专题综述

多模态生成式 AI 探索：从数据合成到内容创造

同济大学 高俊尧 宋子帆 齐鼎 赵才荣

本文是同济大学VILL实验室在多模态生成领域的一系列工作成果，其中语言/视觉数据生成的相关工作发表在NIPS 2025[1]以及CVPR 2025（做口头汇报）[2]，图像/视频生成的相关工作在github上收获了超过400个stars，并发表在TPAMI 2025[3]和ICLR 2025[4]。随着多模态生成式人工智能的持续演进，从“数据学习”向“内容创造”的范式转变已经成为推动新一轮技术革命的关键驱动力。

在语言数据生成方面，VILL实验室重点关注开源大型语言模型（LLMs）在代码生成方向的精调能力提升，针对当前主流Code LLMs常在单一来源的数据集上进行微调，受限于语言数据质量和风格单一性，无法充分激发预训练模型的泛化能力这一问题，提出了AlchemistCoder框架，从根本上重构了代码微调数据构建的方式。在视觉数据生成方面，VILL实验室重点关注数据集蒸馏（Dataset Distillation, DD）在图像分类之外任务中的泛化与适应能力，创新性地将任务知识挖掘与扩散模型（Diffusion Models）生成过程深度融合以克服现有方法在目标检测与图像分割中的任务局限性，所提出的通用视觉数据蒸馏框架UniDD借助合成高质量的视觉数据，在提升数据利用效率的同时，也为低资源场景下的视觉模型训练提供了新范式。在图像生成方面，VILL实验室聚焦开放域的图像风格迁移，提出StyleShot通过构建一个风格感知编码器和一个大规模的风格数据集，高效地提取丰富的风格表示，并结合内容融合编码器以增强图像驱动的风格迁移能力。在视频生成方面，VILL聚焦现有肖像动画方法在非人类角色（如表情包、玩偶等）时常常发生动画效果失真甚至失败这一问题，提出一个无需训练的肖像动画框架FaceShot，利用扩散模型的强大语义对应关系来生成各种角色类型的动画结果。

一、研究背景

1.1 大语言模型

近年来，大语言模型在自然语言处理领域取得了突破性进展。得益于 Transformer 架构和大规模数据的训练，LLMs 展现出强大的语言理解与生成能力，不仅在对话系统、文本摘要、机器翻译、信息抽取等任务中均表现优异，在代码领域的特定变体（Code LLMs）也取得了显著进展。以往的 Code LLMs 通常在单一来源的数据集上进行微调，这些数据集在质量和多样性上存在局限，这可能无法充分激发预训练 LLMs 的潜力。研究指出，尽管多源数据融合是提升模型能力的关键，但草率地混合不同来源的代码语料库，会因其固有的风格、质量和编程范式冲突，反而导致模型性能下降。因此，如何有效整合多源数据，克服其内在冲突，以充分释放基础模型的代码智能，并进一步提升模型的泛化能力，是当前代码大模型微调领域面临的核心挑战。

1.2 扩散模型

扩散模型近年来在生成建模领域取得了显著进展，逐渐成为继 GAN 和 VAE 之后的主流生成方法。其核心思想源于非平衡热力学过程，通过逐步向数据添加噪声构建前向过程，并在反向过程中学习去噪还原数据分布。与其他生成模型相比，扩散模型在图像、音频、3D 建模等多种模态上表现出更强的生成质量和更高的多样性，

