32.26/44.06**	**128.0/239.81**	**2.97/13.04/32.34/44.55**	**127.0/238.77**

相关成果以通讯作者发表在ICMR 2024上

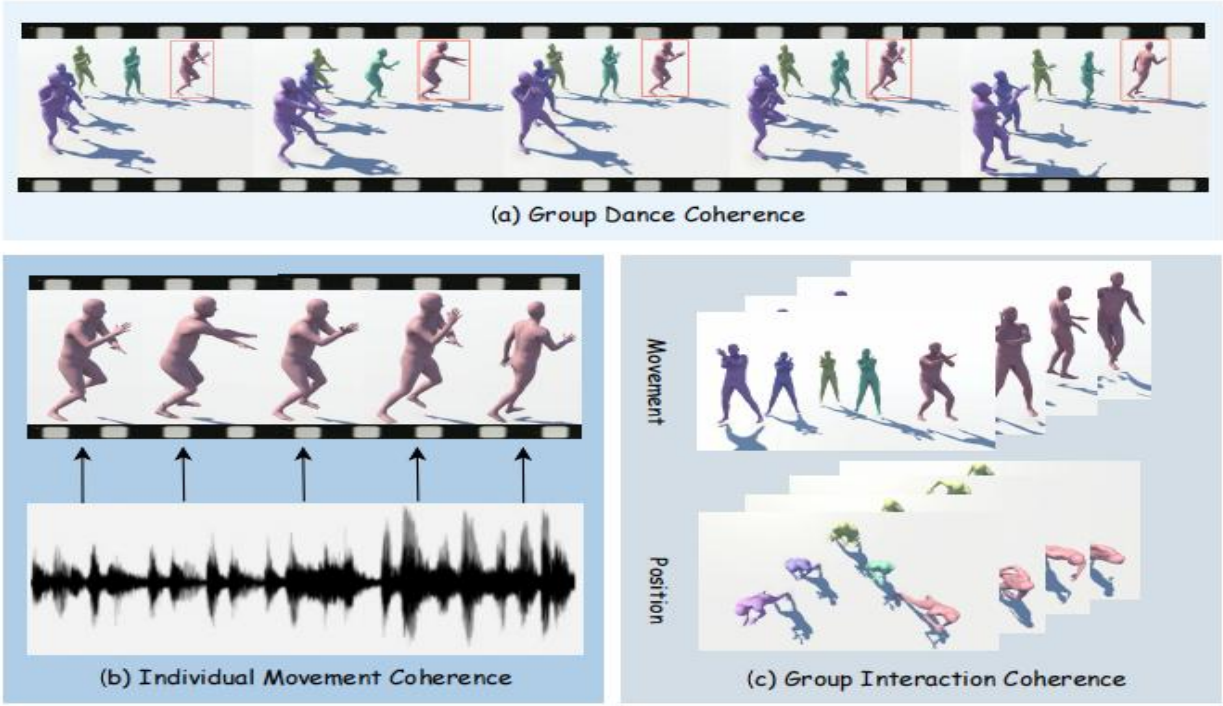
4.5 研究内容：音乐驱动的群舞生成

问题背景

舞蹈和音乐紧密相连，群体舞蹈是舞蹈艺术的重要组成部分。因此，音乐驱动的群体舞蹈生成已成为教育、艺术和体育等多个领域中的一个基本且具有挑战性的任务。现有方法对群舞协调性的探索不足。

