从视觉语言到视觉触觉：多模态智能的关键突破与方法研究

汇 报 人：寇硕、林骅、
李文卓、兰红星

2025/10/24

目录

引言

什么是“多模态”

信息的多样表现形式：

- 文本：书面语言、符号、代码
- 视觉：图像、视频
- 音频：语音、声音
- 触觉：纹理、压力、振动

引言

为什么“多模态”

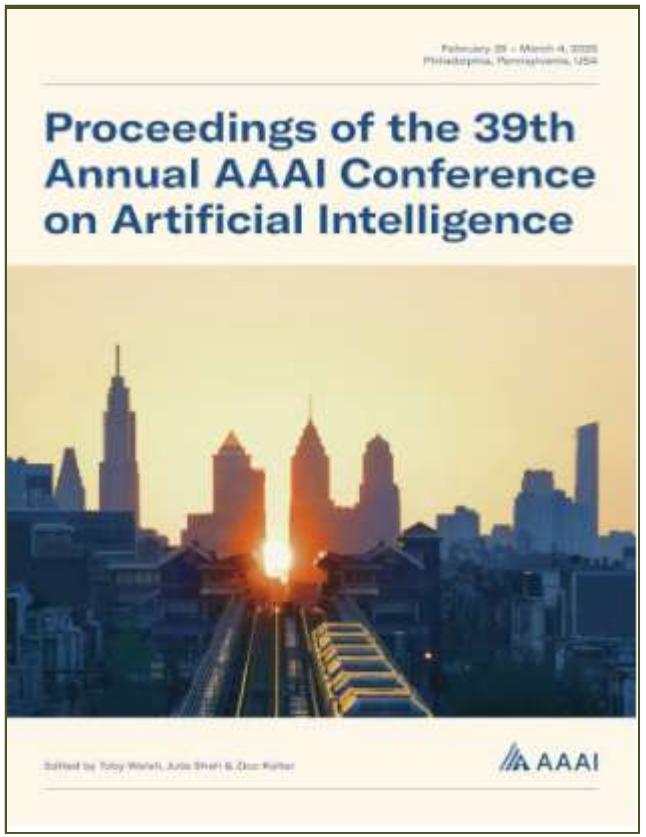
从“单模态”到“多模态”可以实现：

- 模态间信息增强
- 模态间数据转换
- 模态间**特征融合**


Part 01
StoryWeaver

目录

论文介绍



Acknowledgments

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. U21B2037, No. 62072389, No. 62302411), the Natural Science Foundation of Fujian Province of China (No.2021J06003), China Postdoctoral Science Foundation (No. 2023M732948), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

选文原因

真正的“电影级”AI 短片《200 美元》

【保姆级教程】AI动画电影制作，12个作品涨粉50万，手把手教你Comfy…

目前为止看到过的最喜欢、最有生命力的 AI生成视频~

逆天！AI擦边赛道日入5位数：ComfyUI+FLUX+SD终极组合拳，3分…

论文介绍

StoryWeaver漫画生成

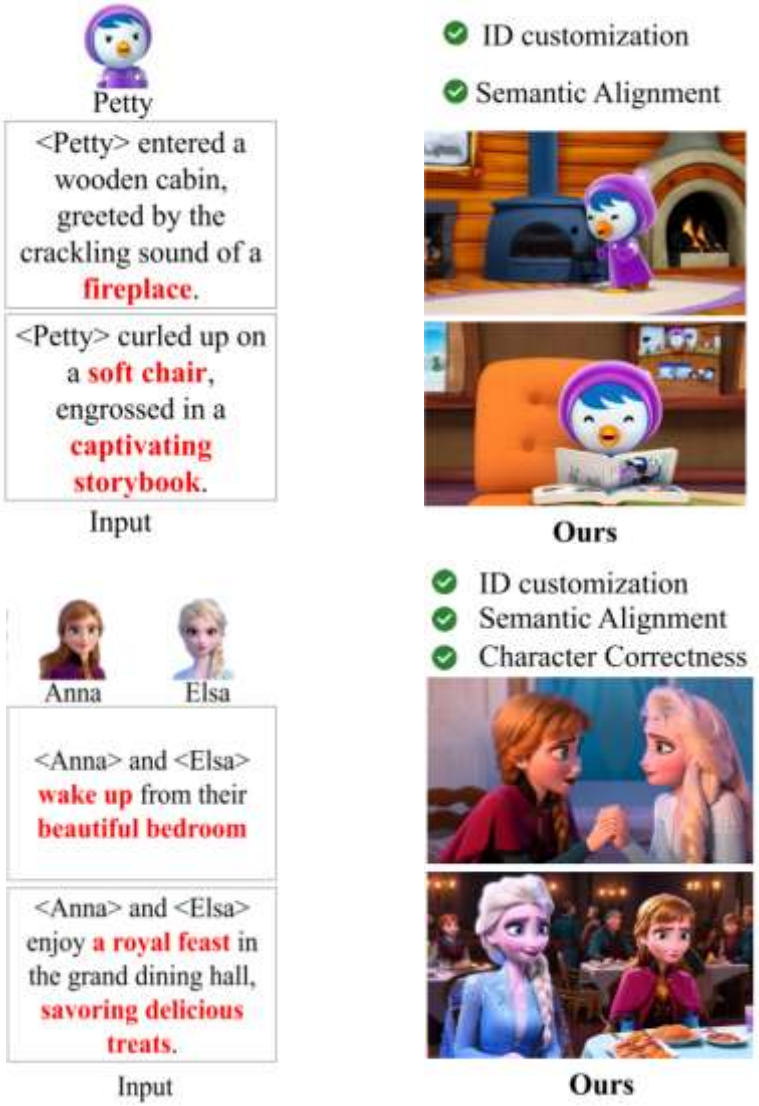
可以根据故事描述生成高质量的漫画

任务介绍

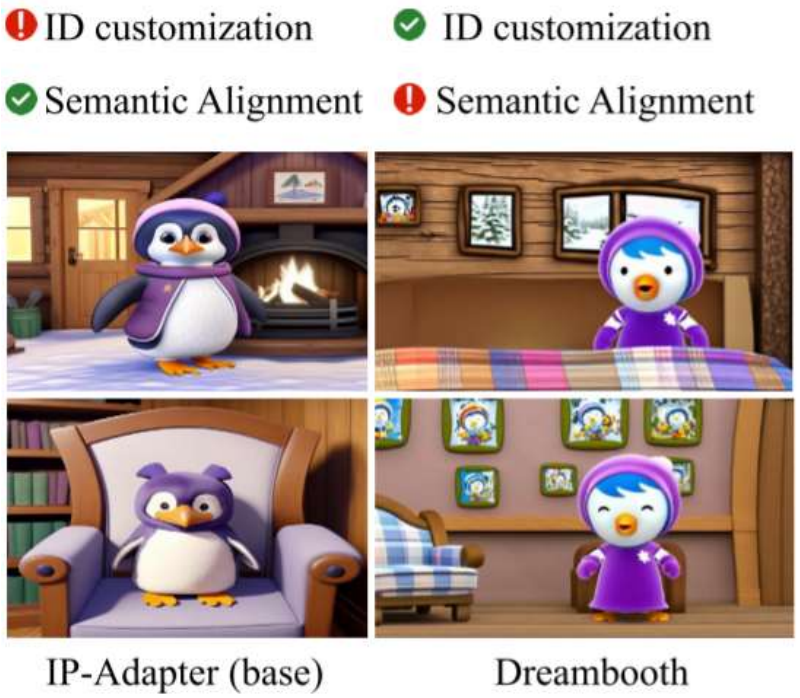
故事生成

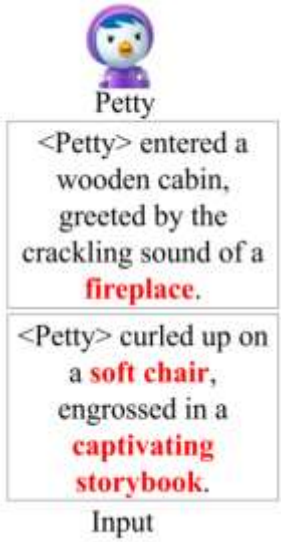
根据文本描述（字幕）生成对应的图片（视频/漫画场景）。需要满足：

- 生成图片与文本描述一致
- 图片间同一角色特征保持
- 图片内不同角色身份隔离



现有方法及其局限





现有角色生成方案
（单角色场景）
(a)适配器方法 （b） 定制化方法

现有角色生成方案
（多角色场景）
固定角色生成位置方法

StoryWeaver的解决方案

$$C_i = V_{cap}(Instruct_c, I_i), i \in [1, N_c], \quad (1)$$

$$\sum Map < O_i, A_i^k >= (SG_i^{(A)}|O_i) = (Parser(C_i)|O_i). \quad (2)$$

$$R(O_i, O_j) = (SG_i^{(R)}|(O_i, O_j)) = (Parser(\mathcal{F}_c))_{i,j}, \quad (3)$$

$$G(O, E) = \{\sum_{i,k} Map < O_i, A_i^k >, \sum_{i,j} R(O_i, O_j)\}. \quad (4)$$

$$\sum_i R_{j,*}, char_j = Parser(V_{cap}(Instruct_e, \mathcal{F})), \quad (5)$$

$$\begin{aligned} O_{char}^j &= \arg\max_{O_i} Sim(char_j, O_i), \\ A_{char}^j &= O_{char}^j \otimes \sum Map < O_i, A_i^k > . \end{aligned} \quad (6)$$

$$W_c^j = O_{char}^j \oplus A_{char}^j, \quad (7)$$

$$W_e = \sum_{j,j'} O_{char}^j \oplus R_{j,j'} \oplus O_{char}^{j'}. \quad (8)$$

$$\mathbb{E}_{f\sim E(\mathcal{F}), T_g, \epsilon\sim\mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(f_t, t, \tau(T_g))\|_2^2\right], \quad (9)$$

人物角色图（C-CG）

$$Attn = \mathcal{M} \cdot V = \text{Softmax}\left(\frac{(W_q f_{\mathcal{F}})(W_k f_T)^T}{\sqrt{d}}\right) \cdot (W_v f_T), \quad (10)$$

$$p_j(x, y) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu_x, y-\mu_y)\Sigma^{-1}(x-\mu_x, y-\mu_y)^T}, \quad (11)$$

$$\mathcal{E}(W_c^j) \rightarrow \{(\Delta\mu_x, \Delta\mu_y), \gamma\}, \quad (12)$$

$$\begin{aligned} p_j(x, y|t_j) &= \frac{1}{2\pi\sqrt{|\hat{\Sigma}|}} e^{-\frac{1}{2}(x-\hat{\mu_x}, y-\hat{\mu_y})\hat{\Sigma}^{-1}(x-\hat{\mu_x}, y-\hat{\mu_y})^T}, \\ \hat{\Sigma} &= \gamma \cdot \Sigma, \hat{\mu_x} = \mu_x + \Delta\mu_x, \hat{\mu_y} = \mu_y + \Delta\mu_y, \end{aligned} \quad (13)$$

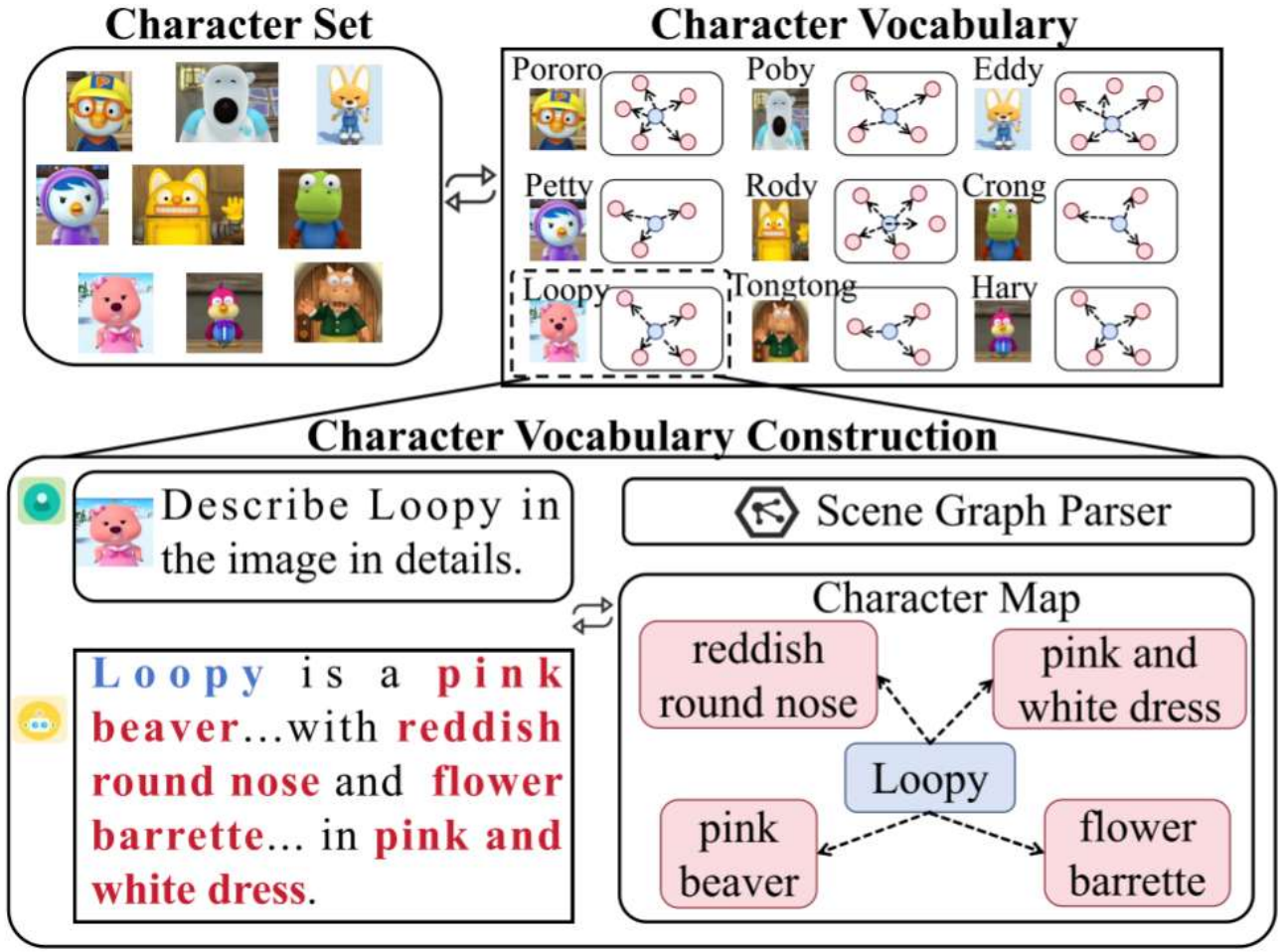
$$[\mathcal{M}_t^i]_{(x,y)} = \begin{cases} [\mathcal{M}_t^i]_{(x,y)} + s, \text{if } [\mathcal{P}^j]_{(x,y)} \geq \beta, \\ [\mathcal{M}_t^i]_{(x,y)} - s, \text{if } [\mathcal{P}^j]_{(x,y)} < \beta. \end{cases} \quad (14)$$