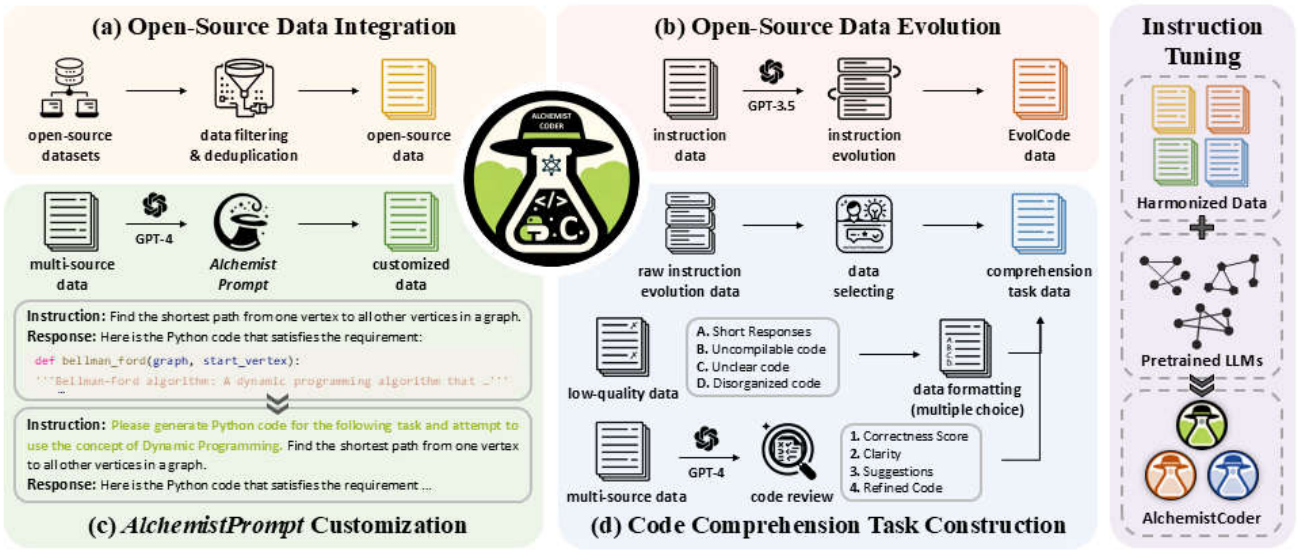

图 1　AlchemistCoder 框架图

尤其在高分辨率图像生成任务中已超越 GAN 的效果。借助扩散模型更清晰的理论基础和更稳定的生成效果，本文挖掘扩散模型在视觉数据生成、图像/视频生成上的进一步潜力，积极探索与任务知识结合的生成范式，推动其在复杂视觉任务中的实用性提升。

二、数据生成

2.1 语言数据生成

首先，为解决多源数据的内在冲突，该研究开创性地引入了 AlchemistPrompts。这是一种基于“事后重标签（hindsight relabeling）”思想的数据特异性提示（data-specific prompts）。如图 1 所示，具体而言，该研究使用一个强大的模型（如 GPT-4）扮演“炼金术士（Alchemist）”的角色，回顾并分析每个“指令-代码响应”数据对，然后生成一个能更精确、更细致地描述该代码响应特点的新指令。例如，如果原始指令是“编写一个最短路径算法”，而代码实现是 Python 版的贝尔曼-福特算法，AlchemistPrompt 会将其优化为“请使用 Python 语言并结合动态规划思想，为以下任务生成代码”。这种方法不仅通过统一的风格协调了不同数据源之间的差异，还通过增强指令与响应的对齐度，将模型的学习过程从“为相似问题克隆不同答案”转变为“学习遵循多样化的精确指令”。

其次，该研究提出将“数据构建过程”本身也作为训练任务，以提升模型的代码理解（code comprehension）能力。除了传统的代码生成任务，他们设计了三项新的代码理解任务并构建了相应数据：1）指令演进（instruction evolution），让模型学习如何将简单指令优化得更复杂、更明确；2）数据过滤（data filtering），向模型展示低质量代码（如编译错误、不合规等）的反例，训练其辨别和避免生成劣质代码；3）代码审查（code review），要求模型评估代码的正确性和清晰度、提出修改建议并给出优化后的代码。

Model	Params	Base Model	HumanEval (+)	MBPP (+)	Average (+)
Closed-source Models					
GPT-3.5-Turbo [33]	-	-	72.6 (65.9)	81.7 (69.4)	77.2 (67.7)
GPT-4-Turbo [34]	-	-	**85.4 (81.7)**	**83.0 (70.7)**	**84.2 (76.2)**
Open-source Models					
Llama 2-Chat [40]	70B	Llama 2	31.7 (26.2)	52.1 (38.6)	41.9 (32.4)
CodeLlama-Python [35]	70B	Llama 2	57.9 (50.0)	72.4 (52.4)	65.2 (51.2)
CodeLlama-Instruct [35]	70B	CodeLlama	**65.2 (58.5)**	**73.5 (55.1)**	**69.4 (56.8)**
CodeLlama-Python [35]	34B	Llama 2	51.8 (43.9)	67.2 (50.4)	59.5 (47.2)
WizardCoder-CL [30]	34B	CodeLlama-Python	73.2 (56.7)	73.2 (51.9)	73.2 (54.3)
DeepSeek-Coder-Instruct [14]	33B	DeepSeek-Coder-Base	**78.7 (67.7)**	**78.7 (59.7)**	**78.7 (63.7)**
StarCoder [22]	15B	-	34.1 (33.5)	55.1 (43.4)	44.6 (38.5)
CodeLlama-Python [35]	13B	Llama 2	42.7 (36.6)	61.2 (**45.6**)	52.0 (41.1)
WizardCoder-SC [30]	15B	StarCoder	**51.9 (45.7)**	**61.9** (44.9)	**56.9 (45.3)**
Llama 2 [40]	7B	-	14.0 (10.4)	26.1 (17.5)	20.1 (14.0)
StarCoder [22]	7B	-	24.4 (21.3)	33.1 (29.2)	28.8 (25.3)
CodeLlama-Python [35]	7B	Llama 2	37.8 (33.5)	57.6 (42.4)	47.7 (38.0)
WizardCoder-CL [30]	7B	CodeLlama-Python	48.2 (42.1)	56.6 (42.4)	52.4 (42.3)
DeepSeek-Coder-Base [14]	6.7B	-	47.6 (41.5)	70.2 (53.6)	58.9 (47.6)
Magicoder-CL [44]	7B	CodeLlama-Python	60.4 (49.4)	64.2 (46.1)	62.3 (47.8)
MagicoderS-CL [44]	7B	CodeLlama-Python	70.7 (60.4)	68.4 (49.1)	69.6 (54.8)
Magicoder-DS [44]	6.7B	DeepSeek-Coder-Base	66.5 (55.5)	75.4 (55.6)	71.0 (55.6)
DeepSeek-Coder-Instruct [14]	6.7B	DeepSeek-Coder-Base	73.8 (69.5)	72.7 (55.6)	73.3 (62.6)
MagicoderS-DS [44]	6.7B	DeepSeek-Coder-Base	76.8 (65.2)	75.7 (56.1)	76.3 (60.7)
AlchemistCoder-L (ours)	7B	Llama 2	56.7 (52.4)	54.5 (49.6)	55.6 (51.0)
AlchemistCoder-CL (ours)	7B	CodeLlama-Python	74.4 (68.3)	68.5 (55.1)	71.5 (61.7)
AlchemistCoder-DS (ours)	6.7B	DeepSeek-Coder-Base	**79.9 (75.6)**	**77.0 (60.2)**	**78.5 (67.9)**