研究内容

- 如何通过有效分解群舞协调性来提高群舞生成效果



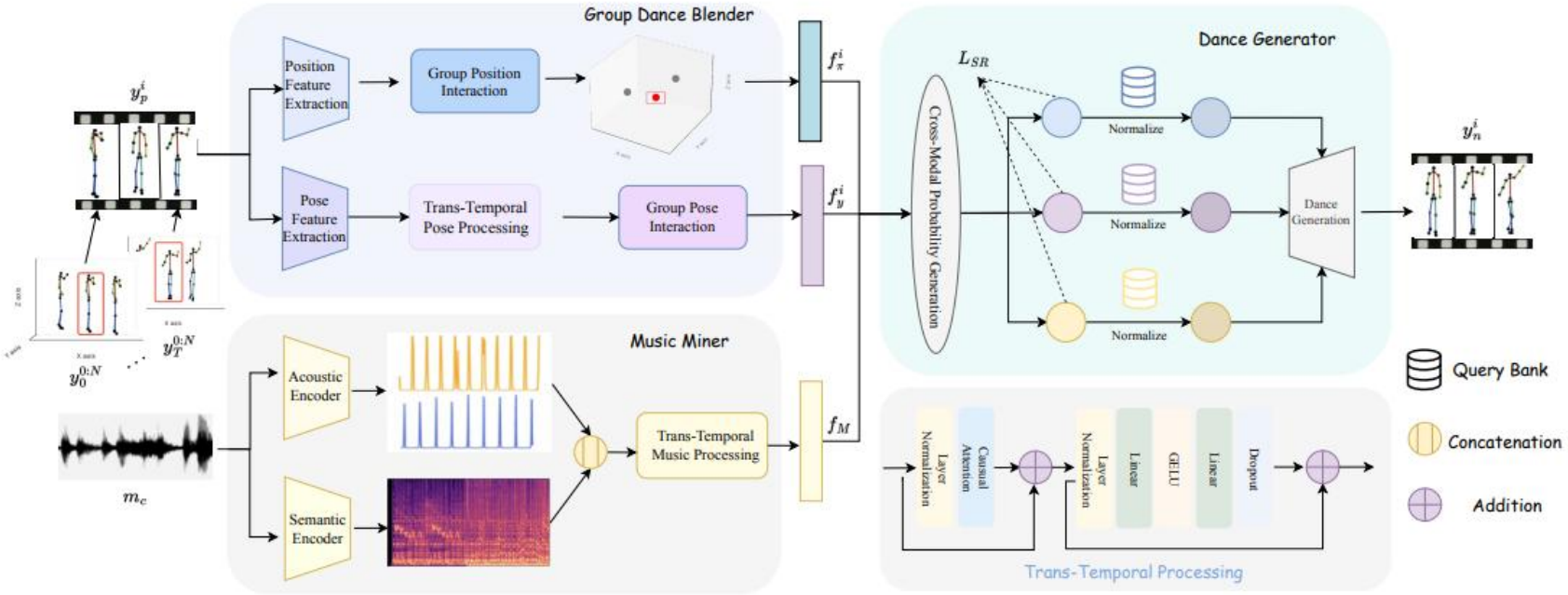
4.5 创新点：Retrieval-Based Group Music2Dance

构建CoDancers，基于Retrieval-Based方法有效建模群舞协调性

传统： 群舞协调性不足

瓶颈： 群舞协调性分解

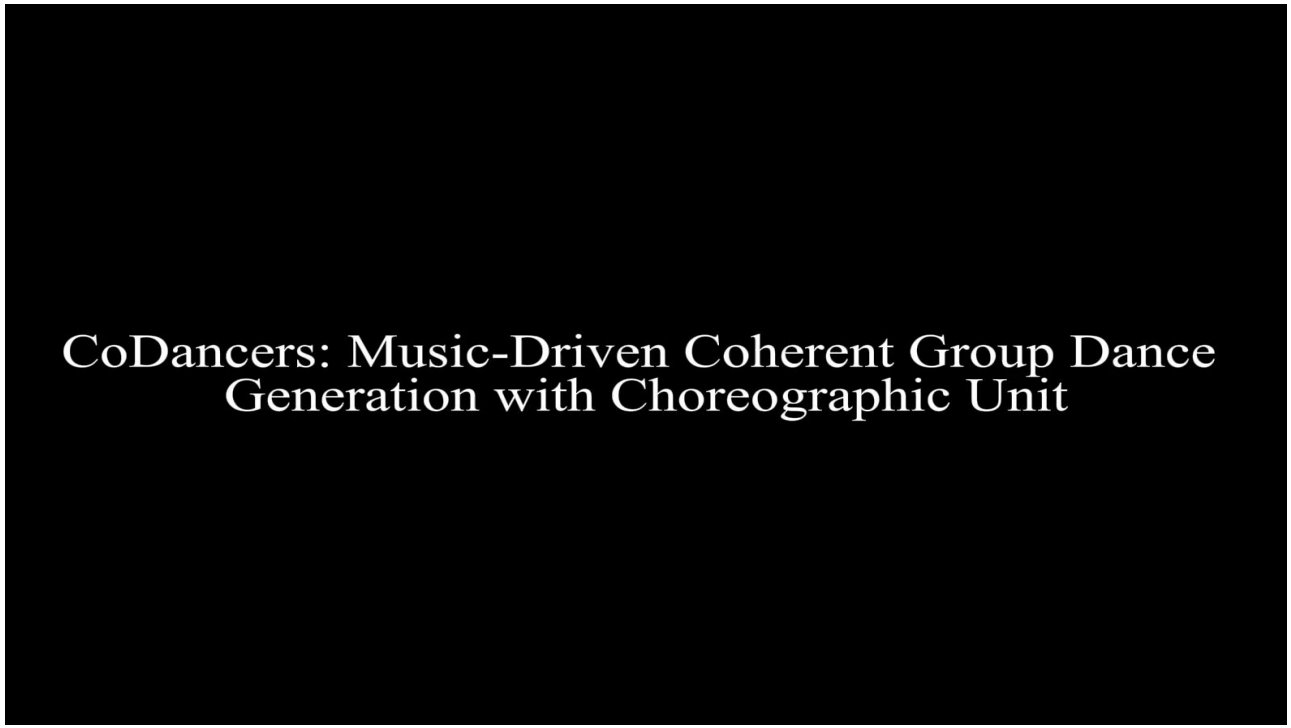
创新： 首次将Retrieval-Based方法应用于Group Music2Dance领域；将群舞协调性分解成群体交互协调性与个体动作协调性并为其做专门化设计



