面向高质量视觉生成的数据、方法与应用

https://github.com/NJU-PCALab

分享：邰颖

时间：2024.12.20

单位：南京大学

团队简介：NJU-PCALab

郜颖

- 南京大学 智能科学与技术学院（副教授、博导）【2023.08～】
- 原腾讯专家研究员（T12），研究组长【2017.04～2023.07】

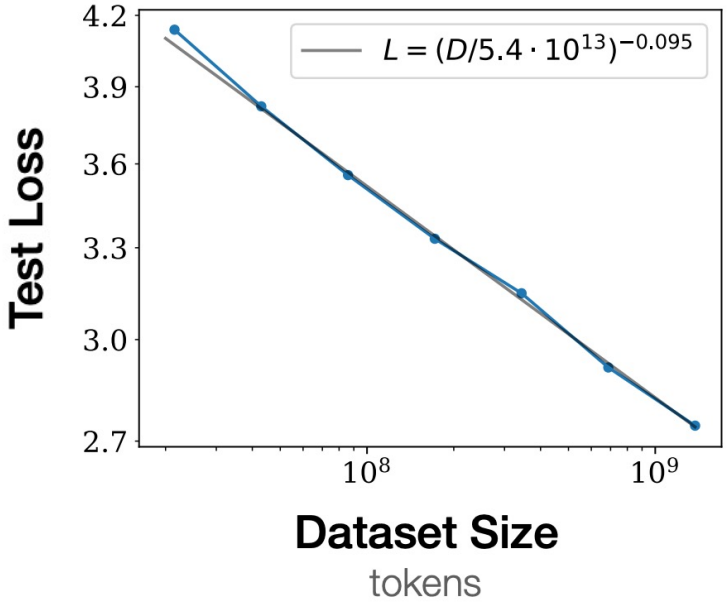
- 姑苏创新创业领军人才计划（2024）
- 南京大学-华为“紫金学者”人才基金（2023）
- 谷歌学术引用12,800余次，两篇代表性工作（独立一作）分别被引用2,600余次和1,900余次