表 1　HumanEval 和 MBPP 上的评估结果

Model	Python	C++	Go	Java	JS	Avg
Llama 2	14.0	11.0	6.1	11.0	14.0	11.2
CodeLlama	31.7	27.4	12.8	25.6	32.9	26.1
AlchemistCoder-L	**56.7**	**31.1**	**25.6**	**45.1**	**41.5**	**37.1**
CodeLlama-Python	37.8	33.5	30.5	39.6	35.4	35.4
MagicoderS-CL	68.3	47.6	39.6	34.8	**57.9**	49.6
AlchemistCoder-CL	**74.4**	**53.1**	**42.7**	**64.0**	52.4	**57.3**
DeepSeek-Coder-Base	47.6	45.1	38.4	56.1	43.9	46.2
MagicoderS-DS	72.6	**63.4**	51.8	70.7	66.5	65.0
AlchemistCoder-DS	**79.9**	62.2	**59.8**	**72.0**	**68.9**	**68.6**

Model	pd	np	tf	sp	skl	torch	plt	All
Llama 2	2.4	7.3	6.7	6.6	2.6	1.5	7.7	5.0
CodeLlama	12.0	**27.7**	17.8	**13.2**	12.2	**20.6**	15.5	17.0
AlchemistCoder-L	**13.4**	22.7	**31.1**	11.3	**25.2**	8.8	**29.0**	**20.2**
CodeLlama-Python	16.2	16.4	15.6	17.9	20.0	22.1	38.7	21.0
MagicoderS-CL	25.1	40.9	35.6	29.3	36.5	38.2	51.0	36.7
AlchemistCoder-CL	**30.9**	**43.6**	**46.7**	**30.2**	**37.4**	**41.2**	**52.3**	**40.3**
DeepSeek-Coder-Base	21.3	35.0	26.7	23.6	34.8	25.0	34.8	28.7
MagicoderS-DS	30.6	46.8	44.2	30.2	33.0	29.7	45.2	37.1
AlchemistCoder-DS	**32.0**	**51.7**	**44.5**	**33.1**	**38.4**	**33.8**	**49.8**	**40.5**

表 2　HumanEval-X 上的评估结果

最终，通过整合高质量的开源数据、指令演进数据，并策略性地融入由 AlchemistPrompts 协调后的数据以及代码理解任务数据，构建了最终的 AlchemistCoder 微调数据集。

如表 1 所示，在主流的 Python 代码生成基准测试 HumanEval 和 MBPP 上，AlchemistCoder 系列模型在其同等规模（6.7B/7B）中取得了全面的领先地位。例如，AlchemistCoder-DS（6.7B）的性能不仅远超其他同尺寸模型，甚至能够媲美或超越参数量更大的模型（如 15B/33B/70B），显著缩小了与 GPT-3.5-Turbo 等闭源模型的差距。同时，如表 2 所示，模型在多语言代码生成（HumanEval-X）和数据科学编程（DS-1000）等任务上也表现出强大的泛化能力。

综上所述，AlchemistPrompts 能够有效降低指令与响应之间的语义偏差，提升了数据质量。另外，该方法不仅提升了模型的代码能力，还在通用语言理解（MMLU）、综合推理（BBH）和数学能力（GSM8K）等非代码任务上取得了显著进步。这说明通过协调化的多源数据微调，能够有效缓解领域微调中常见的“灾难性遗忘（catastrophic forgetting）”问题，从而培养出能力更全面的通用模型。