$$s = s(t) = \alpha \cdot (ln(t+1) + 1), \quad (15)$$

知识增强的空间引导（KE-SG）

StoryWeaver的解决方案

人物角色图（C-CG）



$$C_i = V_{cap}(Instruct_c, I_i), i \in [1, N_c], \quad (1)$$

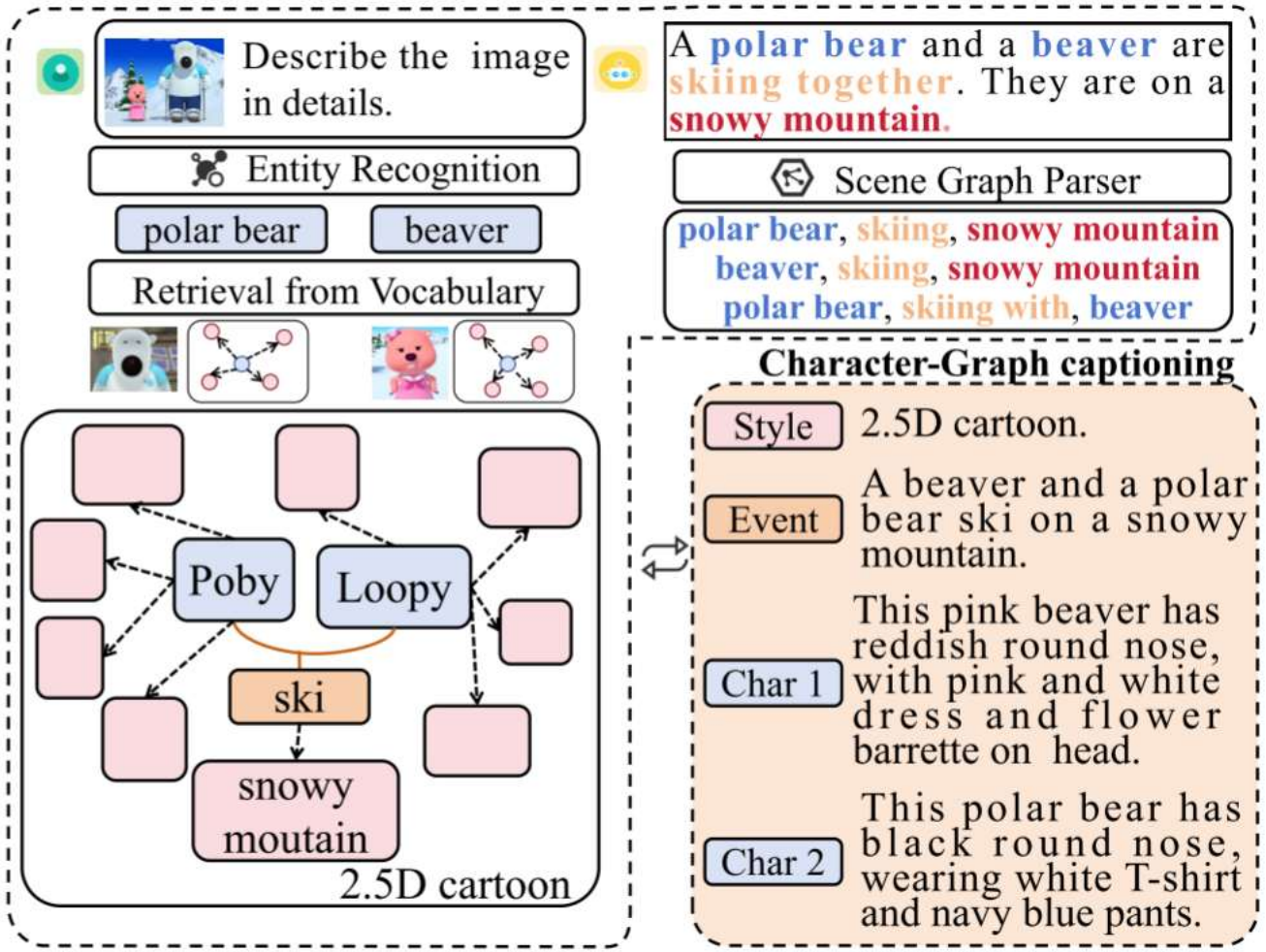
$$\sum Map < O_i, A_i^k >= (SG_i^{(A)}|O_i) = (Parser(C_i)|O_i). \quad (2)$$

$$R(O_i, O_j) = (SG_i^{(R)}|(O_i, O_j)) = (Parser(\mathcal{F}_c))_{i,j}, \quad (3)$$

$$G(O, E) = \{\sum_{i,k} Map < O_i, A_i^k >, \sum_{i,j} R(O_i, O_j)\}. \quad (4)$$

StoryWeaver的解决方案

使用C-CG



$$\sum_j R_{j,*}, char_j = Parser(V_{cap}(Instruct_e, \mathcal{F})), \tag{5}$$

$$O^j_{char} = \arg\max_{O_i} Sim(char_j, O_i),$$
$$A^j_{char} = O^j_{char} \otimes \sum Map < O_i, A^k_i > . \tag{6}$$

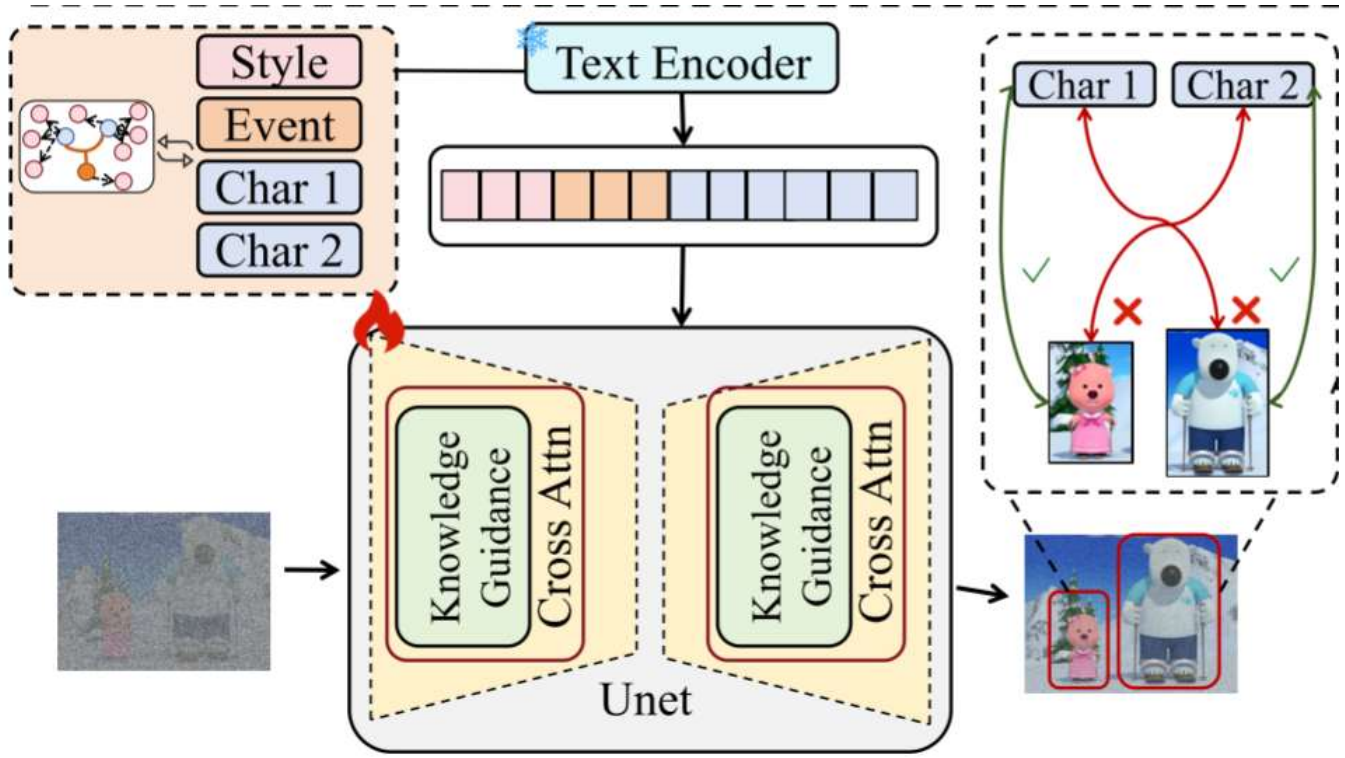
$$W^j_c = O^j_{char} \oplus A^j_{char}, \tag{7}$$

$$W_e = \sum_{j,j'} O^j_{char} \oplus R_{j,j'} \oplus O^{j'}_{char}. \tag{8}$$

$$\mathbb{E}_{f\sim E(\mathcal{F}),T_g,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(f_t, t, \tau(T_g))\|^2_2\right], \tag{9}$$

StoryWeaver的解决方案

知识增强的空间引导（KE-SG）



$$Attn = \mathcal{M} \cdot V = \text{Softmax}\left(\frac{(W_q f_{\mathcal{F}})(W_k f_T)^T}{\sqrt{d}}\right) \cdot (W_v f_T), \tag{10}$$

$$p_j(x,y) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu_x, y-\mu_y)\Sigma^{-1}(x-\mu_x, y-\mu_y)^T}, \tag{11}$$

$$\mathcal{E}(W_c^j) \rightarrow \{(\Delta\mu_x, \Delta\mu_y), \gamma\}, \tag{12}$$

$$p_j(x,y|t_j) = \frac{1}{2\pi\sqrt{|\hat{\Sigma}|}} e^{-\frac{1}{2}(x-\hat{\mu_x}, y-\hat{\mu_y})\hat{\Sigma}^{-1}(x-\hat{\mu_x}, y-\hat{\mu_y})^T},$$

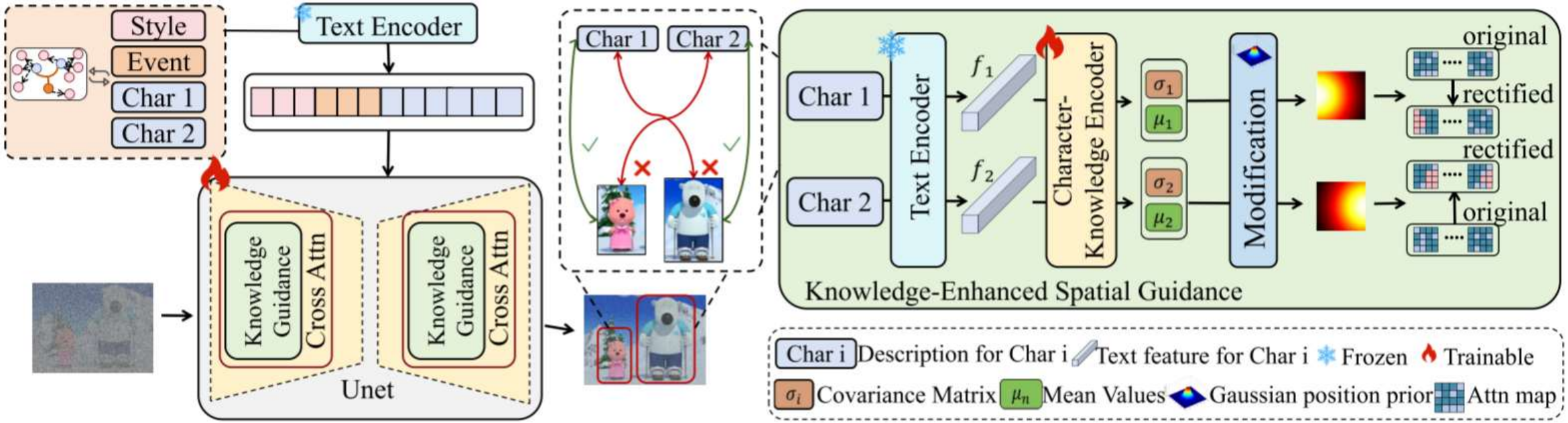
$$\hat{\Sigma} = \gamma \cdot \Sigma, \hat{\mu_x} = \mu_x + \Delta\mu_x, \hat{\mu_y} = \mu_y + \Delta\mu_y, \tag{13}$$

$$[\mathcal{M}_t^i]_{(x,y)} = \begin{cases} [\mathcal{M}_t^i]_{(x,y)} + s, \text{if } [\mathcal{P}^j]_{(x,y)} \geq \beta, \\ [\mathcal{M}_t^i]_{(x,y)} - s, \text{if } [\mathcal{P}^j]_{(x,y)} < \beta. \end{cases} \tag{14}$$

$$s = s(t) = \alpha \cdot (ln(t+1)+1), \tag{15}$$

StoryWeaver的解决方案

训练全流程



实验结果分析

Task	Type	Method	# Params/DB(↓)	Pororo			Frozen		
				DINO-I(↑)	CLIP-I(↑)	CLIP-T(↑)	DINO-I(↑)	CLIP-I(↑)	CLIP-T(↑)
Sin-Char	Adapter-based	StoryGEN	1064 M	52.92	76.03	26.98	46.67	72.61	28.05
		IP-Adapter(base)	1038 M	48.85	76.66	29.98	44.15	78.42	31.69
		IP-Adapter(plus)	1063 M	64.36	81.64	24.88	60.87	84.52	27.15
	Customization-based	LORA	1024 M	54.13	75.19	28.53	49.02	82.77	29.18
		Dreambooth	7118 M	61.85	78.86	26.74	55.01	81.07	27.12
		StoryWeaver(ours)	1017 M	**64.96**	**82.65**	**33.26**	**62.17**	**85.24**	**36.74**

Task	Type	Method	# Params/DB(↓)	Pororo			Frozen		
				CLIP-T(↑)	F-Acc(↑)	C-F1(↑)	CLIP-T(↑)	F-Acc(↑)	C-F1(↑)
Multi-Char	Adapter-based	StoryGEN	1064 M	27.27	19.55	27.17	28.91	12.31	21.79
	Customization-based	Mix-of-Show	1164 M	27.20	30.23	44.03	30.71	18.90	30.62
		LoRA-Composer	1425 M	27.86	27.04	47.36	28.88	27.69	39.72
		StoryWeaver(ours)	1017 M	**34.30**	**40.45**	**59.72**	**34.94**	**34.51**	**44.53**

Table 1: Quantitative comparisons on the single- and multi-character generation with existing methods. Our StoryWeaver obviously merits in semantic alignments with high identity customization compared to existing methods.

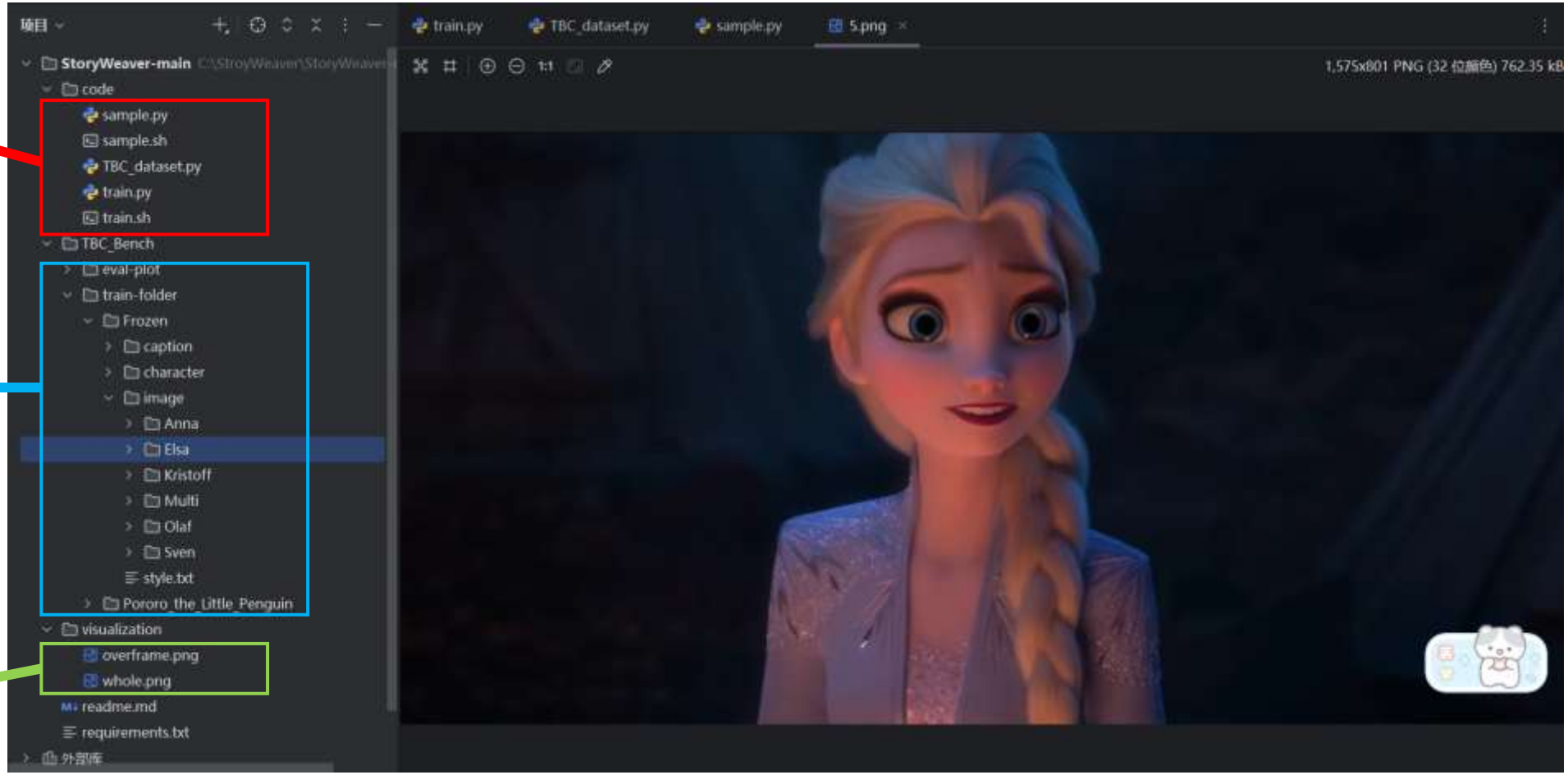
模型在文图一致、角色特征一致、角色身份保留三方面均取得SOTA

复现情况

模型代码

数据集

可视化图表



复现情况

单角色故事生成

Loopy visited Notre-Dame Cathedral,admiring its stunning architecture andintricate details.