4.5 创新点：Retrieval-Based Group Music2Dance

Table 1: Comparisons with state-of-the-art results on AIOZ-GDANCE dataset for Group Music2Dance task.

Method	Individual			Group			US	
	FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓	IMQ↑	GIQ↑
Ground Truth	3.37	0.248	9.88	3.53	86.82	0.021	4.32	3.98
Transflower[32]	37.73	0.217	8.74	81.17	60.78	0.332	-	-
FACT[18]	56.20	0.222	8.64	101.52	62.68	0.321	2.84	2.28
Bailando[28]	39.21	0.242	8.68	50.41	72.35	0.286	3.24	2.42
GDANCER[14]	43.90	0.250	9.23	51.27	**79.01**	0.217	3.31	3.28
CoDancers	**23.98**	**0.254**	**9.48**	**26.10**	74.05	**0.102**	**3.57**	**3.68**



相关成果以通讯作者发表在ICMR 2024上

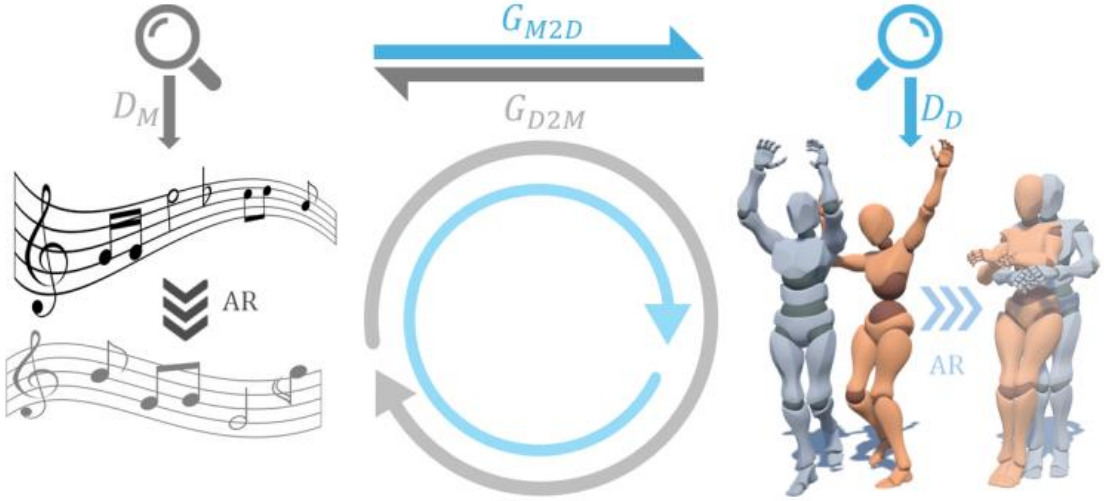
4.6 研究内容：音乐驱动的复杂舞曲群舞生成

问题背景

群体舞蹈是舞蹈艺术的重要组成部分，音乐驱动的群体舞蹈生成任务可广泛应用于音乐教育、虚拟现实等多个领域。现有方法对复杂舞曲群舞生成探索不足，并且缺少有效的复杂舞曲benchmark。

