中国科学院软件研究所学术年会 2025 第十届
暨重点实验室科技活动周

学术论文

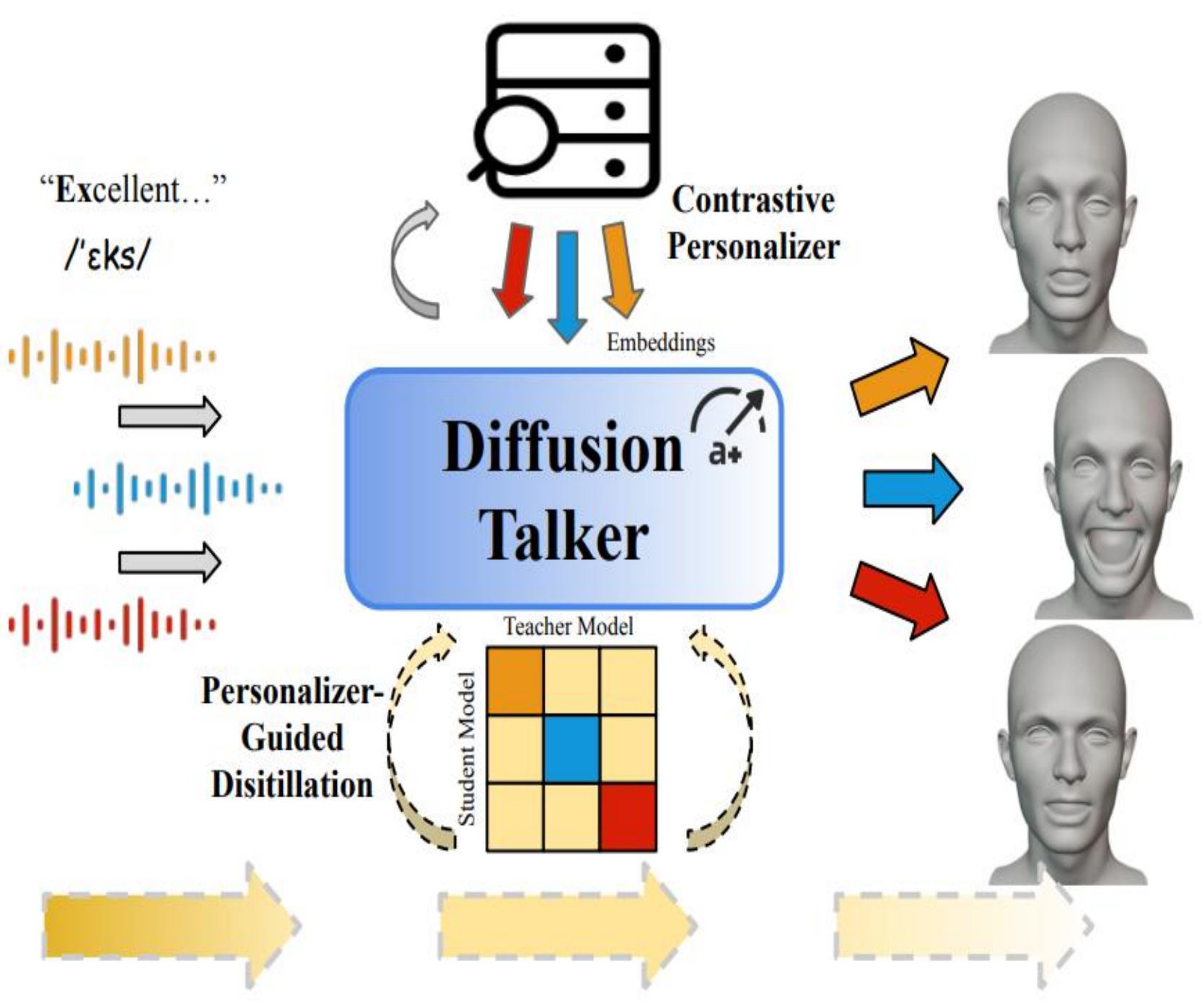
基于个性化引导蒸馏的高效紧凑型语音驱动三维说话头生成

DiffusionTalker: Efficient and Compact Speech-Driven 3D Talking Head via Personalizer-Guided Distillation

陈鹏，韦小宝，陆鸣，陈辉，田丰

IEEE International Conference on Multimedia & Expo 2025 (ICME 2025)