Loopy sat in a charming café , sipping hot chocolate and nibbling on a flaky croissant.

Loopy picnicked in front of the Arcde Triomphe , enjoying delicious French delicacies.

多角色故事生成

Anna and Kristoff sit on the blanket at night happily in the garden and gaze at the starry sky.

Anna and Kristoff look at each other under the beautiful starry night.

Kristoff surprises Anna with a heartfelt serenade by playing Ukulele at night, his voice harmonizing with the night breeze.

总结与展望

本文引发的思考

从文章本身到我们的日常：

- 论文改进：丰富KE-SG策略
- 过去：论文工作流剖析
- 现在：研究方向探索
- 未来：AI对人工取代程度预测

Part 02

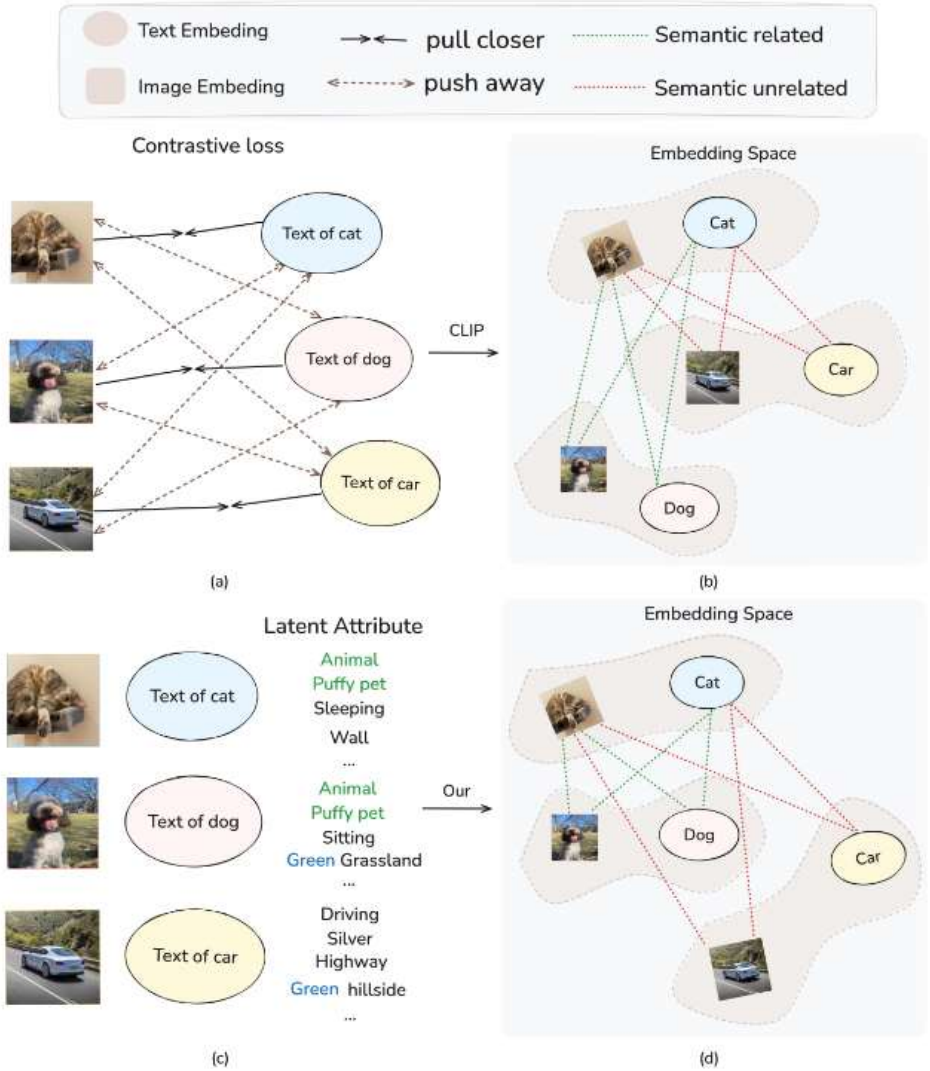
RankCLIP

RANKCLIP: Ranking-Consistent Language-Image Pretraining

Yiming Zhang[*1,2] **Zhuokai Zhao**[*3]
Zhaorun Chen[3] **Zhili Feng**[4] **Zenghui Ding**[1] **Yining Sun**[1,2]
[1]HFIPS, Chinese Academy of Sciences [2]University of Science and Technology of China
[3]University of Chicago [4]Carnegie Mellon University

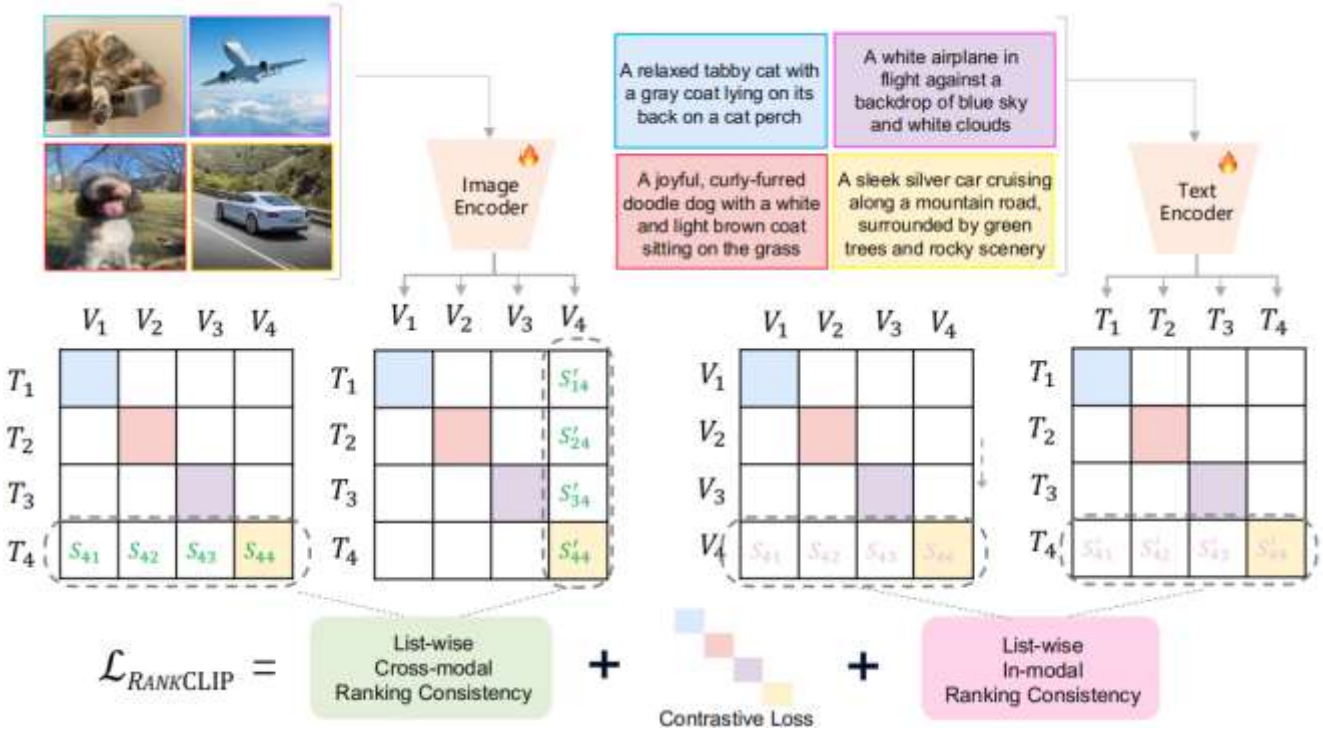
Publication：ICCV 2025

论文概述



- RANK-CLIP: 是基于排名一致性(Ranking)的CLIP与训练方法
- 论文动机: 突破传统对比学习方法(如CLIP)中一对一匹配关系, 但忽略多对多关系
- 论文核心贡献: 提出列表是(List-wise)损失函数, 利用跨模态(Cross-modal)和模态内(In-modal)的排名一致性来捕捉语义关系
- 图像解释: 三个图文对猫(cat)、狗(dog)和汽车(car)。对比损失无法捕捉猫、狗和汽车的潜在相似性（相关或无关）

RankCLIP



- RANKCLIP基于Plackett-Luce(PL)排名模型, 估计图像-文本对的排名优化模态一致性.
- $\pi(d|y_{1:k-1}, y_{ref}, D) = \frac{e^{s_{ij}}}{\sum_{d' \in D \setminus y_{1:k-1}} e^{s'_{ij}}}$
- $P(y, y_{ref}) = \prod_{k=1}^{K} \pi(y_k|y_{1:k-1}, y_{ref}, D)$
- $L_{PL} = -\log P(y|y_{ref})$
- $L_{cross-modal} = -\log P(y_{V-T}, y_{T-V})$
- $L_{in-modal} = -\log P(y_{V-V}, y_{T-T})$
- S_{ij}表示图文对(V_i, T_j)的CLIP得分(余弦值), y_{ref}表示参考排序
- $L_{cross-modal}$可以理解为学习了一个对称的余弦相似度矩阵

$$\mathcal{L}_{\text{RANKCLIP}} = \mathcal{L}_{\text{CLIP}} + \lambda_1 \mathcal{L}_{\text{in-modal}} + \lambda_2 \mathcal{L}_{\text{cross-modal}}$$

实验结果

	ImageNetV2-Matched			ImageNetV2-Threshold			ImageNetV2-Top			ImageNet-R		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP [43]	7.53%	14.99%	19.61%	8.89%	17.22%	21.86%	10.76%	19.80%	24.87%	9.36%	10.56%	19.76%
CyCLIP [19]	7.68%	15.07%	19.11%	9.10%	17.42%	21.94%	11.20%	20.18%	25.34%	9.23%	16.72%	21.64%
ALIP [58]	7.82%	15.56%	19.81%	9.65%	18.31%	22.85%	11.43%	20.88%	26.10%	10.92%	20.27%	26.24%
RANKCLIP	**9.01%**	**16.95%**	**21.12%**	**10.32%**	**19.31%**	**24.13%**	**12.31%**	**22.11%**	**27.17%**	**11.34%**	**20.88%**	**26.94%**

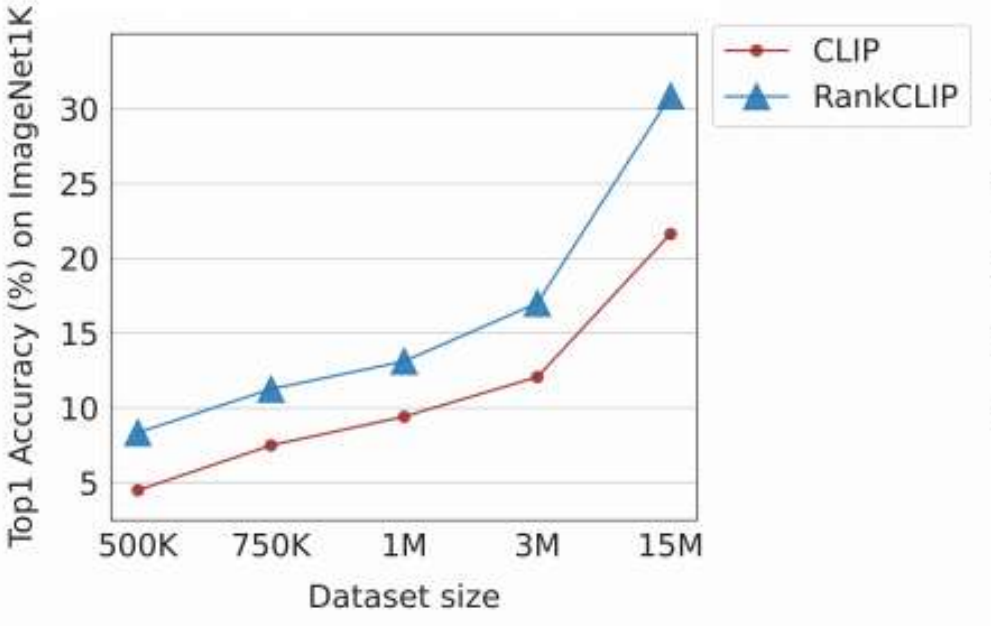
Table 2. Zero-shot top-1, 3, and 5 accuracy on ImageNet1K variants with *natural distribution shifts*. Compared to baselines, RANKCLIP achieves higher accuracies. Notably, these gains are more pronounced than on standard ImageNet1K, highlighting improved robustness.

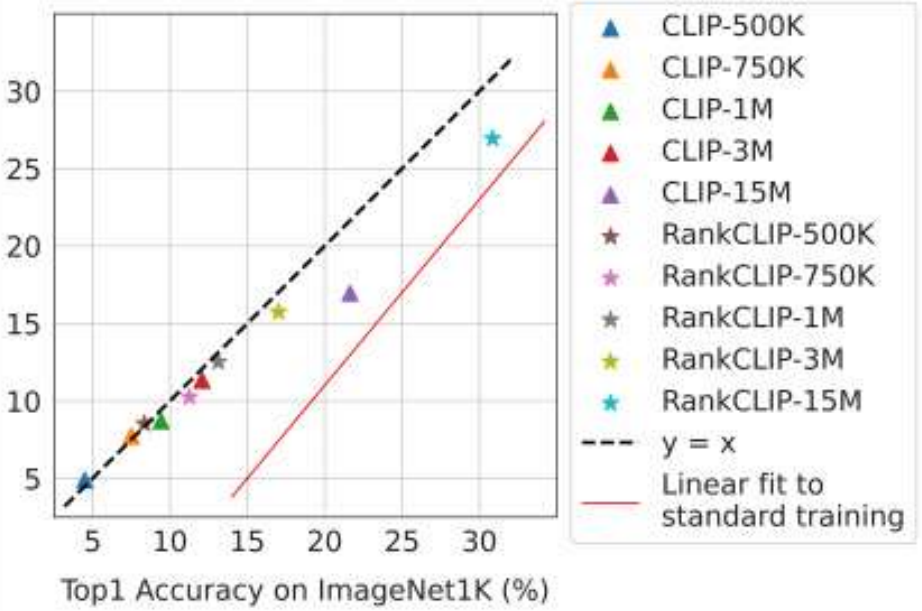
	CIFAR-10	CIFAR-100	DTD	FGVCAircraft	Food101	GTSRB	OxfordPets	SST2	STL10	SVHN	Average
CLIP [43]	77.6%	56.2%	43.2%	22.6%	39.7%	60.0%	40.4%	51.0%	79.0%	**50.5%**	52.0%
CyCLIP [19]	76.8%	54.3%	45.8%	19.2%	37.5%	58.6%	**44.2%**	51.5%	**82.3%**	41.3%	51.2%
ALIP [58]	71.1%	49.1%	**47.1%**	17.4%	36.1%	51.5%	41.9%	53.3%	81.0%	38.3%	48.7%
RANKCLIP	**78.4%**	**56.6%**	42.4%	**23.4%**	**40.2%**	**60.6%**	40.6%	**53.4%**	79.6%	47.7%	**52.3%**

Table 3. Linear probing accuracy on 10 downstream datasets using a ViT backbone.