目前指导博士生5人、硕士生7人、科研助理2人

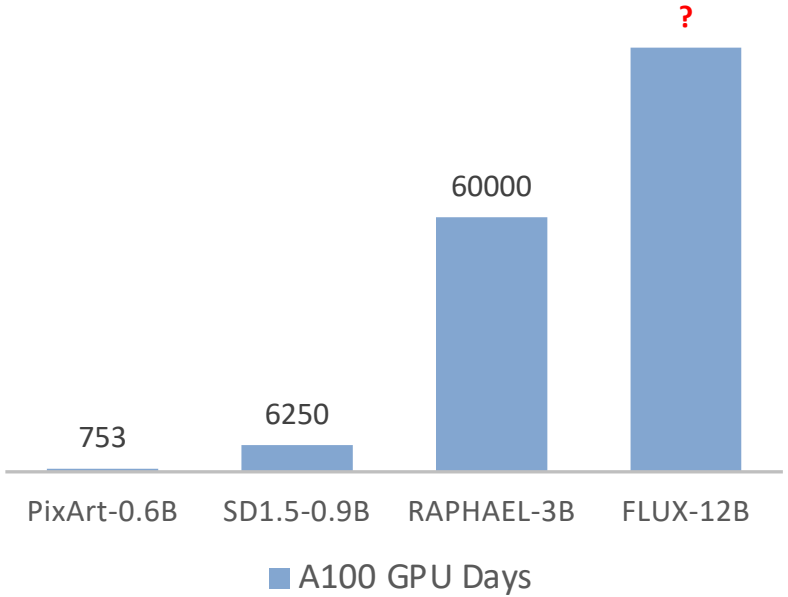
南京大学 NANJING UNIVERSITY

当前研究背景

- 大模型时代，数据极为关键。而开源、高质量、大规模的视频数据是稀缺的
- 主流高校研究机构，预训练大模型的成本过高，低/零成本训练方法是亟需的



OpenAI, Scaling Laws for Neural Language Models, 2020



Huawei, PIXART-α: FAST TRAINING OF DIFFUSION TRANSFORMER FOR PHOTOREALISTIC TEXT-TO-IMAGE SYNTHESIS, 2024

分享内容简介

- 面向高质量视频生成的数据集**OpenVid-1M**

- 一个高质量的文生视频数据集，**最高支持1080p视频生成**
- 提出了一种**多模态视频DiT模型结构(MVDiT)**
- 在视频生成、视频复原、视频插帧、3D/4D生成等任务中被使用

- 面向高质量图像分区生成的方法**RAG-Diffusion**

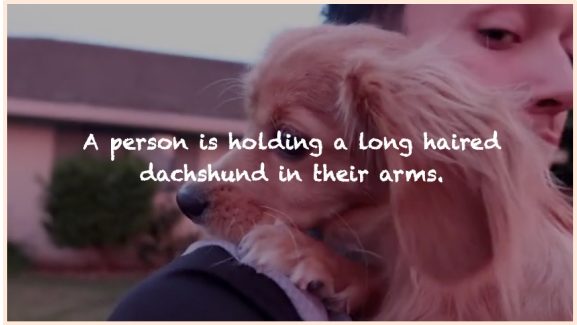
- 无需微调训练：**无缝兼容现有DiT框架**（如Flux, SD3）
- 精确区域控制：相比以往Flux-1.dev实现**准确生成复杂布局**
- 重绘功能：实现修改特定区域而不影响其他部分

OpenVid-1M：与以往流行视频数据集的差异

- 面向广大研究机构可用的、高质量、开放场景、百万量级的文本-视频数据库

WebVid – 10M

分辨率低、有水印



Panda – 70M

短文本、存在静止、运动剧烈、清晰度一般视频

CelebvHQ、UCF101等

受限领域、数据量小、分辨率不够高

OpenVid-1M is a high-quality text-to-video dataset designed for research institutions to enhance video quality, featuring high aesthetics, clarity, and resolution. It can be used for direct training or as a quality tuning complement to other video datasets.

OpenVid-1M：更清晰的视频、更丰富的描述

UCF-101: Biking.

WebVid-10M: Apples in the garden.

Panda-70M: A woman is cutting vegetables on a wooden table.

Existing text-to-video datasets

The video captures a man driving a convertible sports car on a sunny day. He is wearing a red plaid shirt, glasses, and a watch. The car's top is down, and the man is holding the steering wheel with both hands. The car is moving on a road with a clear sky and green hills in the background. The man appears to be enjoying the drive, and the overall atmosphere of the video is cheerful and relaxed.

OpenVid-1M (Ours)

与以往数据库的定量统计分析：高质量的视频、丰富的描述

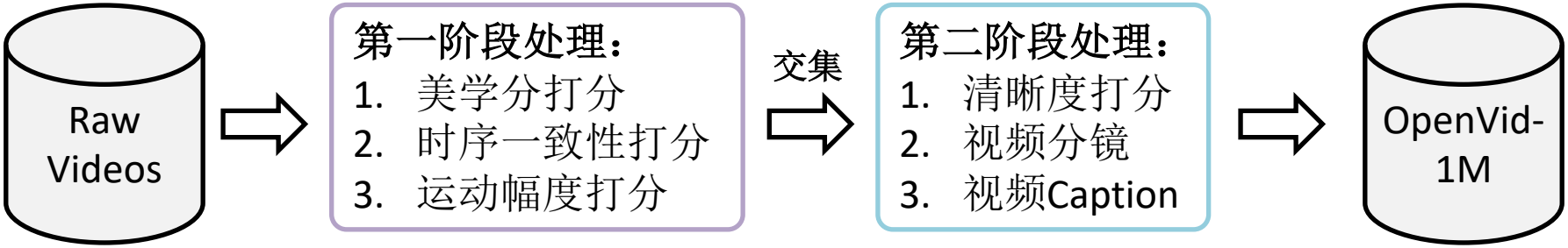
- 百万量级的**OpenVid-1M; 400K 1080P**分辨率的**OpenVidHD**

Table 2: Comparisons with previous text-to-video datasets. Our `OpenVid-1M` is a *million-level, high-quality and open-scenario* video dataset for training high-fidelity text-to-video models.