2.2 视觉数据生成

本节简要介绍 UniDD 的通用视觉数据蒸馏生成框架。如图 2 所示，主要包括通用任务知识挖掘和通用任务驱动扩散两个阶段：

通用任务知识挖掘阶段主要负责从大型真实数据集中提取用于指导合成图像生成的重要信息，包括：

（1）任务特定代理模型训练（Task-Specific Proxy Training）：训练分类器、检测器和分割器等代理模型（TSP 模型）。这些模型在真实数据集上进行训练，以捕获和存储任务特定的信息，例如用于分类的类别信息、用于目标检测的边界框位置以及用于图像分割的像素级掩码数据。UniDD 的代理模型选择非常灵活，可以根据目标数据集的性能需求进行调整，例如 ResNet-18 可用于轻量级分类，而 Faster R-CNN 可用于复杂检测任务。

$$\theta = \arg\min_{\theta} \mathcal{L}_{task}(\mathcal{F}(x), y)$$

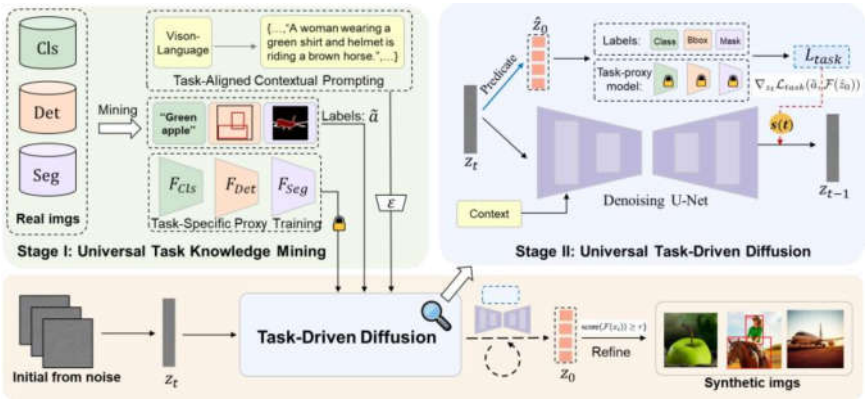

图 2　UniDD 框架图

(2) 任务对齐上下文提示 (Task-Aligned Contextual Prompting): 为了解决以往扩散模型在图像生成中忽视自然语言指导作用的问题，UniDD 引入了任务对齐上下文提示。它利用视觉-语言模型来生成与任务相关的自然语言描述。这些提示不仅描述图像中的主要对象，还包含对象之间的关系以及对象与背景的交互，为生成过程提供更丰富的上下文指导。

通用任务驱动扩散利用第一阶段提取的知识借助扩散模型来指导合成视觉图像，具体过程如下：

(1) 任务驱动扩散图像合成: 在扩散过程的每个采样步骤中，利用训练好的 TSP 模型对预测的干净图像计算任务损失，并根据这个损失的梯度来调整噪声计算任务损失，并根据这个损失的梯度来调整噪声预测。这种调整机制有效地引导扩散模型生成与给定标签(包括类别标签、边界框和像素级分割掩码)高度一致的数据。与以往需要对扩散模型进行微调的方法不同，UniDD 通过 TSP 模型引导生成过程，从而避免了昂贵的扩散模型微调，显著降低了部署成本。

(2) 代理驱动高真实性细化 (Proxy-Driven High-Realism Refinement): 为了确保生成图像的高真实性和准确性，UniDD 采用训练好的 TSP 模型对每张生成的图像进行评分。通过设置明确的阈值，可以过滤掉低质量的图像。剩余的图像会使用相同的 TSP 模型进行重新标注。这种过滤和重新标注的组合过程确保了生成的图像及其标签都具有高质量，并与所需的数据分布紧密匹配。

Dataset	IPC	Method	Accuracy(%)
ImageNet-1K	1000	Full dataset	69.8
	10	TESLA	7.7
		SRe^{2}L	21.3
		D^{4}M	27.9
		RDED	42.0
		MiMxDiff	44.3
		UniDD (Ours)	**50.4 (↑6.1)**
	50	SRe^{2}L	46.8
		D^{4}M	55.2
		RDED	56.5
		MiMxDiff	58.6
		UniDD (Ours)	**62.8 (↑4.2)**

表 3 ImageNet-1K 上的评估结果

Methods	Object Detection			
	Pascal VOC		MS COCO	
	mAP	AP50	mAP	AP50
Ratio	0.5%		0.25%	
Random	0.8±0.2	3.1±0.4	0.5±0.1	1.7±0.3
Uniform	0.9±0.1	3.4±0.3	0.8±0.2	2.4±0.5
K-Center	0.5±0.1	2.1±0.3	0.4±0.1	1.5±0.2
Herding	0.6±0.2	2.4±0.2	0.5±0.1	1.8±0.4
UniDD (Ours)	**8.5±0.4**	**22.3±0.6**	**4.5±0.3**	**10.3±0.4**
Ratio	1%		0.5%	
Random	4.2±0.5	13.7±0.6	3.7±0.2	10.1±0.3
Uniform	5.7±0.2	17.7±0.4	3.4±0.4	9.5±0.6
K-Center	3.6±0.6	12.3±0.3	3.2±0.5	9.3±0.5
Herding	3.5±0.5	11.9±0.5	3.5±0.3	9.7±0.3
UniDD (Ours)	**16.8±0.5**	**38.9±0.7**	**7.1±0.4**	**16.9±0.3**
Ratio	2%		1%	
Random	12.4±0.4	34.3±0.5	7.2±0.8	17.3±0.9
Uniform	13.8±0.3	36.2±0.4	7.4±0.5	17.6±0.5
K-Center	10.9±0.6	29.3±0.6	6.1±0.3	15.4±0.6
Herding	10.4±0.4	28.7±0.7	6.7±0.4	16.3±0.7
UniDD (Ours)	**23.9±0.5**	**48.5±0.6**	**10.8±0.4**	**22.5±0.5**
Full	51.4±0.8	80.3±0.4	32.6±0.7	51.4±0.8