Method	Vision Backbone	ImageNet1K			MSCOCO						Linear Probing Avg. Acc.
					Image-to-Text Retrieval			Text-to-Image Retrieval			
		Top-1	Top-3	Top-5	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP [43]	RN50	21.6%	36.9%	44.9%	15.6%	36.4%	48.4%	6.7%	15.2%	20.1%	64.2%
RANKCLIP		**30.9%**	**49.4%**	**57.6%**	**19.5%**	**42.6%**	**54.8%**	**7.5%**	**16.2%**	**21.6%**	**68.9%**
CLIP [43]	ViT-B/32	20.7%	35.0%	42.4%	11.9%	29.4%	40.8%	5.1%	12.9%	17.9%	60.7%
RANKCLIP		**26.2%**	**41.4%**	**48.9%**	**13.8%**	**33.8%**	**45.9%**	**6.0%**	**13.6%**	**18.6%**	**61.3%**

- 图1：在ImageNet1K数据集的自然分布偏移变体上，RANKCLIP模型实现了零样本训练下的前1%、3%和5%准确率。与基准模型相比，其准确率显著提升。
- 图2：基于ViT骨干网络对10个下游数据集进行线性探测的精度评估，可见RankCLIP平均成绩高于其他模型。
- 图3：MSCOCO跨模态检索及线性探测任务中，基于不同视觉主干网络训练的零样本评估结果。





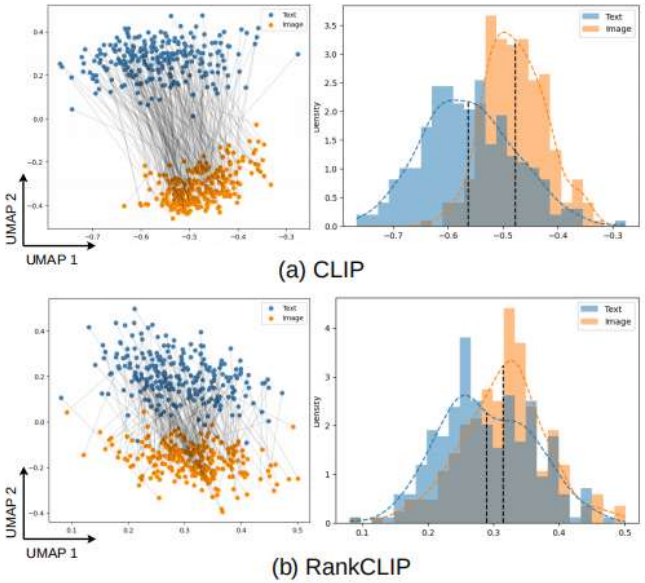
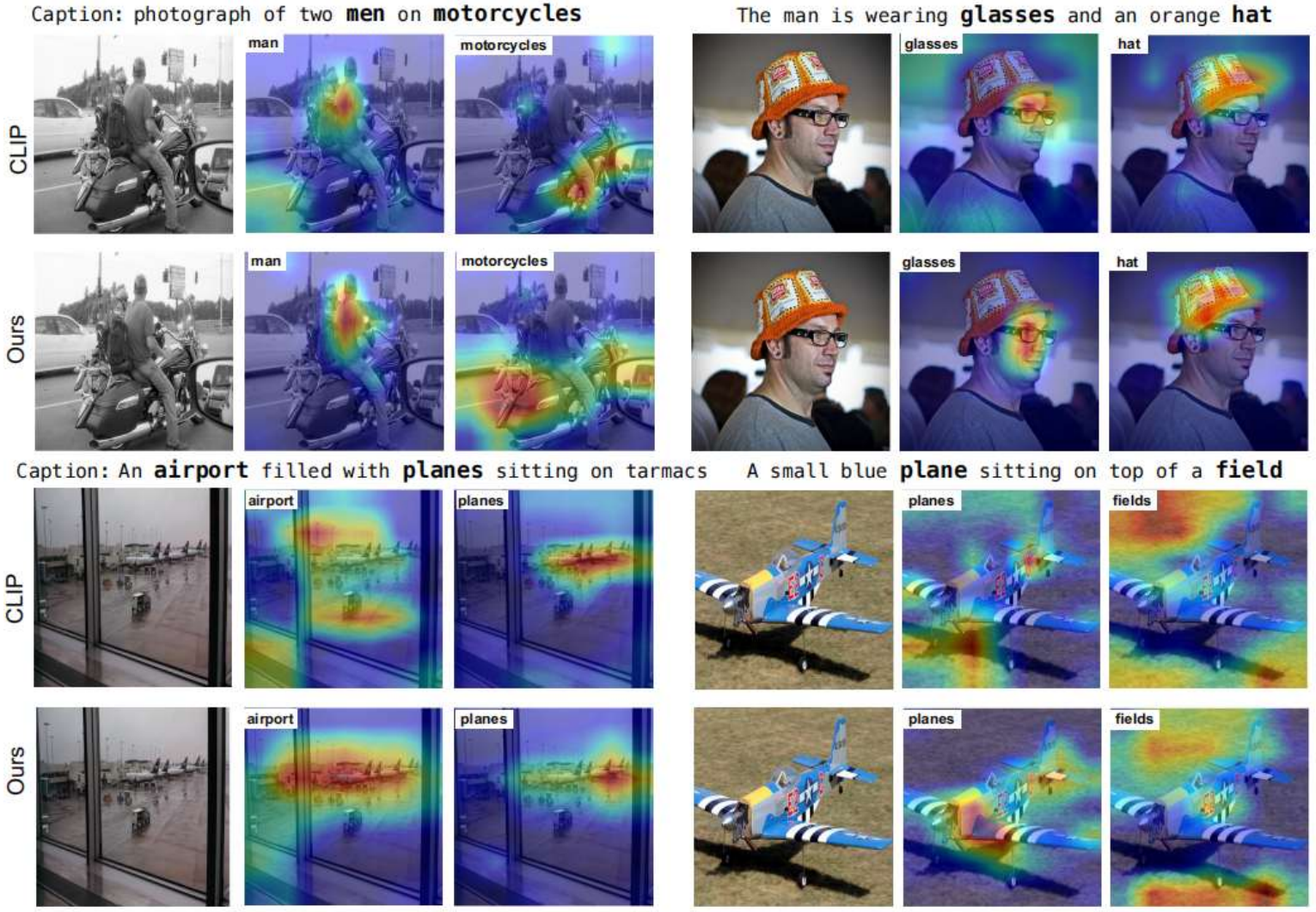


Figure 6. Scatter and histograms plots illustrating modality gaps of (a) CLIP and (b) RANKCLIP.

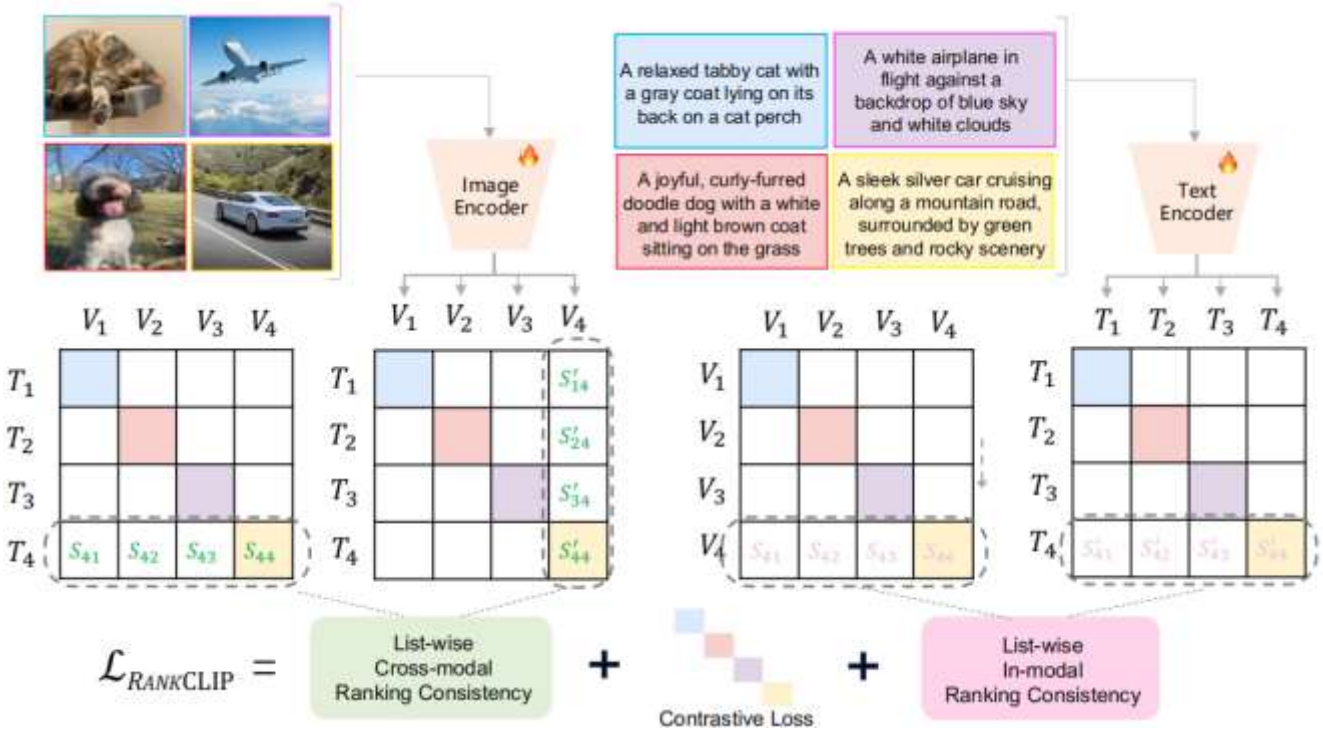
- 左上图, 针对CLIP和rankclip模型在不同数据规模下的消融实验。在ImageNet1K数据集上，从CC3M数据集中随机抽取不同规模数据时，rankclip的零样本top-1分类准确率始终显著优于CLIP。
- 右上图：在ImageNet1K（横轴）和ImageNet1K-R（纵轴）数据集上，零样本top-1分类准确率。rankclip展现出更优的鲁棒性和准确度。
- 左下图：展示嵌入向量在空间中的分布均匀性，高对齐度表明配对，嵌入向量间存在强关联性，低均匀性则意味着分布分散且效率低下。

效果可视化展示



MSCOCO数据集字幕中不同物体的RankCLIP和CLIP分类激活图对比显示：RankCLIP对部分名词的识别精度更高，且能精准定位名词对应的区域

RankCLIP



- RANKCLIP基于Plackett-Luce(PL)排名模型, 估计图像-文本对的排名优化模态一致性.
- $\pi(d|y_{1:k-1}, y_{ref}, D) = \frac{e^{s_{ij}}}{\sum_{d' \in D \setminus y_{1:k-1}} e^{s'_{ij}}}$
- $P(y, y_{ref}) = \prod_{k=1}^{K} \pi(y_k|y_{1:k-1}, y_{ref}, D)$
- $L_{PL} = -\log P(y|y_{ref})$
- $L_{cross-modal} = -\log P(y_{V-T}, y_{T-V})$
- $L_{in-modal} = -\log P(y_{V-V}, y_{T-T})$
- S_{ij}表示图文对(V_i, T_j)的CLIP得分(余弦值), y_{ref}表示参考排序
- $L_{cross-modal}$可以理解为学习了一个对称的余弦相似度矩阵

$$\mathcal{L}_{\text{RANKCLIP}} = \mathcal{L}_{\text{CLIP}} + \lambda_1 \mathcal{L}_{\text{in-modal}} + \lambda_2 \mathcal{L}_{\text{cross-modal}}$$

Part 03

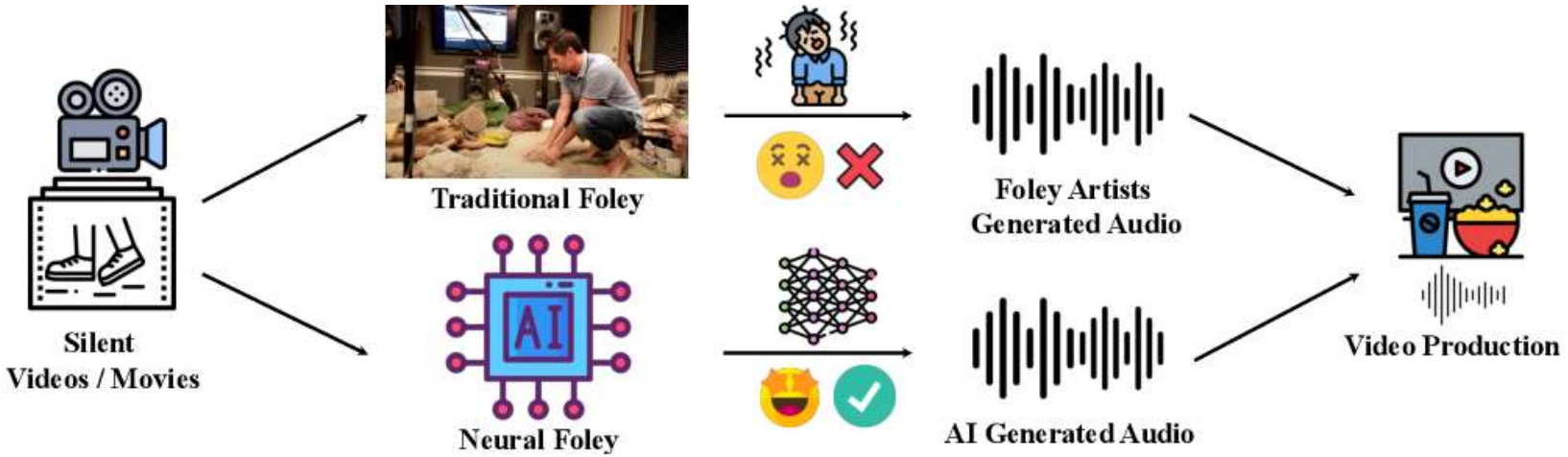
Diff-Foley

目录

研究背景

1.1 什么是视频-音频生成?

视频-音频生成 (Video-to-Audio Generation, V2A) 是指从无声视频自动生成与视觉内容语义一致、时序同步的音频内容的技术。



应用场景

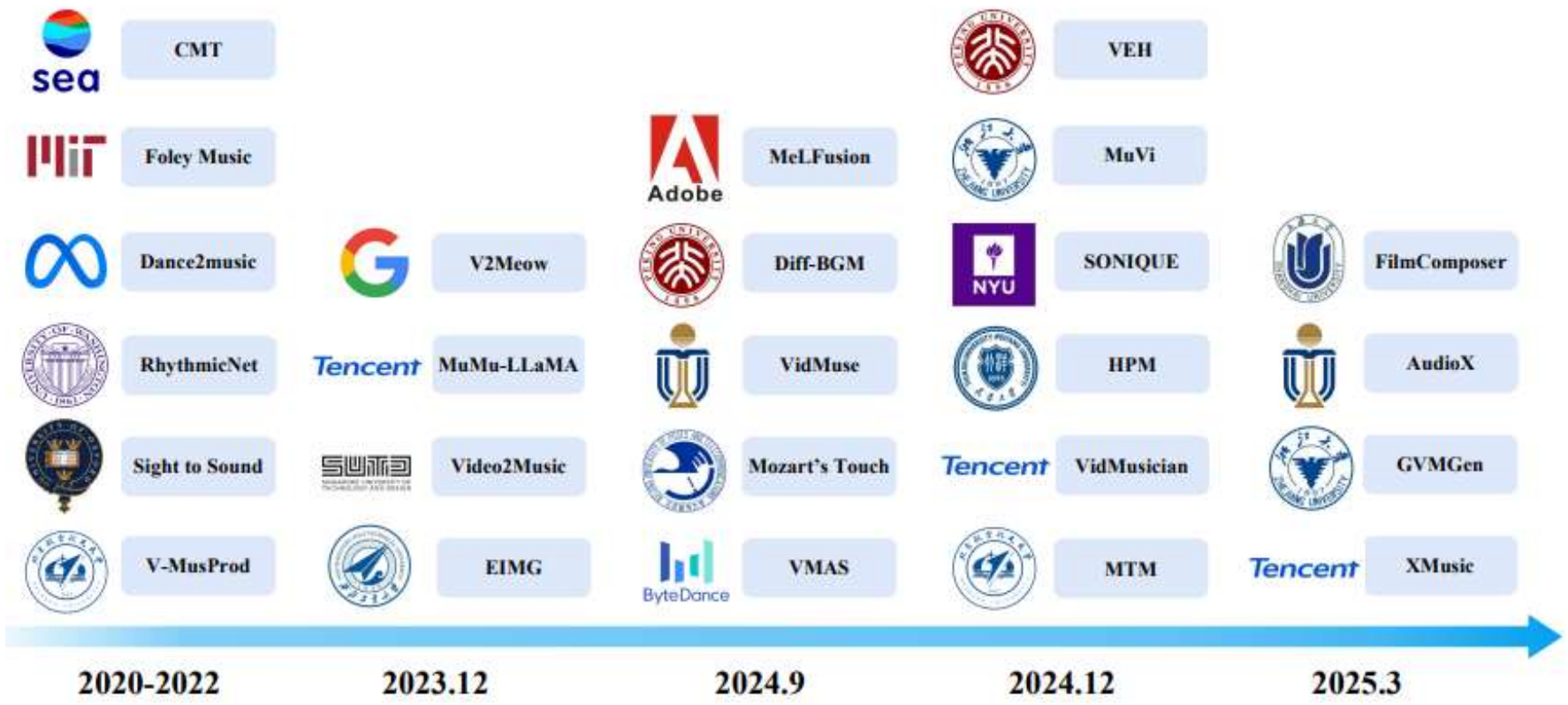
- 电影与视频制作
- 虚拟现实与元宇宙
- 游戏开发
- 辅助内容创作

核心挑战

- 语义一致性
- 音频质量
- 时序同步
- 多样性与真实感

研究背景

1.2技术演进历程



早期方法 (2016-2020)