研究内容

- 如何通过有效分解舞蹈协调性来提高复杂舞曲群舞生成效果
- 如何构建具有强舞者交互的大型复杂舞曲群舞数据集
- 如何构建全面评价群舞生成效果的指标



4.6 创新点：基于舞蹈运动学的Group Music2Dance

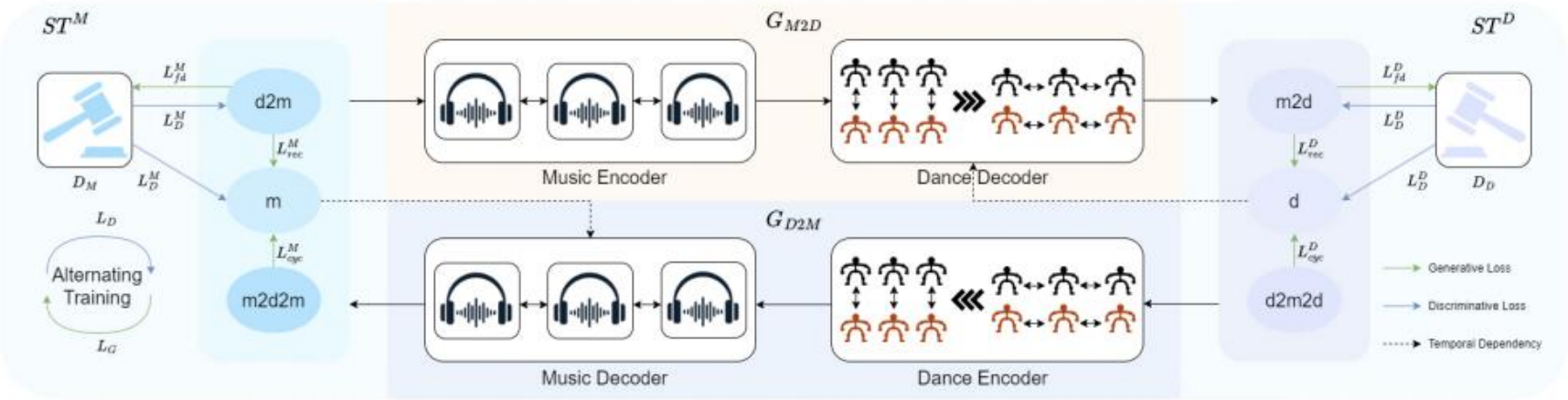
构建CoheDancers，探索基于舞蹈运动学的建模方法

传统： 复杂舞曲的群舞生成能力不足

瓶颈： 缺少benchmark

创新：

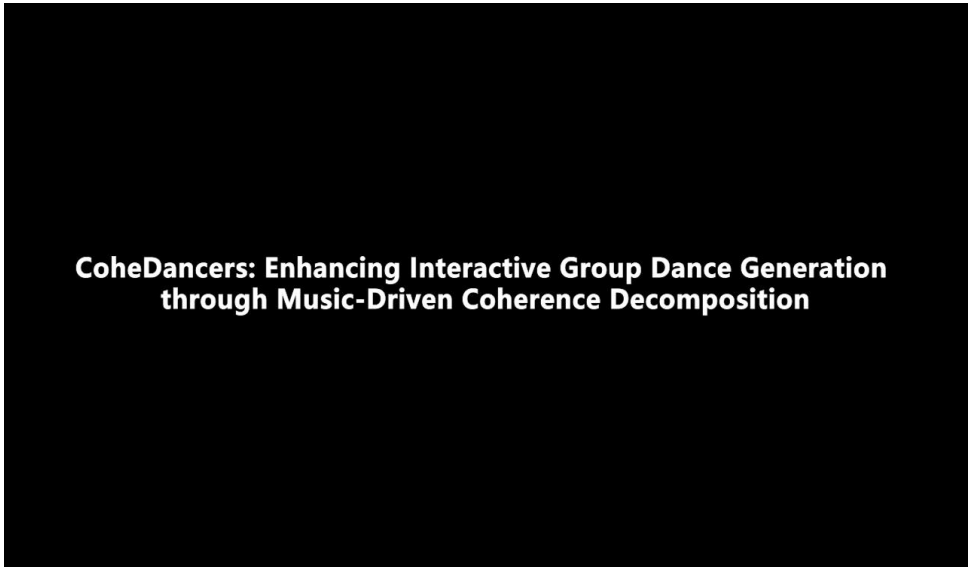
- 提出了大型带有复杂舞者交互数据集I-Dancers
- 提出了基于舞蹈运动学和检索模型的全面的评价体系
- 把舞蹈协调性分解成自然性、连贯性、对应性并对其进行专门化设计