表 4 Pascal VOC 和 MS COCO 上的评估结果

如表 3、表 4 所示，我们在 ImageNet-1K、Pascal VOC 和 MS COCO 等多个基准数据集上进行了广泛实验，结果表明 UniDD 超越了现有最先进的方法。特别是在 ImageNet-1K 数据集上，当每类图像数 (IPC) 为 10 时，UniDD 相较于之前的基于扩散的方法，性能提升了 6.1%，同时显著降低了部署成本。这一成果为视觉数据生成在更多样化任务中的应用提供了新的思路和方法支持。生成数据的可视化如图 3 所示。

三、图像/视频生成

3.1 风格迁移

图像风格迁移的目标是将一幅参考图像的风格应用到另一幅内容图像上，广泛应用于艺术创作、相机滤镜等场景。然而，现有的风格迁移方法往往依赖于测试时的风格调优，这不仅增加了计算和存储的开销，还可能导致模型过拟合于单一的参考图像。为了解决这些问题，

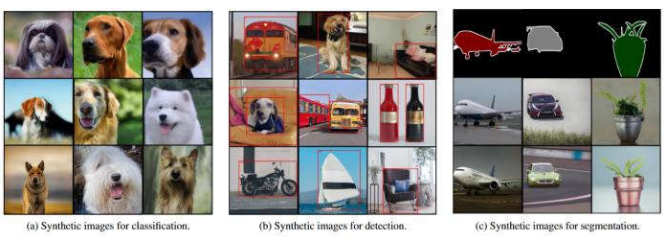

图 3 生成视觉数据的可视化

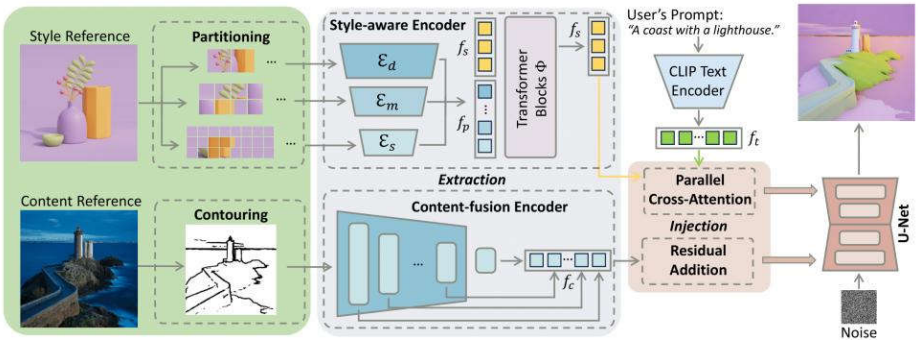

图 4　StyleShot 框架图

本节简要介绍 StyleShot，一种通用风格迁移方法。如图 4 所示，主要包括风格感知编码器、内容融合编码器以及风格数据集三个模块：

一个良好的风格表示对于无测试时调优的情况下的通用风格迁移是非常重要的。由此 StyleShot 通过构建一个风格感知编码器来提取丰富的风格嵌入。与传统的 CLIP 编码器不同，StyleShot 采用了一种多尺度补丁提取策略，可以从不同大小的图像补丁中获取低级和高级风格特征，并且通过 partition 的方式打散内容信息。在获取多尺度补丁后，StyleShot 利用不同深度的网络提取不同尺度的 style 信息，并且使用 transformer 模块进行融合学习得到最终的风格特征。通过这种方式，风格感知编码器能够捕捉到更细腻的风格细节。

为了进一步整合内容信息，实现图像驱动的风格迁移，StyleShot 训练了一个内容融合编码器。首先使用 Contour 的处理方式在保留内容的同时去除内容图像上的风格信息，然后构造一个 ControlNet[5]-like 结构的网络提取空间信息。

最后，StyleShot 构造了一个大规模的风格数据集以提升模型学习富有表现力和广义的风格表征的能力。该数据集一共包含了几万种风格、几百万张风格图片以及对应的福根文本描述。

Human	StyleCrafter	DEADifff	StyleDrop	InST	StyleAligned	StyleShot
text ↑	9.7%	19.3%	6.0%	12.7%	8.0%	**44.3%**
image ↑	14.3%	8.0%	4.0%	6.3%	17.3%	**50.0%**
CLIP	**StyleCrafter**	**DEADifff**	**StyleDrop**	**InST**	**StyleAligned**	**StyleShot**
text ↑	0.202	**0.232**	0.220	0.204	0.213	0.219
image ↑	**0.706**	0.597	0.621	0.623	0.680	0.640