基于传统机器学习和简单CNN,生成质量有限,缺乏时序建模能力。

GAN时代 (2018-2022)

利用GAN生成更真实的音频,但训练不稳定,模式崩溃问题突出。

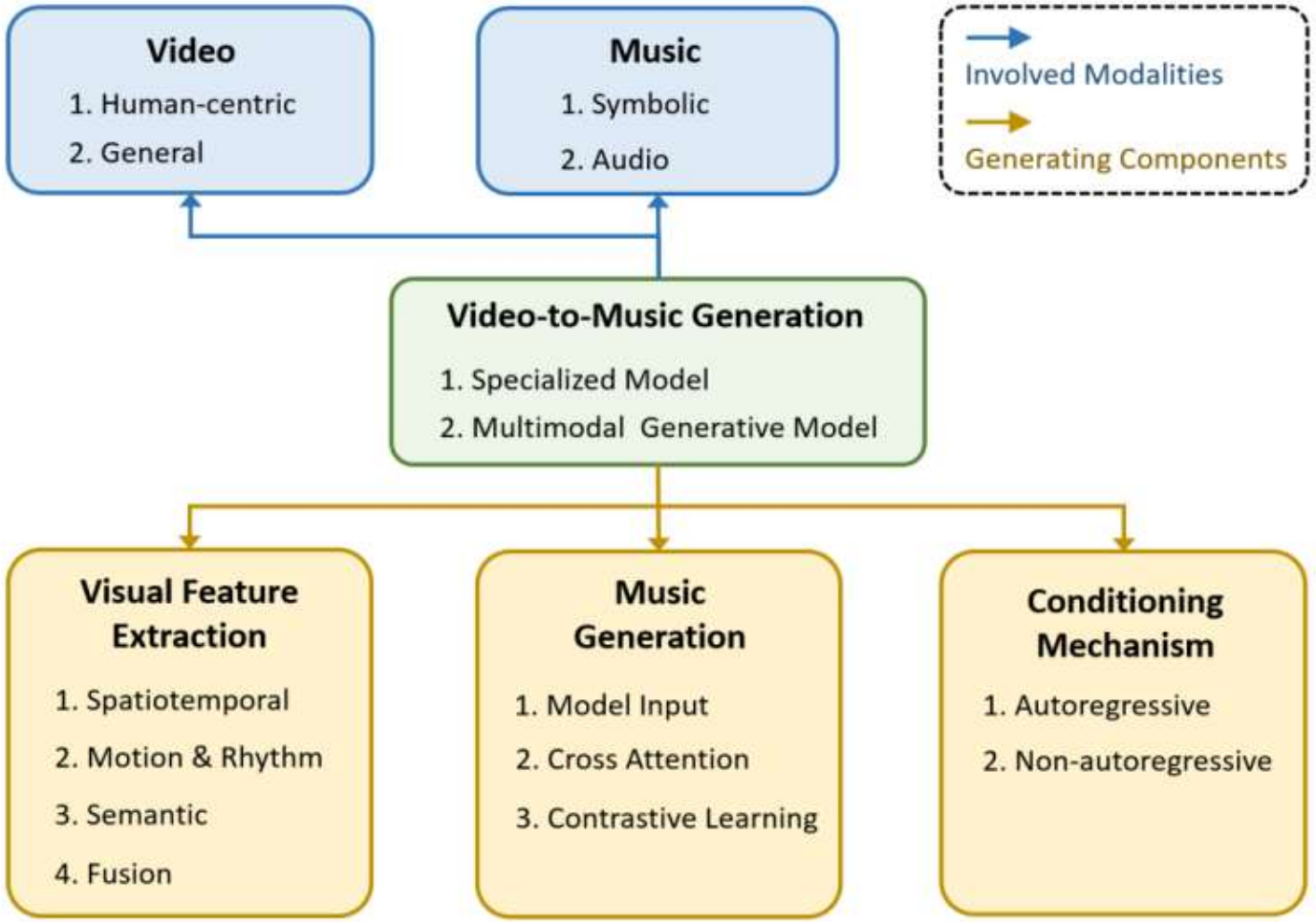
扩散、自回归模型 (2023-至今)

基于扩散模型的方法成为主流,显著提升生成质量和稳定性。

目录

方法综述

2.1 主流技术路线



1. 视觉特征提取

使用预训练视觉编码器(如CLIP、DINOv2)提取视频帧的视觉特征

2. 跨模态对齐

通过对比学习建立视觉-音频特征的语义对应关系

3. 条件音频生成

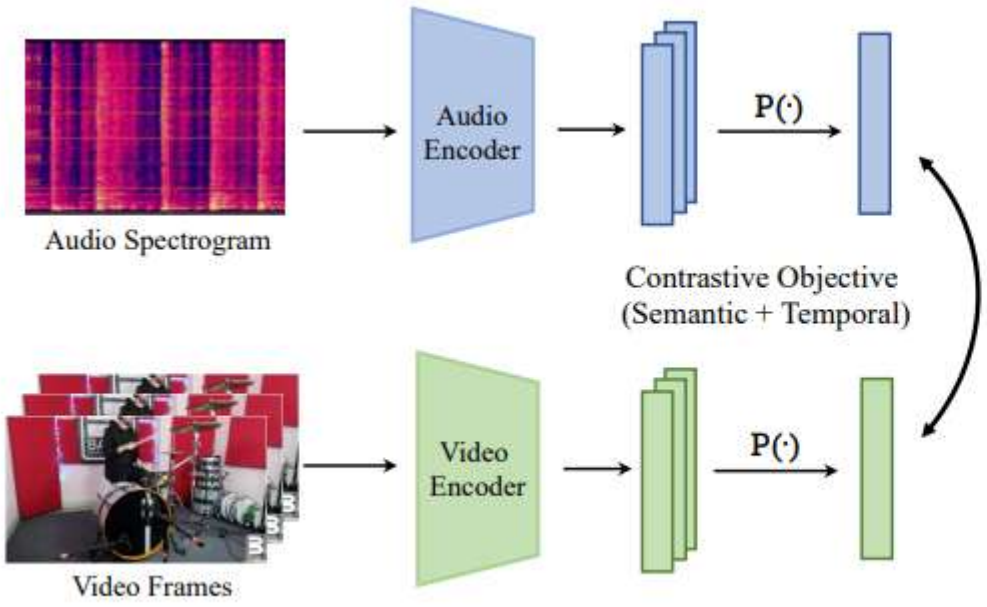
基于视觉条件,利用扩散模型或自回归模型生成音频

4. 时序同步优化

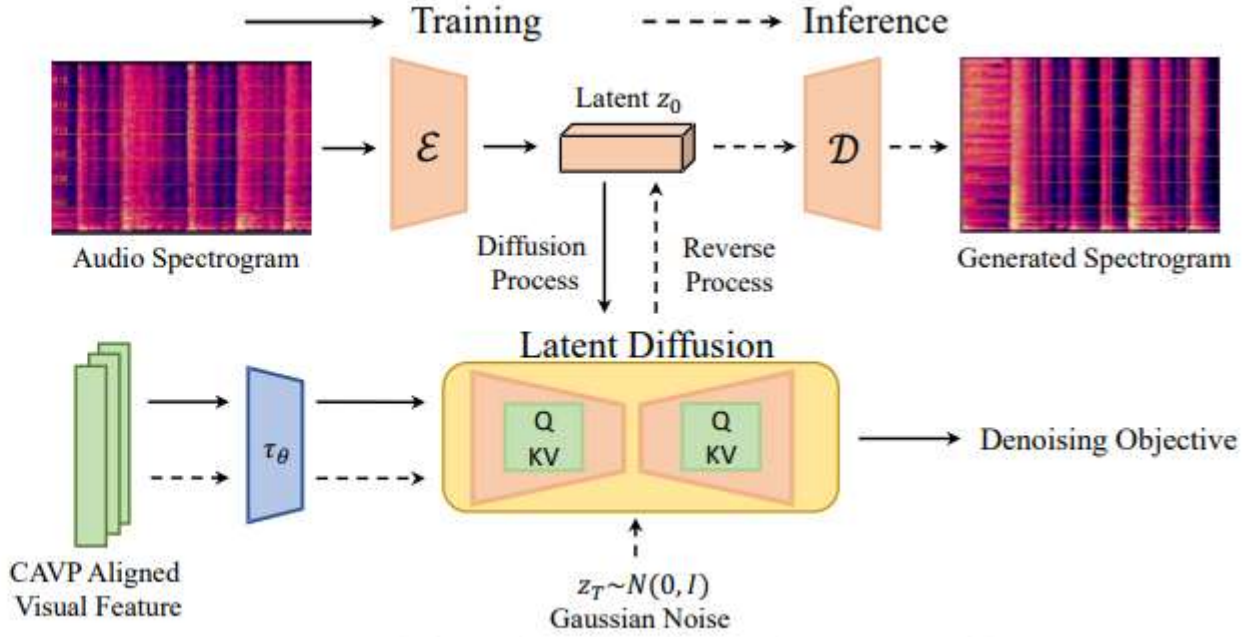
通过注意力机制和对齐损失确保音视频时序一致性

目录

代表工作——Diff-Foley



Contrastive Audio-Visual Pretraining (CAVP)



LDM Training with Aligned Visual Representation

Stage 1: CAVP

目标: 学习时序和语义对齐的音视频特征

方法: 对比学习最大化同视频内音视频特征的相似度

输出: 对齐的视觉编码器和音频编码器

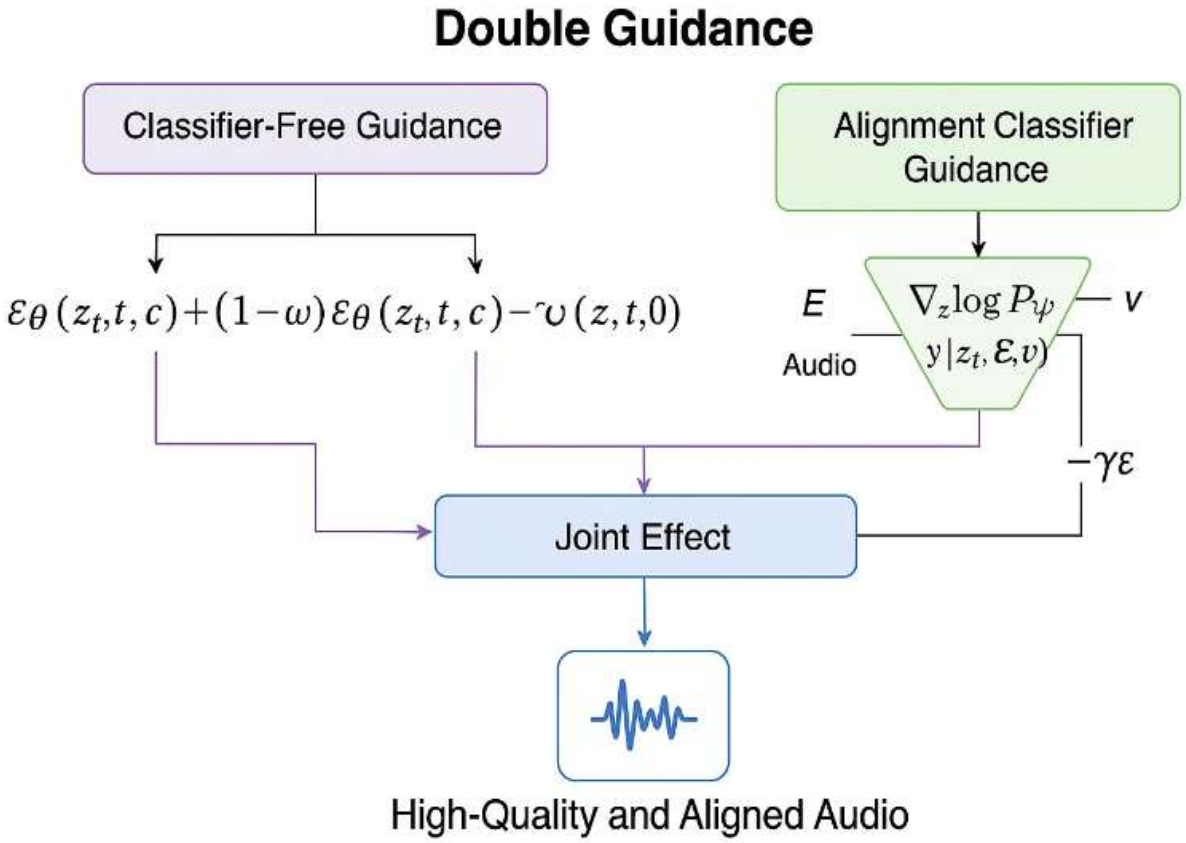
Stage 2: LDM训练

目标: 基于视觉条件生成高质量音频

方法: 在频谱潜在空间训练条件扩散模型

条件: CAVP对齐的视觉特征通过交叉注意力机制注入

代表工作——Diff-Foley



- ϵ_θ 为模型预测的噪声；
- $\omega \in [0,1]$ 为平衡参数，用于控制模型对条件 c 的依赖强度。

Double Guidance (双重引导)

Classifier-Free Guidance: 通过条件和无条件输出的差异增强条件控制

Alignment Classifier Guidance: 使用对齐分类器进一步提升音视频相关性

联合作用: 两种引导机制协同工作,实现高质量且对齐的音频生成

实验结果

Model	Visual Feature	FPS	Guidance	Metrics				Infer. Time↓
				IS ↑	FID ↓	KL ↓	Acc (%) ↑	
SpecVQGAN [21]	RGB + Flow	21.5	✘	30.01	**8.93**	6.93	52.94	5.47s
SpecVQGAN [21]	ResNet50	21.5	✘	30.80	9.70	7.03	49.19	5.47s
Im2Wav [41]	CLIP	30	CFG (✔)	39.30	11.44	**5.20**	67.40	6.41s
Diff-Foley (Ours)	CAVP	**4**	CFG (✔)	53.34	11.22	6.36	92.67	**0.38s**
Diff-Foley (Ours)	CAVP	**4**	Double (✔✔)	**62.37**	9.87	6.43	**94.05**	**0.38s**

主要结果

VGGSound数据集

IS达到62.37,大幅超越SpecVQGAN基线(IS 30.01)