联系方式（陈鹏，chenpeng23@mails.ucas.ac.cn）



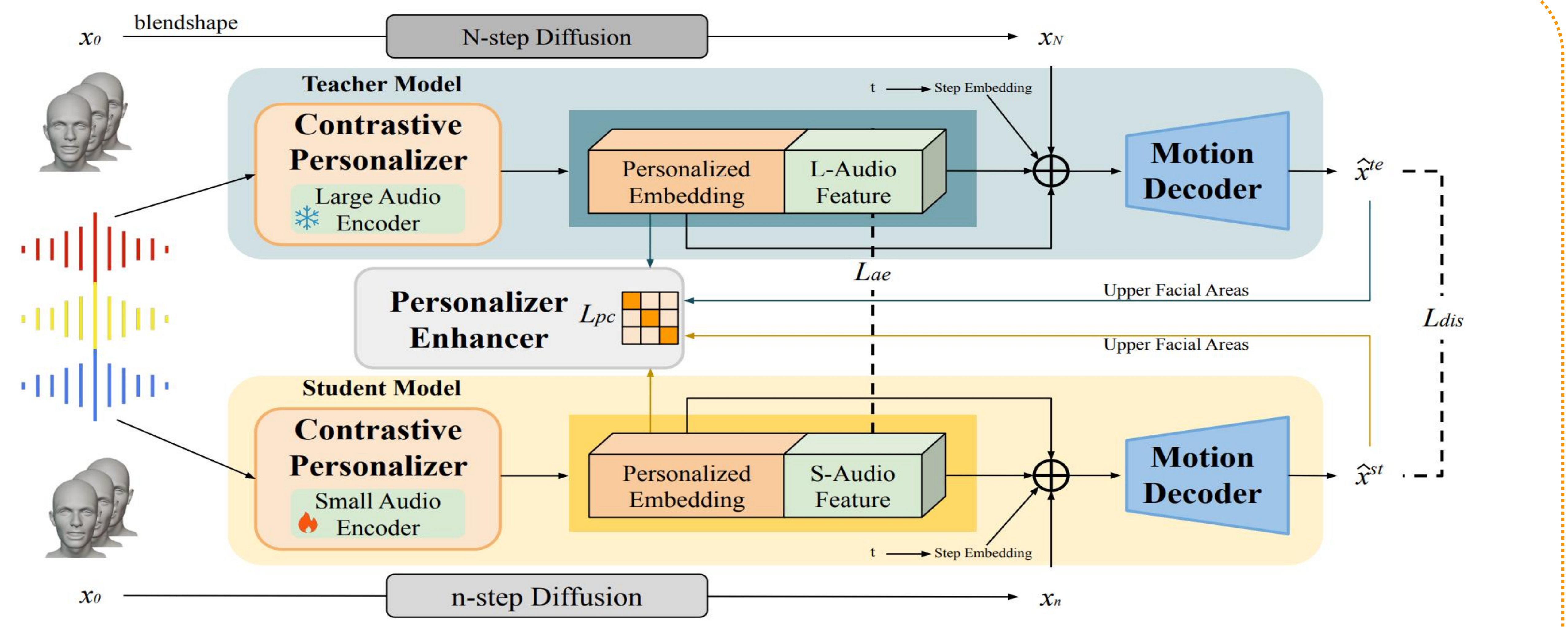
已有三维说话头方法存在以下局限：

1. 推理速度慢：现有方法通常难以实现实时性能，尤其是基于长步数扩散的模型。

2. 存储需求过高：多数方案依赖大型预训练音频编码器，导致模型参数冗余。

3. 身份与情感融合不足：相同文本，不同个体的面部表情和说话风格存在独特性，并随情绪状态动态变化。

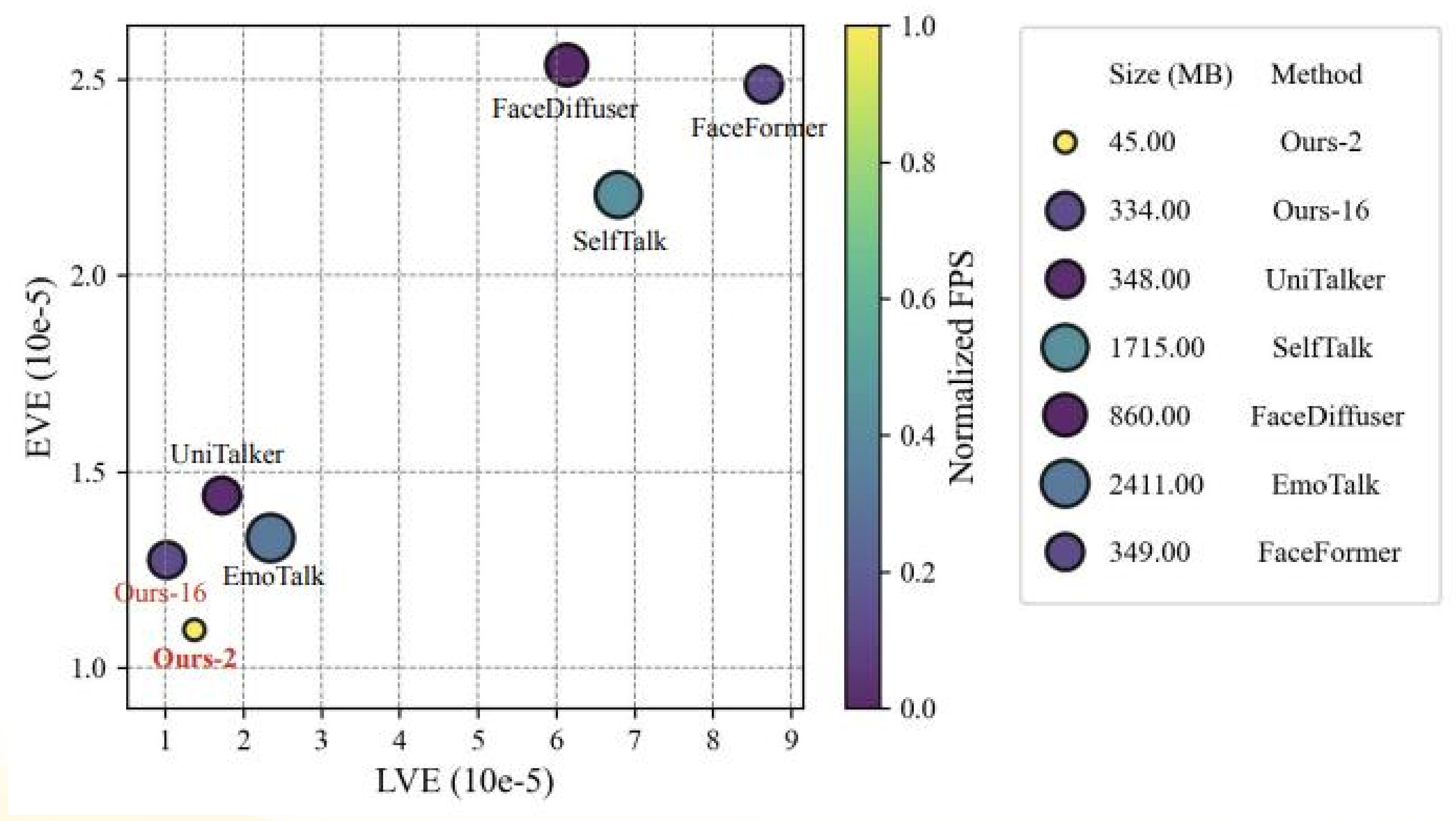
关键问题：在虚拟现实、增强现实等场景中，高效和紧凑且富有个性化的三维说话人是提升人机交互沉浸感的关键。



我们通过个性化引导的蒸馏方法，减少扩散模型的推理步数以加速生成，并压缩模型体积以实现轻量化。经蒸馏得到的2步模型在情感表现与唇形准确度上超越现有最优方法，同时实现最快的推理速度与最少的参数量。

Dataset	Method	EVE ↓ ($\times 10^{-5}$)	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)	FPS↑	Size(MB)↓
3D-ETF (Test)	FaceFormer	2.487	8.656	11.909	426.01	349
	EmoTalk	1.331	2.347	2.196	1063.13	2411
	FaceDiffuser	2.537	6.137	8.172	10.33	860
	SelfTalk	2.206	6.796	11.622	1428.57	1715
	UniTalker-B	1.439	1.727	3.004	68.75	348
	Ours-16	1.275	**1.021**	**2.023**	440.01	334
	Ours-2	**1.097**	1.375	2.177	**3632.15**	**45**

与现有方法相比，DiffusionTalker在唇部精度、情感表达精度、推理速度和模型大小上均表现最优。



整体比较

Dataset	Method	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)	Zero Shot
VOCASET (Test)	FaceFormer	1.170	2.493	✗
	FaceDiffuser	0.973	1.754	✗
	SelfTalk	0.967	**1.049**	✗
	UniTalker-B	**0.814**	1.396	✗
	Ours-2	0.857	1.198	✓

零样本测试

Settings	EVE↓ ($\times 10^{-5}$)	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)
Ours-2	**1.097**	1.375	**2.177**
w/o identity embedding	1.223	1.412	2.301
w/o emotion embedding	1.968	1.547	2.570
w/o enhancer	1.712	**1.116**	2.512
w/o distillation	1.305	1.647	2.618

消融实验