表 5　StyleShot 中的评估结果

如表 5 所示，实验结果表明，StyleShot 能够有效捕捉各种风格特征，从颜色和纹理等基本元素到布局、结构和阴影等复杂元素，最终生成与文本提示一致的风格化图像。这展现了我们风格感知编码器在提取丰富且富有表现力的风格嵌入方面的有效性。

另外，得益于我们的内容融合编码器，StyleShot 还擅长将风格迁移到内容图像上。如图 5、6 所示，我们的 StyleShot 可以将任何风格（甚至包括光影、点画

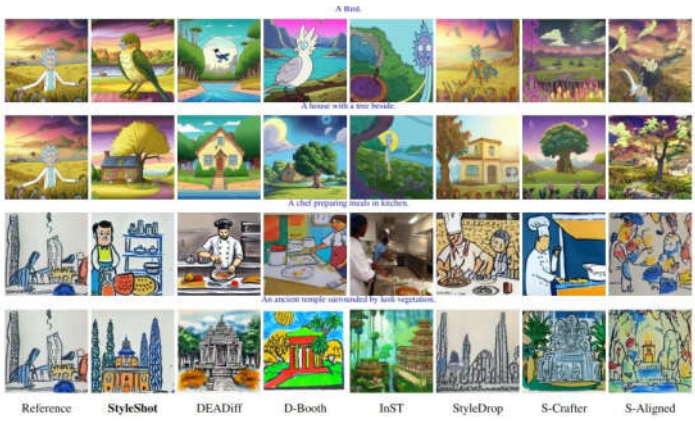

图 5　文本驱动风格迁移可视化结果

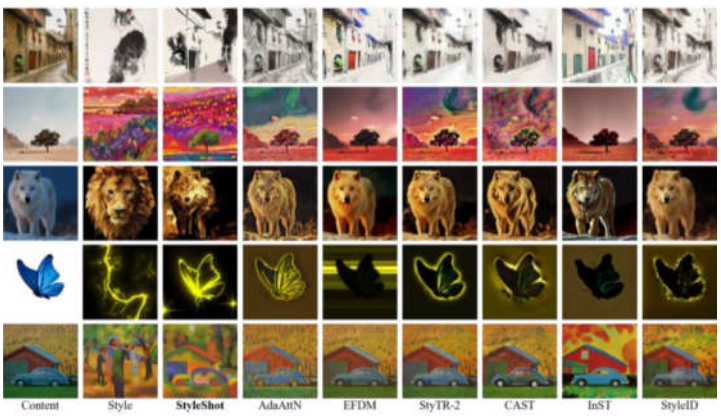

图 6　图像驱动风格迁移可视化结果

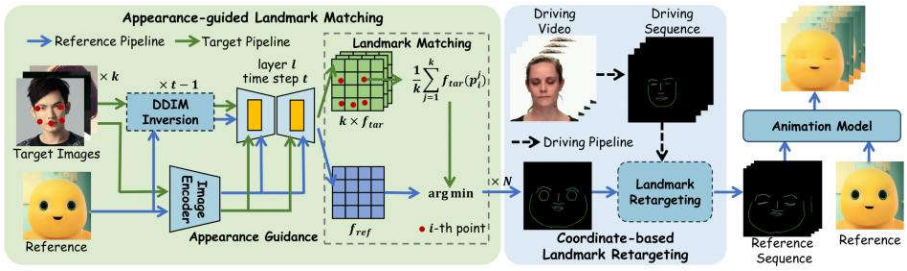

图 7　FaceShot 框架图

法、低多边形和平面等复杂高级风格）迁移到各种内容图像（例如人物、动物和场景）上，而基线方法主要擅长绘画风格，而难以处理这些高级风格。这证明了内容融合编码器在实现卓越风格迁移性能的同时，还能保持内容图像的结构完整性。

3.2 肖像动画

当前的肖像动画方法大致由面部关键点序列控制驱动。这类方法已经表现出了良好的动作可控性与面部保真能力，特别是在人脸驱动任务中能够生成自然、稳定的动画效果，然而由于其核心依赖的人脸关键点检测以及动作序列生成模块大多是在真实高质量面部数据集上进行监督训练，在处理非人类角色（如表情包、玩具、卡通角色）时，往往无法准确识别关键点分布和迁移面部动作，导致动画生成阶段出现结构错位、嘴型崩塌等现象。为了解决这些问题，本节简要介绍 FaceShot，一个无需训练的肖像动画生成框架。如图 7 所示，主要包括语义引导关键点匹配、坐标系建模动作变换以及肖像动画三个模块：