时序同步性

对齐准确率显著提升,能准确捕捉关键时刻的声音事件

泛化能力

通过下游微调展示出良好的实用性和泛化能力

结论与局限

✓ SOTA性能

在V2A任务上达到当时最佳性能

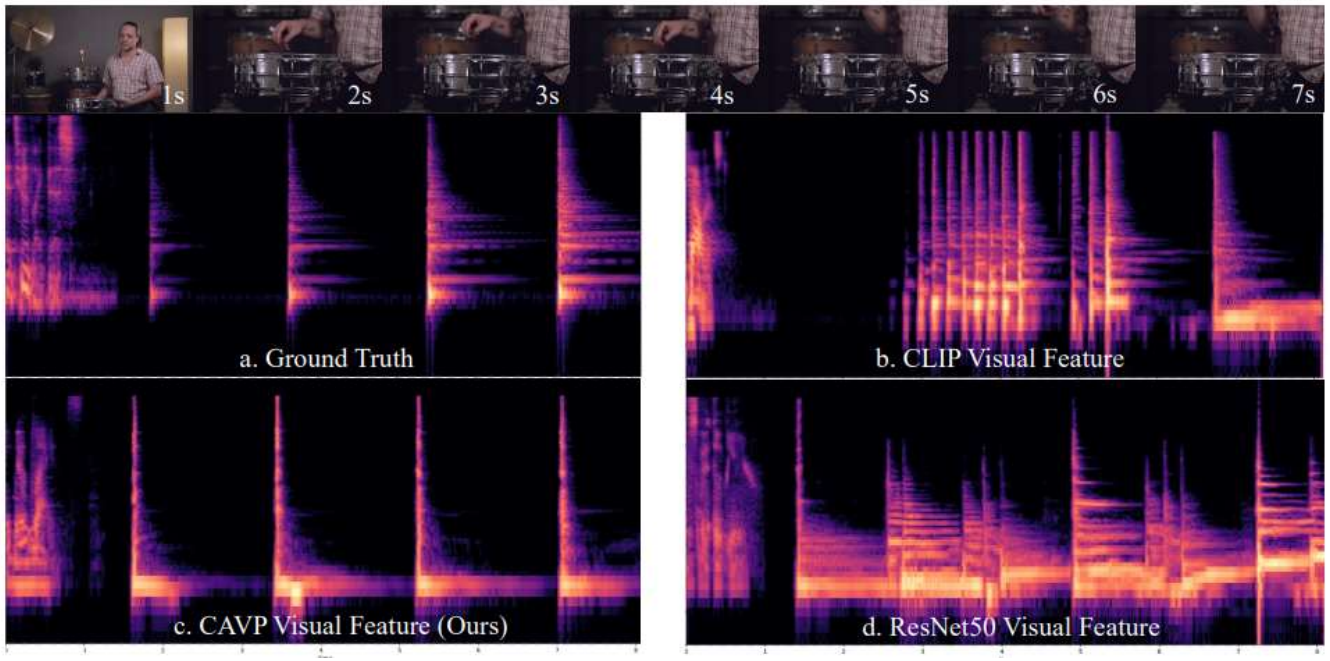
✓ 强时序对齐

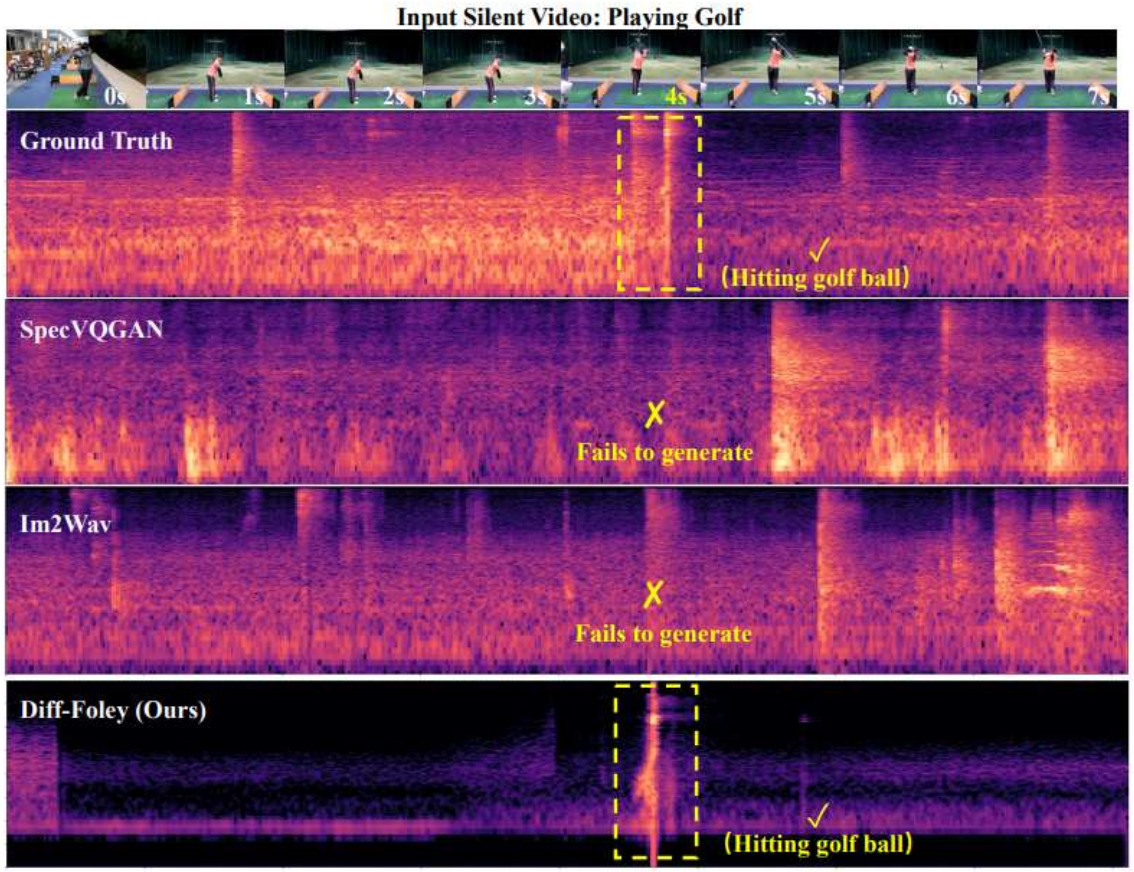
CAVP显著提升音视频时序同步性

⚠ 短视频局限

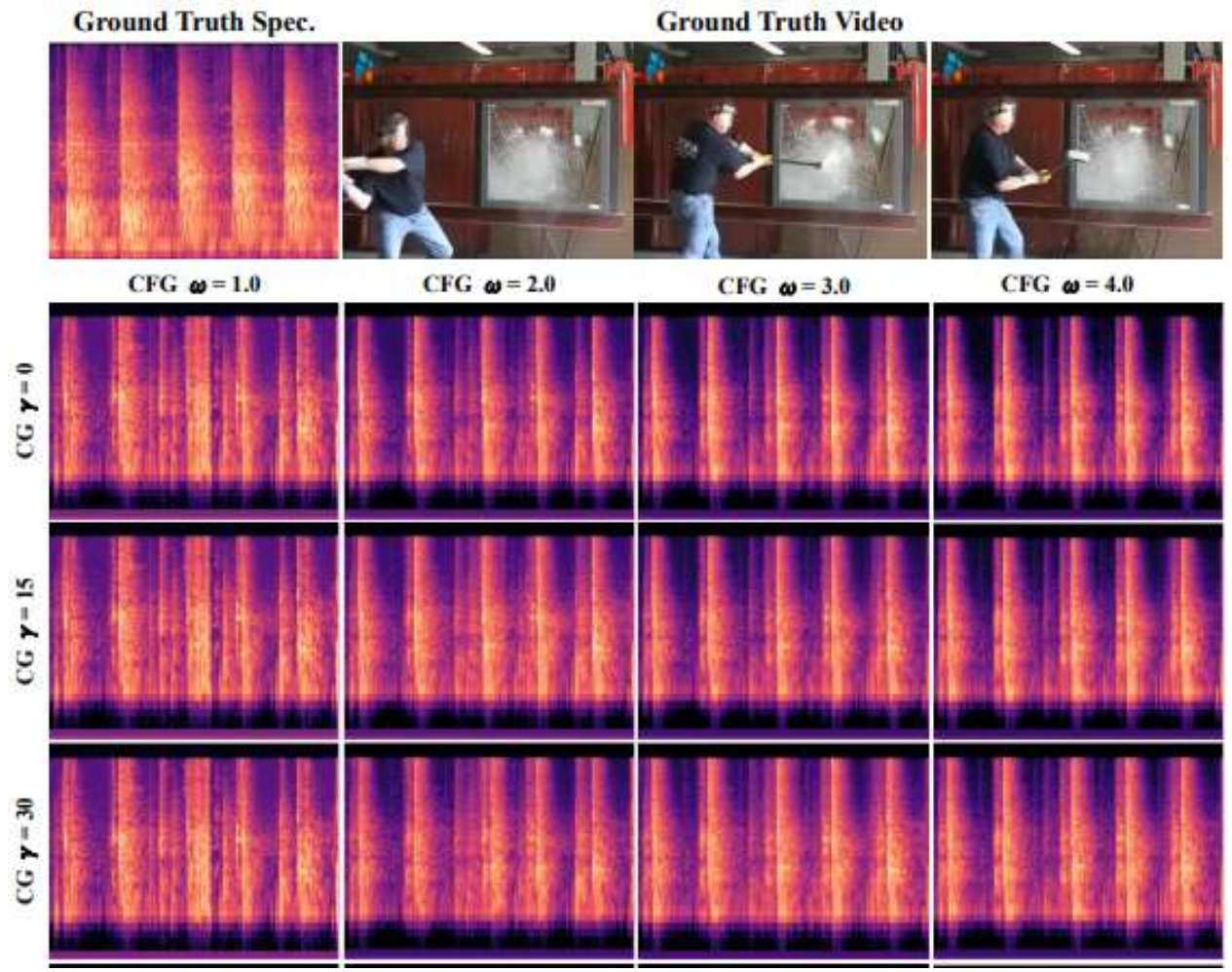
主要针对10秒以内短视频,长视频生成仍具挑战

音频可视化结果


1s
2s
3s
4s
5s
6s
7s
a. Ground Truth
b. CLIP Visual Feature
c. CAVP Visual Feature (Ours)
d. ResNet50 Visual Feature


Input Silent Video: Playing Golf
0s
1s
2s
3s
4s
5s
6s
7s
Ground Truth
(Hitting golf ball)
SpecVQGAN
Fails to generate
Im2Wav
Fails to generate
Diff-Foley (Ours)
(Hitting golf ball)

音频可视化结果


Ground Truth Spec.
Ground Truth Video
CFG ω = 1.0
CFG ω = 2.0
CFG ω = 3.0
CFG ω = 4.0
CG γ = 0
CG γ = 15
CG γ = 30

目录

复现结果

目录

未来展望

长视频生成

处理长时序依赖,实现分钟级视频的连贯音频

实时生成

优化推理速度,支持直播和互动应用

可控性增强

精细控制音乐风格、情感等属性

多模态融合

整合文本、深度等更多模态信息

技术趋势

- Transformer与扩散模型深度融合
- 整流流匹配加速推理
- 世界模型实现音视频联合建模

Part 04

RDP：基于视觉-触觉慢快学习的接触密集型操作策略

目录

研究背景

领域问题

人类能够借助视觉与触觉完成复杂的接触密集型任务，并具备高度反应能力，如快速响应外部变化和自适应接触力控制，但这对机器人而言仍具挑战性。

多数遥操作系统采用传统遥操作收集人类示范数据，难以提供细粒度触觉/力反馈，难以获取具有细粒度触觉反馈的高质量动作数据，这限制了可执行任务的类型。

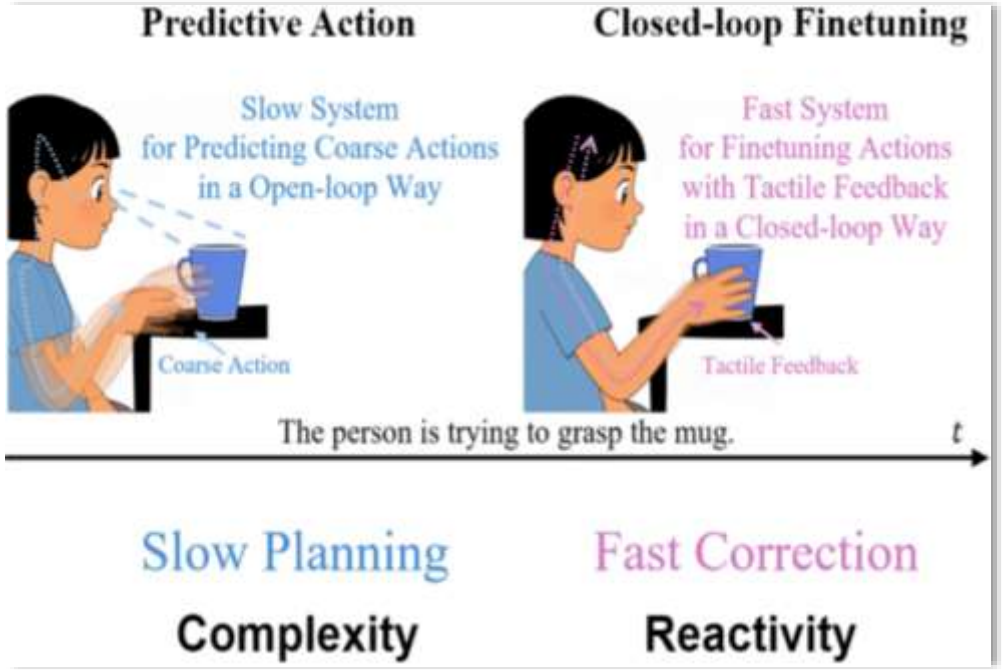
现有视觉模仿学习方法依赖**动作分块**来建模复杂行为，缺乏在分块执行期间实时响应触觉反馈的能力。

现有的一些工作如DP、π_0都是视觉模仿学习，神经网络预测/更新一次的频率都只有1-2Hz，虽然可以建模一些很复杂的行为但是难以快速响应一些触觉的信号。而一些位置/力的Controller(impedance /admittance control)则能够做到高达上千Hz控制频率但是无法完成一些复杂的动作依赖于人为指定。

目录

研究动机

神经科学领域的一些研究成果表明，人类在执行接触密集型任务时，其过程可分为两个组成部分：
1）前馈/预测性动作；
2）基于感官反馈（如触觉信号）的闭环微调。



受此启发，我们致力于开发一种能模拟人类在执行复杂接触密集型任务时控制模式的机器人学习系统。

Roland S Johansson and J Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks
Nature Reviews Neuroscience

目录

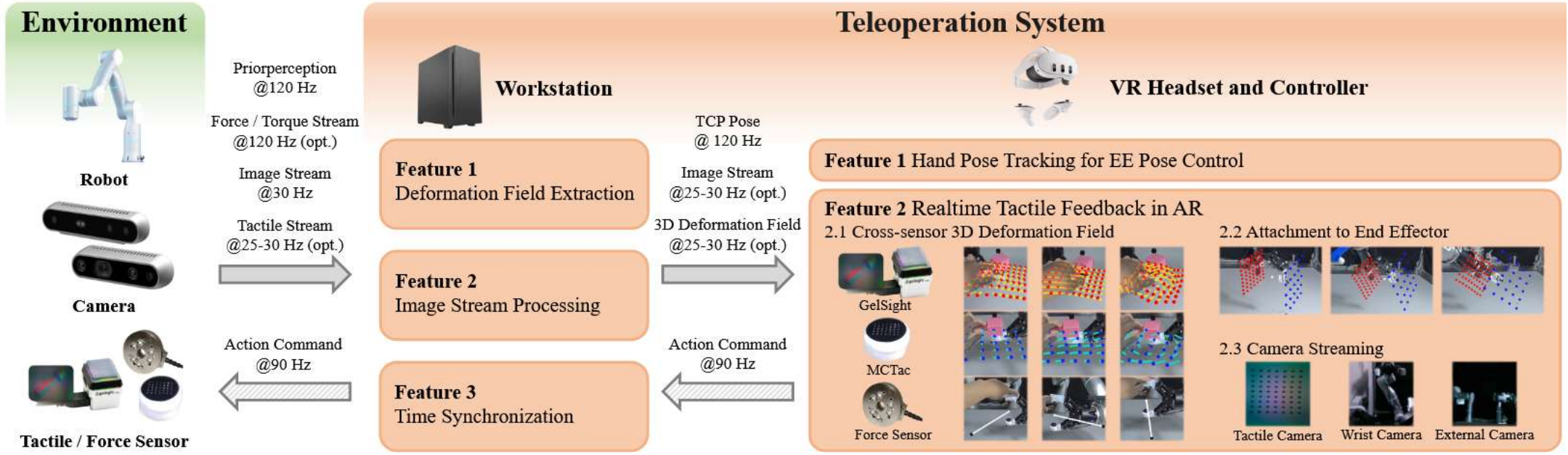
硬件平台

力触觉传感器介绍

Name	Gelsight Mini	Mc-Tac (customized)	force / torque sensor
Type	optical tactile	optical tactile	force / torque
Resolution	**very high (8MP)**	**high (2MP)**	low (6-dim)
Latency	high (>50ms)	medium (>20ms)	**low(<2ms)**
Durability	low	low	**high**