Dataset	Scenario	Video clips	Average length (seconds)	Duration (hours)	Resolution	Caption
UCF101	Action	13K	7.2	2.7	320×240	N/A
Taichi-HD	Human	3K	-	-	256×256	N/A
SkyTimelapse	Sky	35K	-	-	640×360	N/A
FaceForensics++	Face	1K	-	-	Diverse	N/A
WebVid	Open	10M	18.7	52k	596×336	Short
ChronoMagic	Metamorphic	2K	11.4	7	Diverse	Long
CelebvHQ	Portrait	35K	6.6	65	512×512	N/A
OpenSoraPlan-V1.0	Open	400K	24.5	274	512×512	Long
Panda	Open	70M	8.5	166k	Diverse	Short
`OpenVid-1M` (Ours)	Open	1M	7.2	2.1k	Diverse	Long
`OpenVidHD-0.4M` (Ours)	Open	433K	9.6	1.2k	1920×1080	Long

南京大学 NANJING UNIVERSITY

OpenVid数据收集流程：22,184 A100 GPU hours



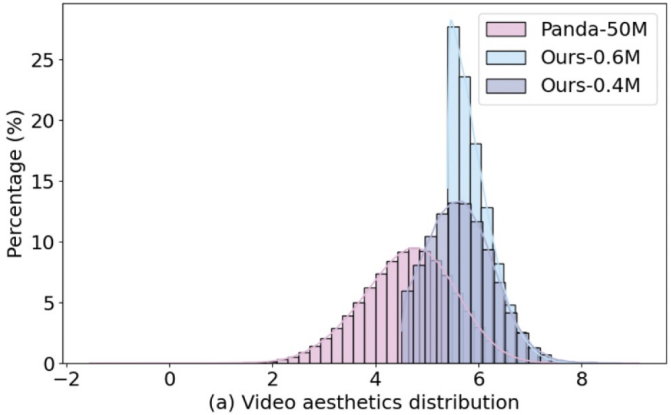
Pipeline	Tool	Computation Resources	Processing Time (hours)	Remark
Aesthetics score	LAION Aesthetics Predictor	32 A100	320	Get high aesthetics score set S_A
Temporal consistency	CLIP (Radford et al., 2021)	48 A100	173	Obtain moderate consistency set S_T
Motion difference	UniMatch (Xu et al., 2023)	48 A100	59	Obtain moderate amplitude of motion Set S_M
Intersection of qualified videos	Intersection	-	-	Obtain intersection: $S_I = S_A \cap S_T \cap S_M$
Clarity assessment	DOVER-Technical (Wu et al., 2023)	8 A100	25	Obtain clear and high-quality video set S
Clip extraction	Cascaded Cut Detector (Blattmann et al., 2023)	8 A100	30	Split multi-scene videos: $\widetilde{S} = Detector(S)$
Video caption	LLaVA-v1.6-34b (Liu et al., 2023a)	8 A100	46	Obtain long captions for the videos