基于扩散模型的强大泛化能力，DIFT[6]发现不同图像的语义相近区域的预训练扩散模型特征是相似的，可以直接进行关键点匹配。然而人脸和非人类角色之间存在较大的域差异，常常会造成错误匹配。为了拉近不同域在特征空间的差异，FaceShot 结合 IP-Adapter[7]，将参考图像作为外观引导注入扩散过程。另外 FaceShot 还构建了包含眼睛、嘴巴、眉毛等五个部分的外观图库，自动选取相近域作为辅助目标，进一步缓解非人类角色与人类语义空间之间的域间差异。实现对非人类角色的面部关键点的精准定位。

为了更精准地捕捉驱动视频中的面部动作，FaceShot 构建了全局与局部坐标系，用于显式建模并迁移整体与局部表情变化。具体而言，FaceShot 利用参考图中面部轮廓两端点定义全局坐标系，通过原点位置与坐标轴角度的变化建模头部的整体位移与旋转；同时，在眼、眉、嘴、鼻等局部区域分别建立子坐标系，通过点在各自坐标系中的相对变化，刻画细节动作的动态变形。

最后将对应肖像的关键点序列输入到任意的关键点驱动的肖像动画模型中，就可以得到最终的动画结果。图 8 和表 6 的实验结果显示，FaceShot 在非人类角色上表现出色。相比现有方法，FaceShot 在身份保持

Methods	Metrics				User Preference		
	ArcFace ↑	HyperIQA ↑	Aesthetic ↑	Point-Tracking ↓	Motion ↑	Identity ↑	Overall ↑
FaceVid2Vid	0.525	33.721	4.267	6.944	3.58	3.83	4.52
FADM	0.633	39.402	4.522	6.993	1.93	2.04	1.96
X-Portrait	0.490	52.357	4.754	7.301	1.47	1.63	1.57
Follow Your Emoji	0.612	52.056	4.906	6.960	6.91	6.67	6.74
AniPortrait*	0.634	55.951	4.928	6.367	5.84	5.64	5.39
MegActor*	0.613	40.191	4.855	7.183	6.53	6.75	6.26
LivePortrait*	0.893	53.587	5.092	7.474	7.33	7.08	7.11
MOFA-Video	0.695	52.272	4.952	14.985	3.27	3.04	3.18
FaceShot	**0.848**	**53.723**	**5.036**	**6.935**	**8.14**	**8.32**	**8.27**

表 6　FaceShot 中的评估结果

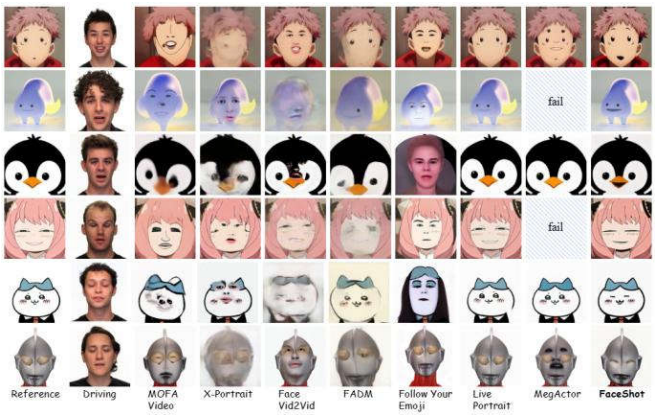

图 8　FaceShot 中的可视化结果

图 9　将 FaceShot 作为插件

图 10　非人类驱动视频可视化结果

（ArcFace）、图像质量（HyperIQA）和动作还原（Point Tracking）等多个指标上均取得领先，尤其在结构不规则、风格差异大的角色（如玩偶、卡通形象、动物）上表现更为稳定。现有方法常常因关键点检测不准或驱动迁移失真而导致动画崩坏、嘴型错位等问题，而 FaceShot 利用语义引导的关键点匹配与坐标建模动作迁移，显著提升了角色动作的还原度和连贯性。

除了作为独立的肖像动画生成框架，FaceShot 还展现出出色的模块化扩展能力。如图 9 所示，在插件化应用方面，FaceShot 可作为关键点序列预测模块集成到现有的关键点驱动方法中（如 MOFA-Video 和 AniPortrait），显著提升其在非人类角色上的动画稳定性与结构一致性。

此外，FaceShot 还支持从非人类驱动视频中提取动作信号，并将其迁移到任意参考角色，实现跨类别、跨风格的开放域角色动画。如图 10 所示，这一能力打破了传统肖像动画对人类驱动数据的依赖，展示了 FaceShot 向通用、开放场景扩展的广阔潜力。

四、总结与未来展望

本文系统梳理了同济大学VILL实验室在多模态生成式人工智能领域的一系列研究成果，涵盖语言数据生成、视觉数据生成、图像风格迁移与肖像动画等多个子方向。实验室针对不同模态下生成技术的关键挑战，提出了具有创新性的技术方案，并在多个顶级会议如NIPS、CVPR、ICLR中发表相关成果，部分工作更是获得了广泛的开源社区认可，展现了团队在该领域的前瞻性研究能力与工程实现水平。