4.6 创新点：基于舞蹈运动学的Group Music2Dance

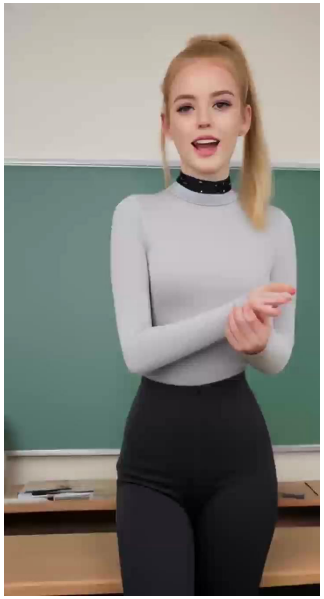
Table 2: Quantitative evaluation results for Group Music2Dance on AIOZ-GDANCE-P2, AIOZ-GDANCE-P3, and I-Dancers. We compare CoheDancers with Bailando(G) (Siyao et al. 2022), CoDancers (Yang et al. 2024a), FACT(G) (Li et al. 2021), and GDanceR (Le et al. 2023b), and explore the impacts of Dance Synchronization strategy (DS), Auto-Regressive-based Exposure Bias Correction (EBC) stragegy and Adversarial Training Strategy (AT) .

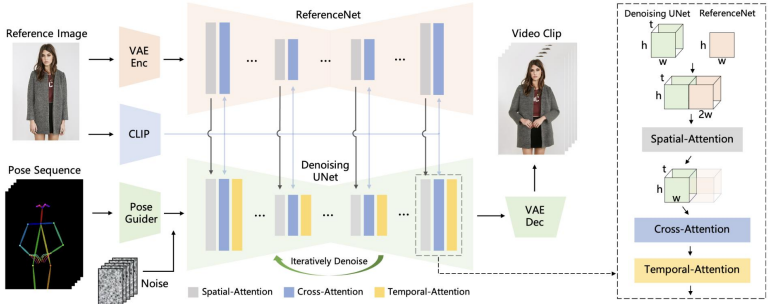
Datasets	Methods	FID ↓	M-Dist ↓	MM-dist ↓	Div ↑	MDA ↑	GDA ↑
AIOZ-GDANCE-P2	**Real Motions**	00.00	00.00	18.87	19.26	0.388	0.533
	Bailando(G)	100.45	16.20	20.66	17.21	0.326	0.288
	CoDancers	88.14	17.18	20.82	**18.03**	0.332	0.267
	FACT(G)	123.59	16.58	20.50	16.43	**0.408**	0.368
	GDanceR	104.09	15.11	20.15	16.70	0.402	0.337
	CoheDancers	**71.18**	**13.67**	**19.40**	17.54	0.405	**0.399**
AIOZ-GDANCE-P3	**Real Motions**	00.00	00.00	18.17	19.25	0.382	0.555
	Bailando(G)	**67.38**	15.65	19.87	18.19	0.337	0.281
	CoDancers	73.09	16.61	20.40	**18.16**	0.336	0.284
	FACT(G)	170.96	16.42	20.47	13.52	0.395	0.350
	GDanceR	121.87	15.52	19.98	15.41	0.391	**0.469**
	CoheDancers	81.63	**14.66**	**19.31**	17.30	0.402	0.402
I-Dancers	**Real Motions**	00.00	00.00	19.74	20.03	0.389	0.574
	Bailando(G)	58.31	18.50	21.35	18.72	0.321	0.273
	CoDancers	56.21	18.81	21.30	18.96	0.333	0.282
	FACT(G)	77.19	16.95	20.57	17.82	**0.396**	0.318
	GDanceR	76.11	16.16	20.27	17.66	0.387	**0.334**
	- *DS*	70.37	16.12	20.19	17.98	0.391	0.311
	- *AT*	55.92	15.71	20.20	18.40	0.392	0.308
	- *EBC*	130.78	15.87	20.24	15.34	0.367	0.499
	CoheDancers	**37.01**	**15.50**	**20.09**	**19.25**	0.390	0.313



相关成果以通讯作者发表在AAAI 2025上

4.7 研究内容：扩散范式下的2D pose驱动





animate anyone 轻量化改进

- 时序一致性提高
- 算法稳定性提高
- 动作复杂性提高

感谢聆听

报告人：范肇心