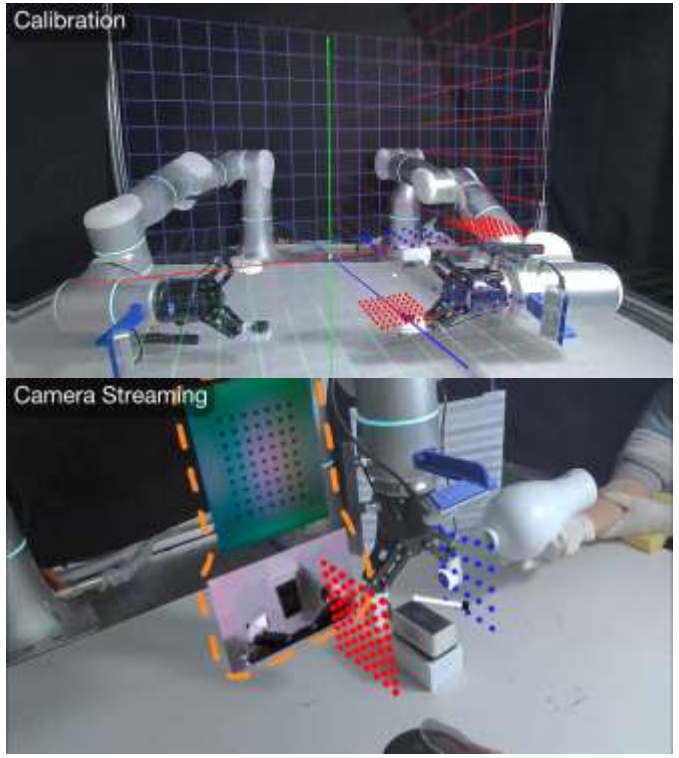
硬件平台

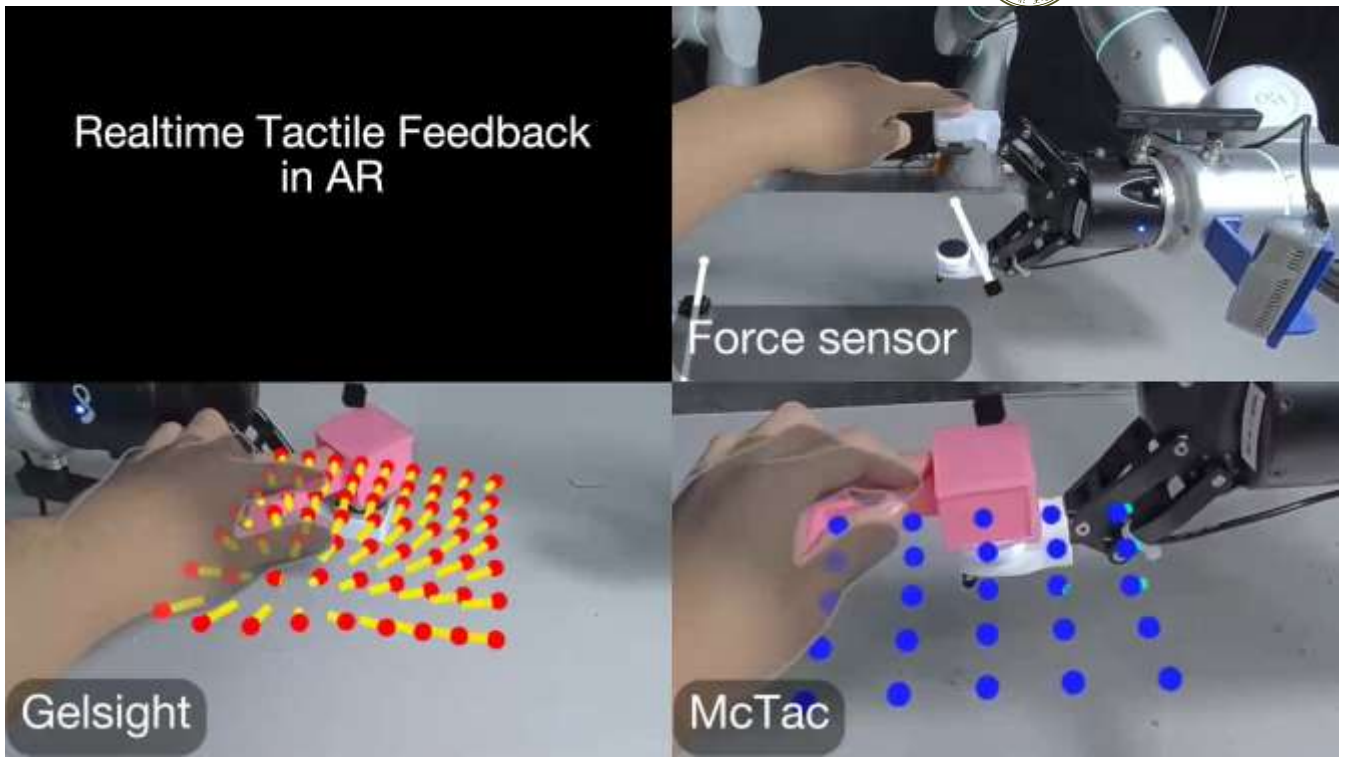
TactAR：一种通过AR提供实时触觉反馈的低成本遥操作系统



- **Meta Quest3**提供实时触觉/力反馈
- 采用三维形变场作为不同传感器触觉/力反馈的统一表征，在AR中渲染该形变场并将其绑定至虚拟空间的机器人末端执行器。
- 系统支持AR内触觉传感器与RGB相机的视频流传输，使用户在遥操作过程中能获取触觉图像、法向力、剪切力及扭矩等丰富的接触信息。

硬件平台



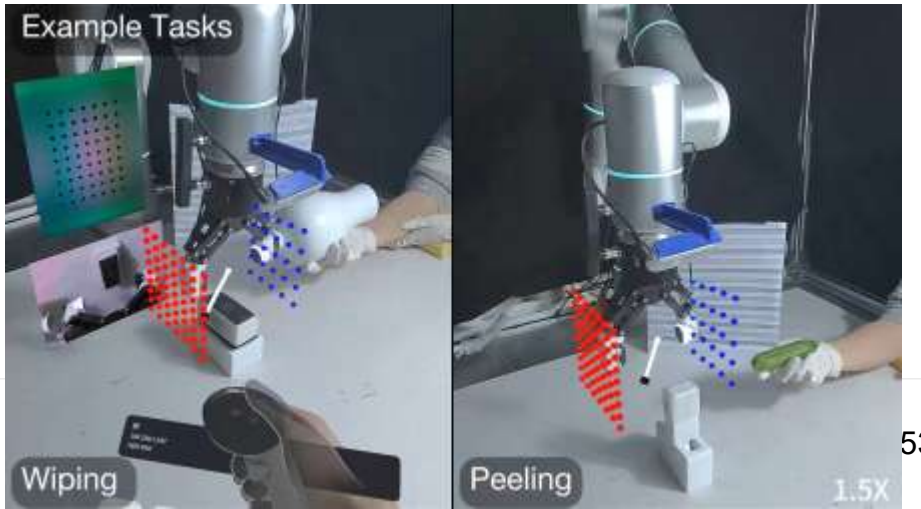


TactAR系统注重通用性与易用性

1.实时的力触可视化反馈；

2.支持多类触觉/力/电/磁传感器，可轻松部署于不同机器人实体；

3.具备高性价比。Meta Quest3仅需500美元

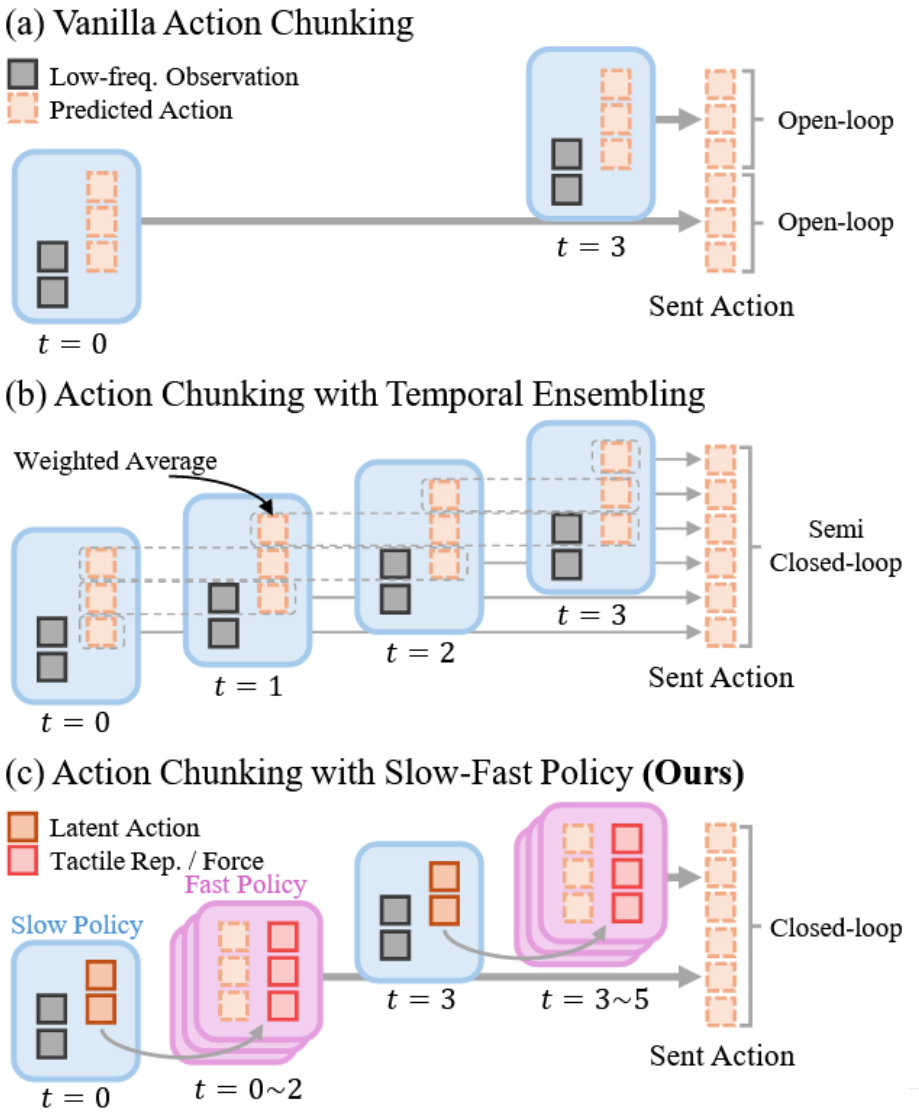


目录

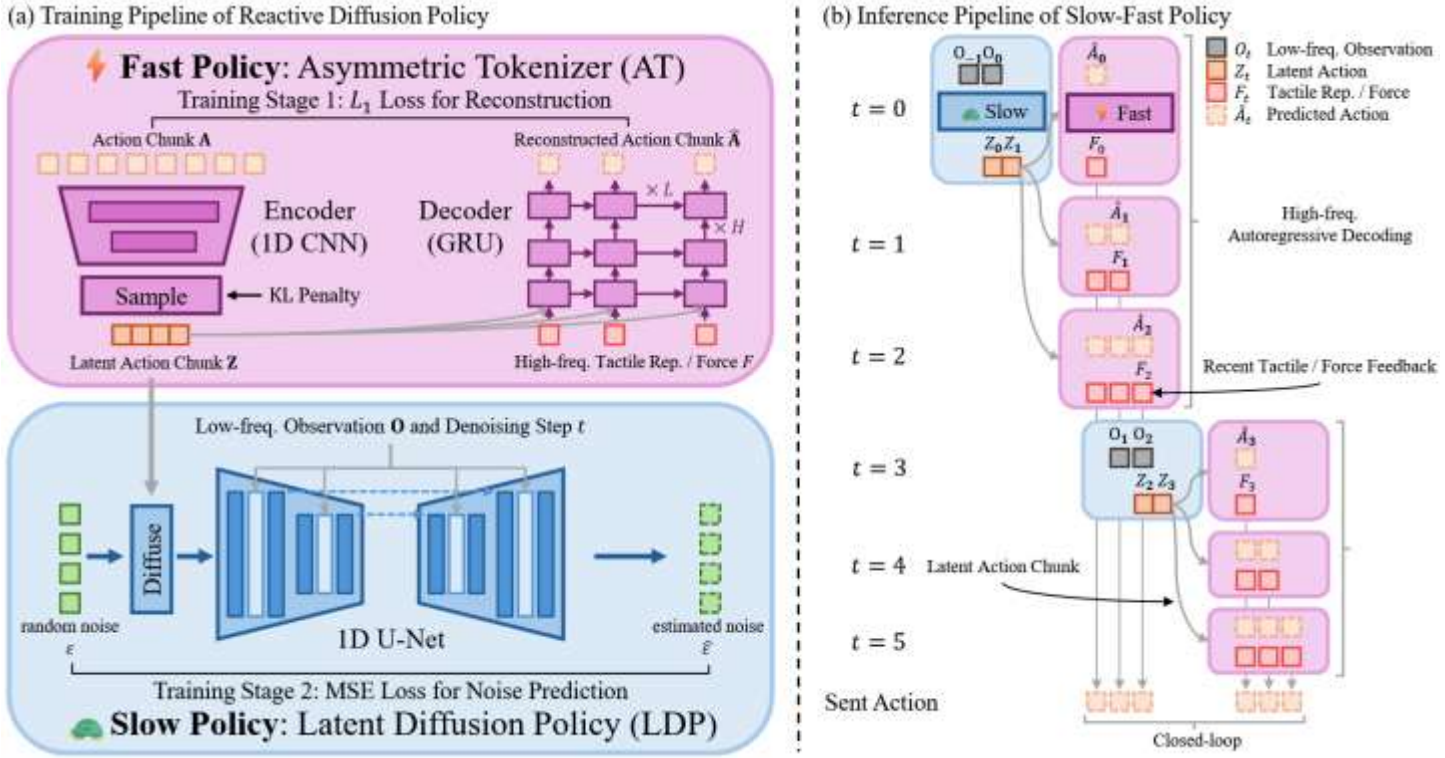
模型算法

RDP采用**双层架构：**
（1）在低频下预测潜在空间中高层动作分块的慢速潜在扩散策略；
（2）在高频下实现闭环触觉反馈控制的快速非对称分词器

先前研究表明，预测动作序列或动作块能有效保持动作时序一致性并处理非马尔可夫动作或空闲动作，从而提升策略学习性能。但此类方法在执行动作块时相当于**开环策略**，无法实时响应触觉等高频信号。VISK采用**时序集成**缓解该问题，如图5所示，**该方法通过对同一时间步的多次预测结果进行聚合，在闭环控制与序列一致性间取得平衡。**但该方案会削弱策略对多模态分布和非马尔可夫动作的建模能力，易导致策略**陷入局部最优。**此外，策略性能对**时序集成的平滑系数非常敏感**，限制了其适用性。