在语言数据生成方面，AlchemistCoder通过构建协调化多源微调框架，有效缓解了Code LLMs在多源数据整合中的冲突问题，并进一步提升了模型的泛化能力和理解能力，不仅在HumanEval和MBPP等标准评测上取得领先，还在跨语言、多任务场景中展现了出色性能。该工作展示了通过精细化提示与任务设计进行数据重构的巨大潜力。

在视觉数据生成方面，UniDD以任务知识挖掘和任务驱动扩散为核心，构建了一种通用的视觉数据蒸馏框架，实现了高质量、高效率的数据合成，显著降低了模型训练对真实数据的依赖，为低资源条件下的视觉模型训练开辟了新路径。尤其是在图像分类、检测和分割等多任务场景中展现了良好适应性，验证了其在实际应用中的可行性和可扩展性。

在图像与视频生成方向，StyleShot和FaceShot分别从风格迁移与动画生成出发，打破了传统方法对测试时调优和人脸驱动数据的依赖。StyleShot通过构建风格感知与内容融合编码器，成功实现了任意风格与任意内容之间的高质量迁移；而FaceShot则通过语义引导的关键点匹配与精细的动作建模，实现了对非人类角色的稳定驱动，并具备良好的插件化集成能力，支持开放域角色间的动作迁移。

展望未来，生成式人工智能将在更多实际应用中发挥更广泛作用。VILL实验室将继续围绕“多模态、高质量、开放域”的核心目标推进研究工作。一方面，在语言生成方面将进一步探索指令构造、模型行为监督与对齐等关键问题，提升模型对人类意图的理解与响应能力；另一方面，在视觉生成方面，将聚焦于大规模多模态数据的结构化建模与跨域泛化能力的提升，拓展生成模型在医疗、工业、娱乐等真实场景中的落地空间。同时，结合扩散模型与结构建模的前沿方法，探索更多有效的通用生成范式，为生成式AI系统的稳定性、可控性与通用性奠定更坚实的基础。

责任编委　王金甲

参考文献

[1] Zifan Song, Yudong Wang, Wenwei Zhang, Kuikun Liu, Chengqi Lyu, Demin Song, Qipeng Guo, Hang Yan, Dahua Lin, Kai Chen, Cairong Zhao. "AlchemistCoder: Harmonizing and Eliciting Code Capability by Hindsight Tuning on Multi-source Data. " Advances in Neural Information Processing Systems(NIPS), 2025, 2185—2214.

[2] Ding Qi, Jian Li, Junyao Gao, Shuguang Dou, Ying Tai, Jianlong Hu, Bo Zhao, Yabiao Wang, Chengjie Wang, Cairong Zhao. "Towards Universal Dataset Distillation via Task-Driven Diffusion." Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10557-10566

[3] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, Cairong Zhao. "Styleshot: A snapshot on any style." TPAMI:2025.

[4] Junyao Gao, Yanan Sun, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, Cairong Zhao. "FaceShot: Bring Any Character into Life." The Thirteenth International Conference on Learning Representations (ICLR).

[5] Lvmin Zhang, Anyi Rao, Maneesh Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836-3847

[6] Luming Tang and Menglin Jia and Qianqian Wang and Cheng Perng Phoo and Bharath Hariharan. "Emergent Correspondence from Image Diffusion". Thirty-seventh Conference on Neural Information Processing Systems (NIPS). 1363—1389.

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, Wei Yang. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. " arXiv preprint arxiv:2308.06721.

高俊尧

同济大学计算机科学与技术学院 2022 级博士研究生，导师为赵才荣教授。在 ICLR, ICML, NIPS, ICCV, CVPR, AAAI, TPAMI, IJCV 等国际期刊会议发表论文 10 余篇，主要研究方向为图像/视频生成、行人再识别、隐私安全。

Email: junyaogao@tongji.edu.cn

宋子帆

同济大学博士生，师从赵才荣教授，本科毕业于同济大学。在 NeurIPS、ICML、AAAI 等国际期刊会议发表一作论文 4 篇，主要研究方向为多模态学习、Data-centric AI、大模型微调。

Email: 2111139@tongji.edu.cn

齐 鼎

同济大学博士生，师从赵才荣教授。在 NeurIPS、CVPR 国际会议发表一作论文 2 篇，主要研究方向为数据集蒸馏、Data-centric AI。

Email: 2011267@tongji.edu.cn

赵才荣

工学博士，同济大学计算机科学与技术学院教授，博士生导师，计算机智能教研室主任。曾任香港理工大学兼职研究员（2016-2017）。目前担任上海市计算机学会计算机视觉专委会主任，中国图象图形学学会青工委秘书长，中国人工智能学会粒计算与知识发现专委常委，中国计算机学会杰出会员，担任 IEEE TMM Guest Editor、《中国图象图形学报》、《计算机科学》青年编委。已在 TPAMI、IJCV、《中国科学.信息科学》、CVPR、ICML、NIPS 等发表学术论文 50 余篇，研究成果获 2022 年上海市科技进步一等奖（序 4/13）以及 2023 年上海市自然科学二等奖（序 1/4）。

Email: zhaocairong@tongji.edu.cn