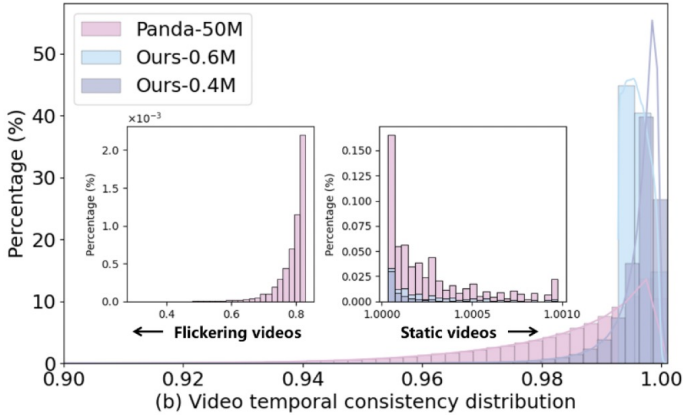
统计信息：更高的美学，剔除极端的运动/时序视频，更丰富的文本描述

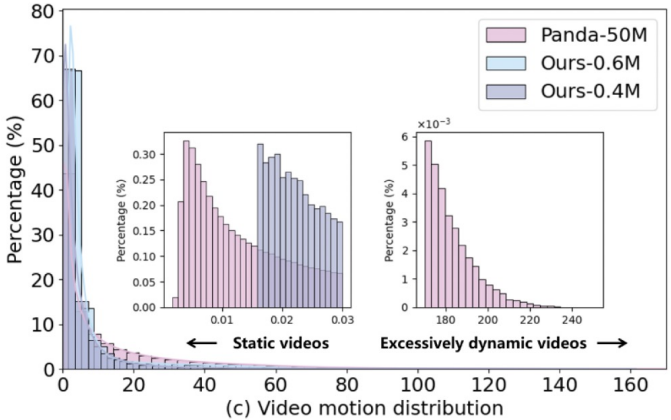
美学/画质 卓越

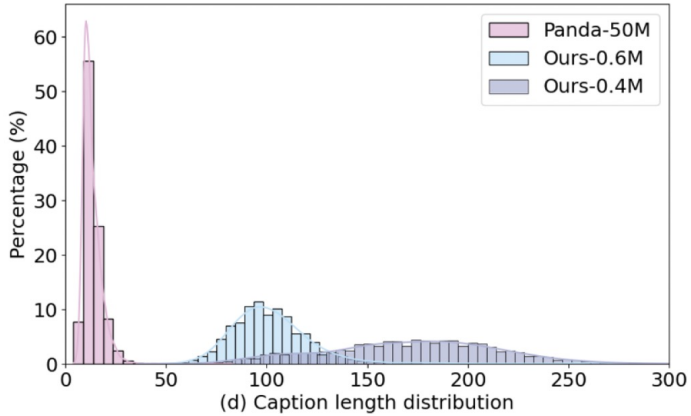
时序/运动 合理

文本描述 丰富





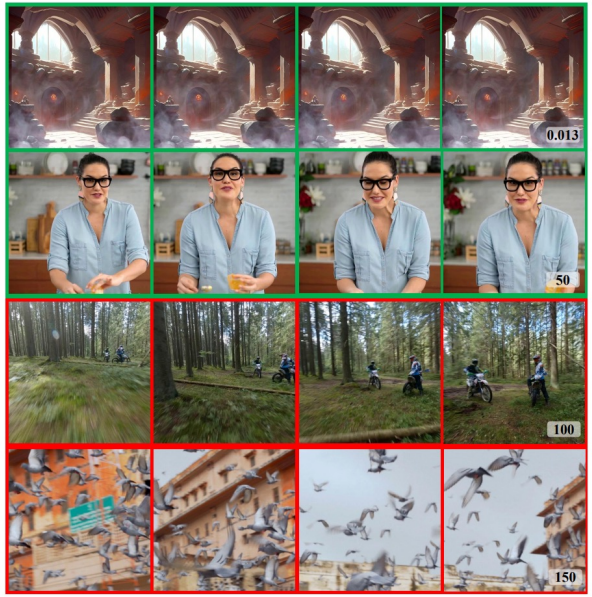




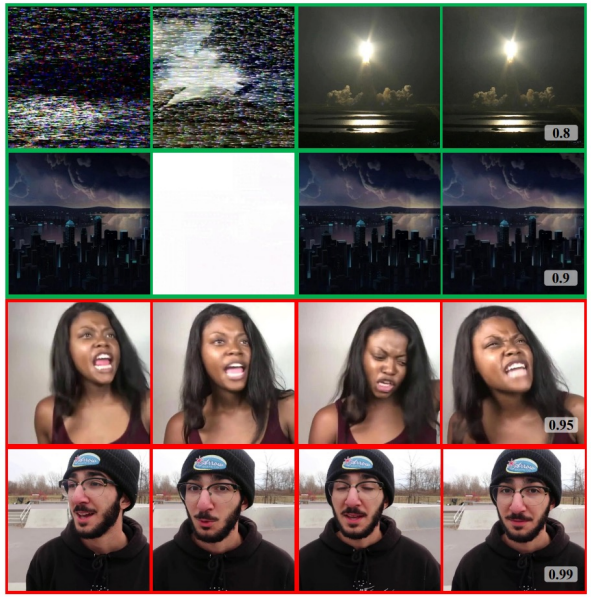
统计信息：严格的质量筛选 & 视频场景、长度分布

(a)

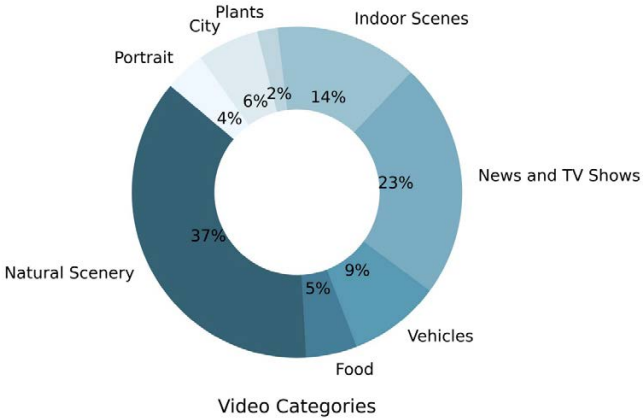
(b)



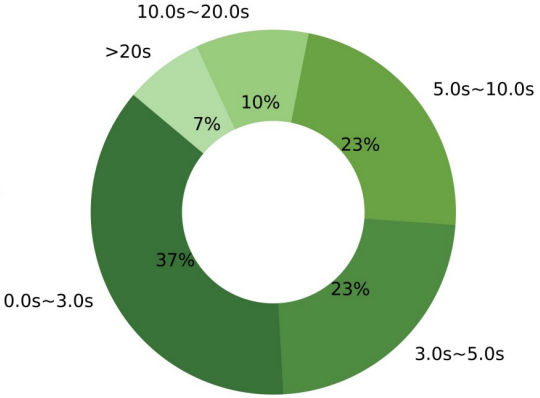
(c)



(d)



Video Categories



Video Durations

Figure 10: Visualizations of the videos with varying (a) clarity, (b) aesthetic, (c) motion, and (d) temporal consistency scores.

Multi-modal Video DiT（MVDiT）

- 多模态视频MVDiT：充分学习视频与文本语义特征

Multi-Modal Self-Attention Module:

$$\mathbf{F}^s_{\text{SAL}} = \text{SAL}(\text{Concat}(\text{AdaLN}(\mathbf{X}, \mathbf{t}_1), \text{AdaLN}(\mathbf{Y}, \mathbf{t}_1)))$$

$$\text{AdaLN}(\mathbf{X}, \mathbf{t}_1)) = \gamma^1_1 \text{LayerNorm}(\mathbf{X}) + \beta^1_1.$$

$$\mathbf{X}^s_{\text{SAL}}, \mathbf{Y}^s_{\text{SAL}} = \text{Split}(\mathbf{F}^s_{\text{SAL}})$$

$$\mathbf{X}^s = \mathbf{X} + \alpha^1_1 \mathbf{X}^s_{\text{SAL}}, \mathbf{Y}^s = \mathbf{Y} + \alpha^2_1 \mathbf{Y}^s_{\text{SAL}}.$$