模型算法

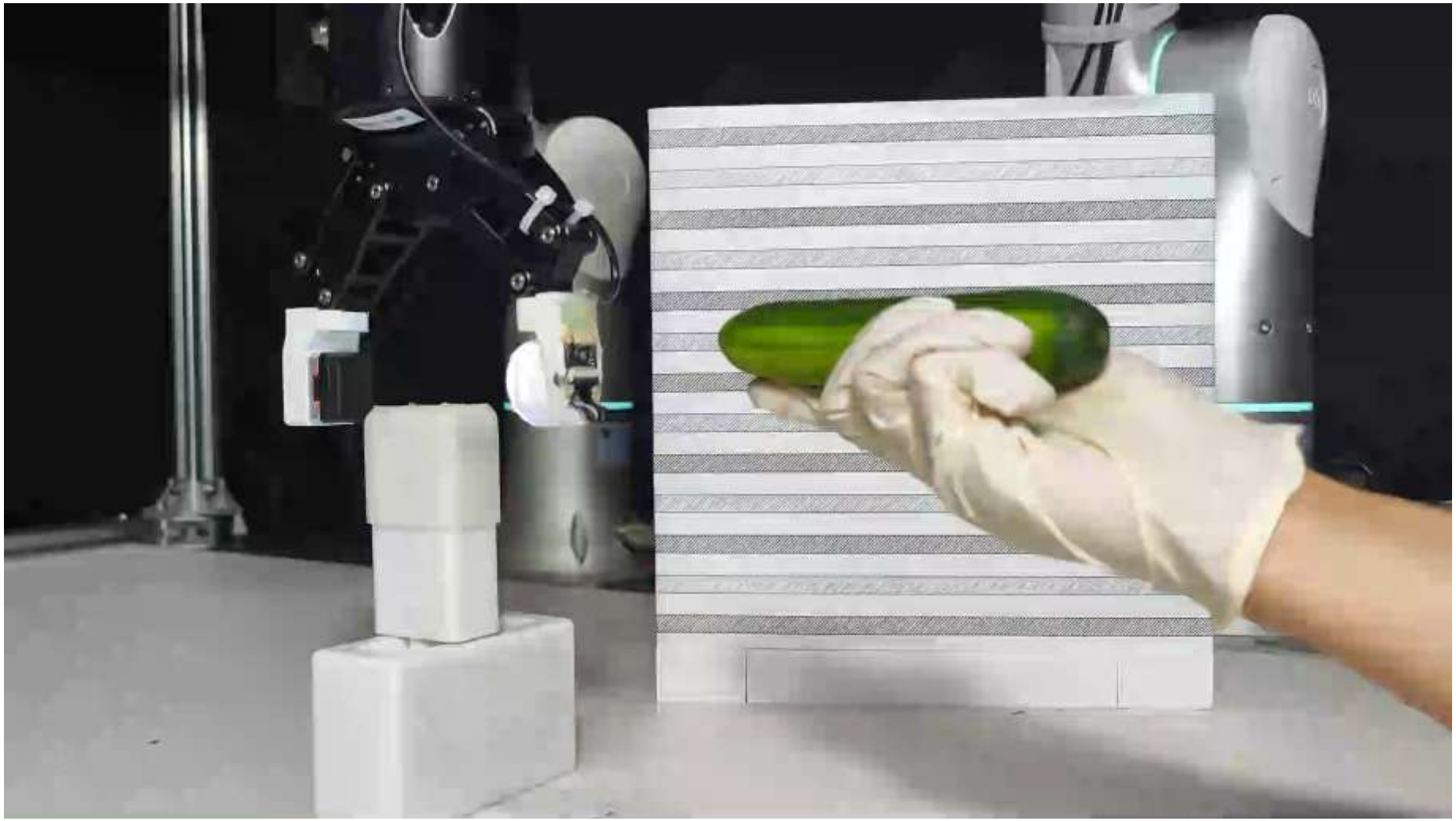


采用**相对末端执行器轨迹**进行**动作表征**：不直接计算连续帧间的增量动作（可能导致较大累积误差），而是通过计算相对于基准

首先通过训练**非对称分词器**将原始动作块编码至潜空间，其解码器将瞬时触觉表征与潜动作块分离作为输入；
非对称的含义是**仅在解码器中将触觉表示作为输入。**这种刻意设计的结构不对称性是为了确保**潜在动作片段仅保留高层反馈策略，而精确位置则由解码器借助触觉信息进行预测。**
策略学习阶段，慢速潜扩散策略(LDP)以类似**扩散策略**的方式根据观测数据预测潜动作块；下采样后的潜在表示降低了计算成本；
AT中的不对称设计可将具有挑战性的反应行为排除在潜在动作片段之外，从而降低潜在扩散策略在低频观测下的学习难度并增强其泛化能力
推理阶段低频采样潜动作块，在每个块内执行动作时，最新触觉表征实时输入AT解码器以预测下一帧实际动作

目录

实验结果与结论



实验结果与结论

TABLE II: Policy Performance for Peeling Task

	No Perturb.	Perturb. before Contact	Perturb. after Contact	All
DP	0.56	0.58	0.19	0.44
DP w. tactile img.	0.60	0.49	0.16	0.41
DP w. tactile emb.	0.48	0.55	0.15	0.39
RDP (GelSight)	0.98	0.93	0.80	0.90
RDP (MCTac)	**1.00**	0.84	0.79	0.88
RDP (Force)	0.99	**0.98**	**0.88**	**0.95**

TABLE I: Inference Time of Different Modules on RTX 4090

Diffusion Policy	Slow Policy (LDP)	Fast Policy (AT)
120ms	100ms	< 1ms

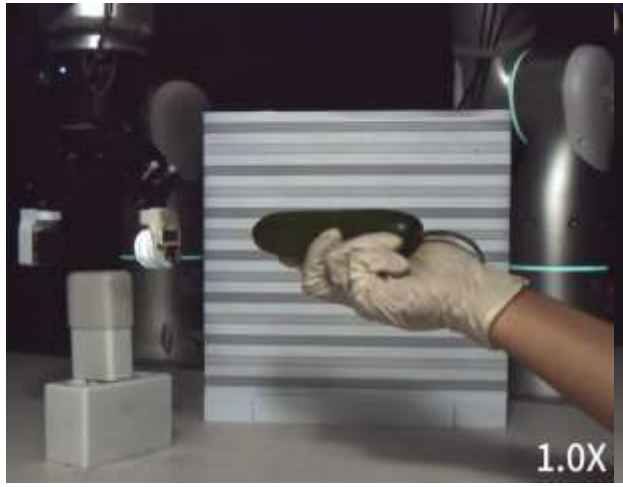
TABLE III: Policy Performance for Wiping Task

	No Perturb.	Perturb. before Contact	Perturb. after Contact	All
DP	0.75	0.70	0.25	0.57
DP w. tactile emb.	0.60	0.75	0.15	0.50
RDP (GelSight)	0.85	**0.95**	0.50	0.77
RDP (Force)	**0.95**	0.85	**0.80**	**0.87**

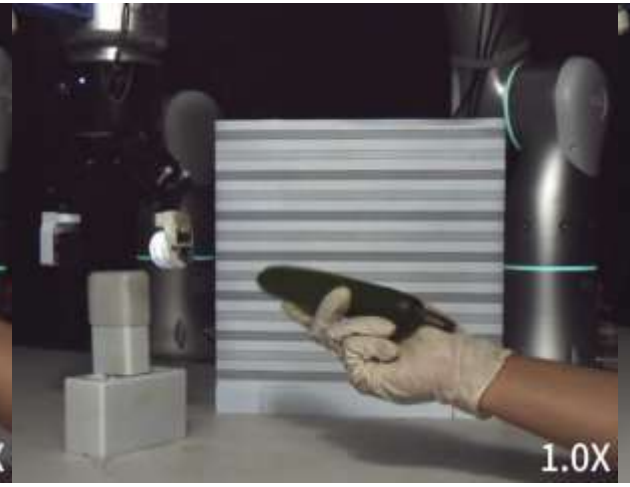
TABLE IV: Policy Performance for Bimanual Lifting Task

	Soft Paper Cup			Hard Paper Cup			All
	Clamp	Lift	Score	Clamp	Lift	Score	Score
DP	0%	0%	0.00	0%	0%	0.00	0.00
DP w. tactile emb.	10%	10%	0.10	20%	10%	0.05	0.08
RDP (GelSight + MCTac)	**100%**	**100%**	0.55	**90%**	80%	0.40	0.48
RDP (Force)	**100%**	90%	**0.80**	**90%**	**90%**	**0.60**	**0.70**

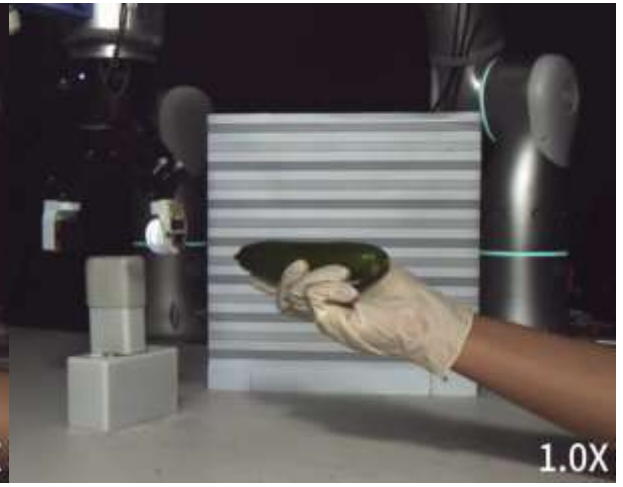
实验结果与结论



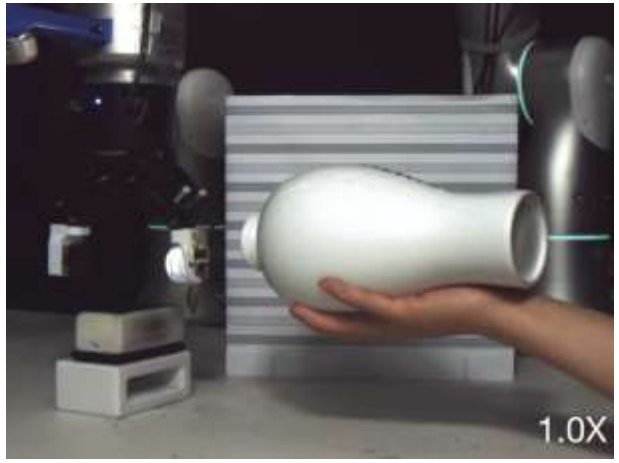
DP w. tactile emb.



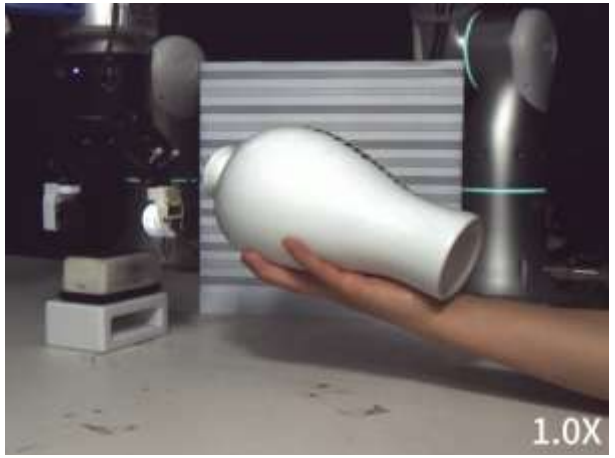
DP w. tactile emb.



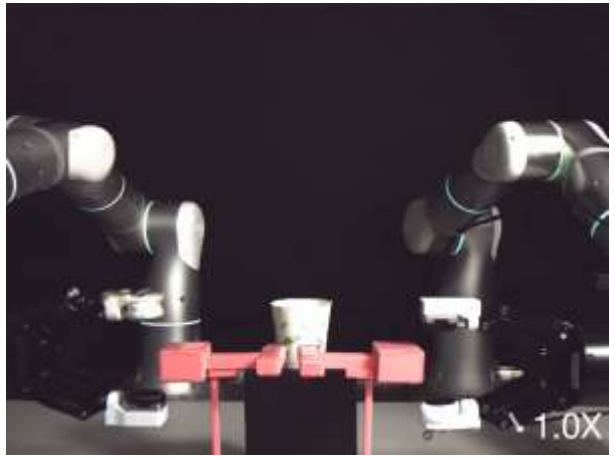
DP w. tactile img.



DP w. tactile emb.

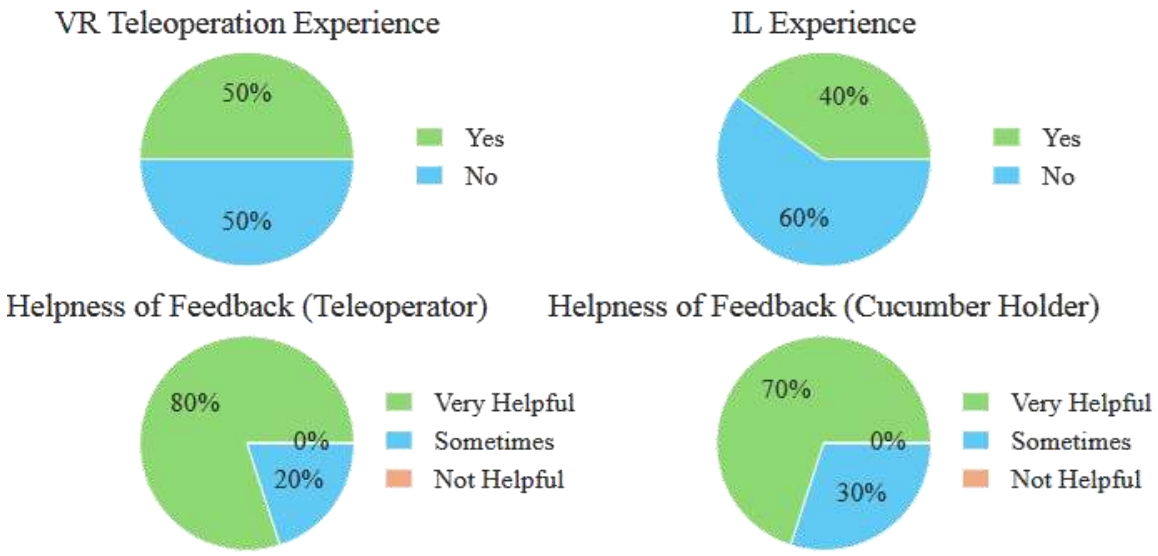


DP

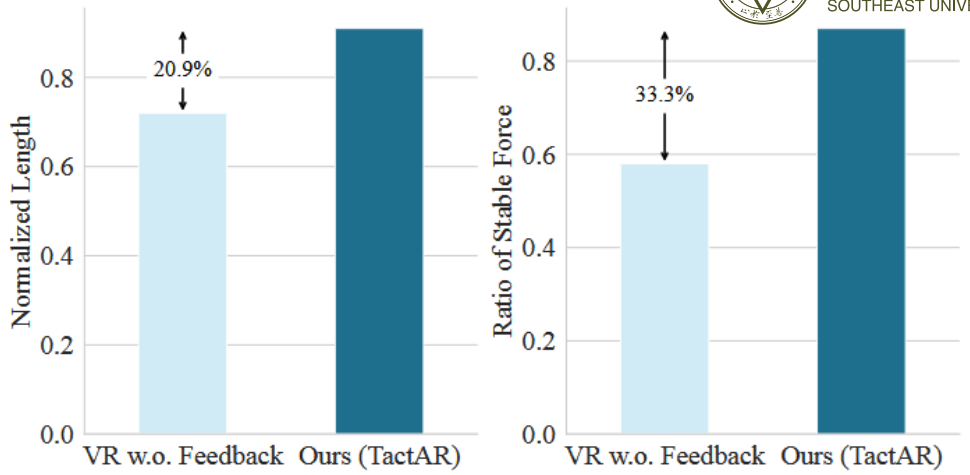


DP w. tactile emb.

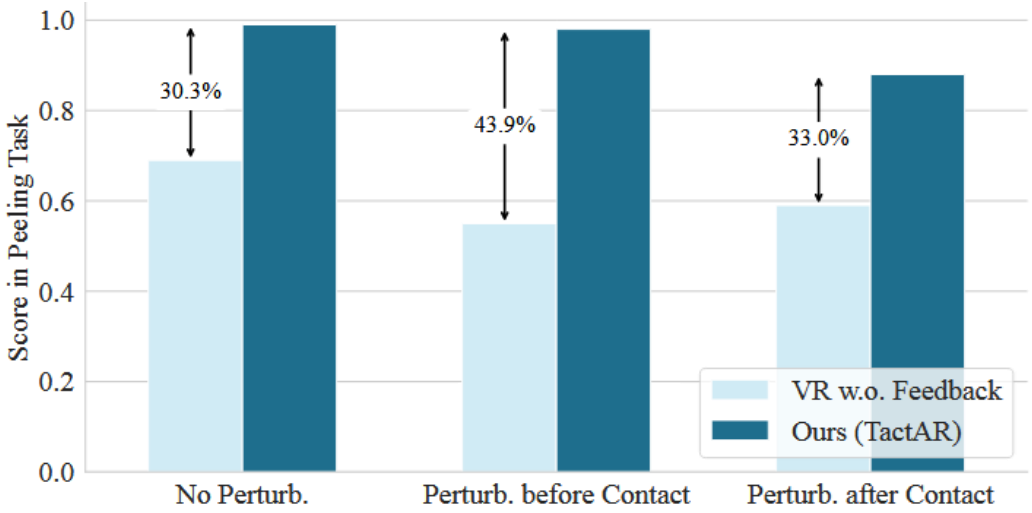
实验结果与结论



用户调研



数据质量对比



策略表现对比

目录

局限性与未来工作

1.TactAR系统专为二指夹爪设计
-> 扩展至灵巧手中
2.RDP算法中的快速策略目前仅能响应高频触觉/力输入信号，尚无法快速处理高频图像输入
-> 加入视觉输入，可以实现需要高频视觉反应的任务如打乒乓球
3.RDP算法目前仅适用于单任务场景
-> 扩展至多任务的VLA
4.虽然TactAR能在AR中提供一定程度的触觉/力反馈，但其直观性和效率仍不及人手直接操作
->进一步降低传感器与系统延迟来提升远程操作效率

Q 问答 & A

请老师同学们批评指正

汇 报 人：寇硕、林骅、
李文卓、兰红星

2025/10/24