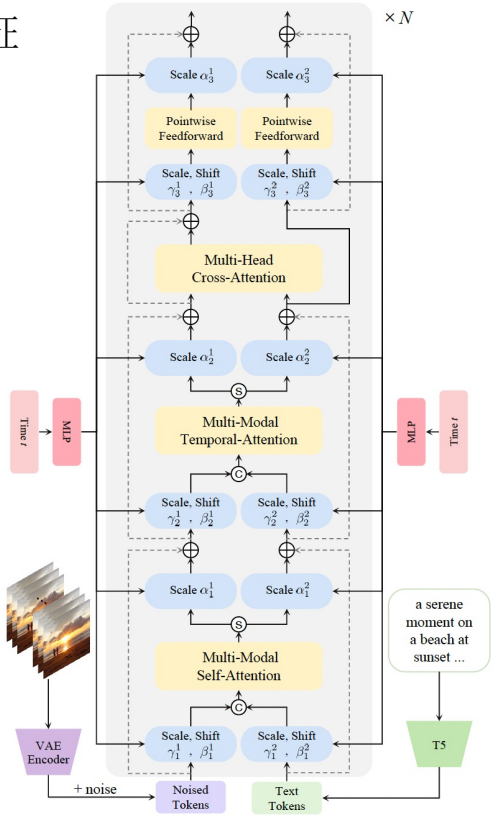
Multi-Modal Temporal-Attention Module:

$$\mathbf{X}^t_{\text{TAL}}, \mathbf{Y}^t_{\text{TAL}} = \text{Split}(\text{TAL}(\text{Concat}(\text{AdaLN}(\mathbf{X}^s, \mathbf{t}_2), \text{AdaLN}(\mathbf{Y}^s, \mathbf{t}_2))))$$

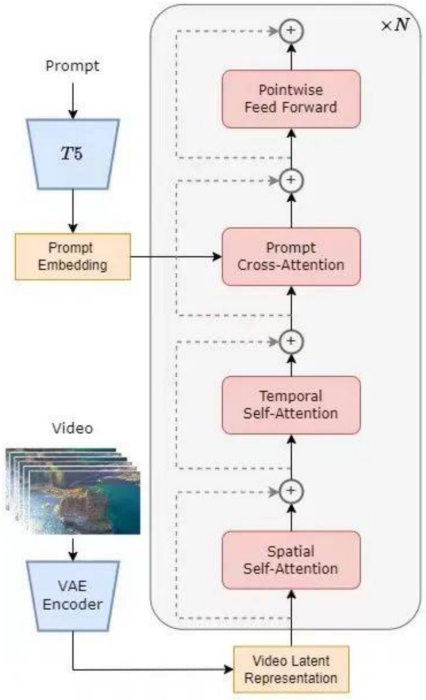
$$\mathbf{X}^t = \mathbf{X}^s + \alpha^1_2 \mathbf{X}^t_{\text{SAL}}, \mathbf{Y}^t = \mathbf{Y}^s + \alpha^2_2 \mathbf{Y}^t_{\text{SAL}}.$$

Multi-Head Cross-Attention Module:

$$\mathbf{X}^c = \text{CAL}(\mathbf{X}^t, \mathbf{Y}^t) + \mathbf{X}^t.$$



MVDiT



STDiT in OpenSora[1]

[1] https://github.com/hpcaitech/Open-Sora

效果与训练收敛分析

- **效果分析**：在开源学术界方案中，使用**最少的32卡GPUs，实现最高的分辨率1024px，和最好的清晰度**
- **收敛分析**：更多的卡对训练收敛有帮助（Middle）；50K steps左右的模型效果趋于稳定

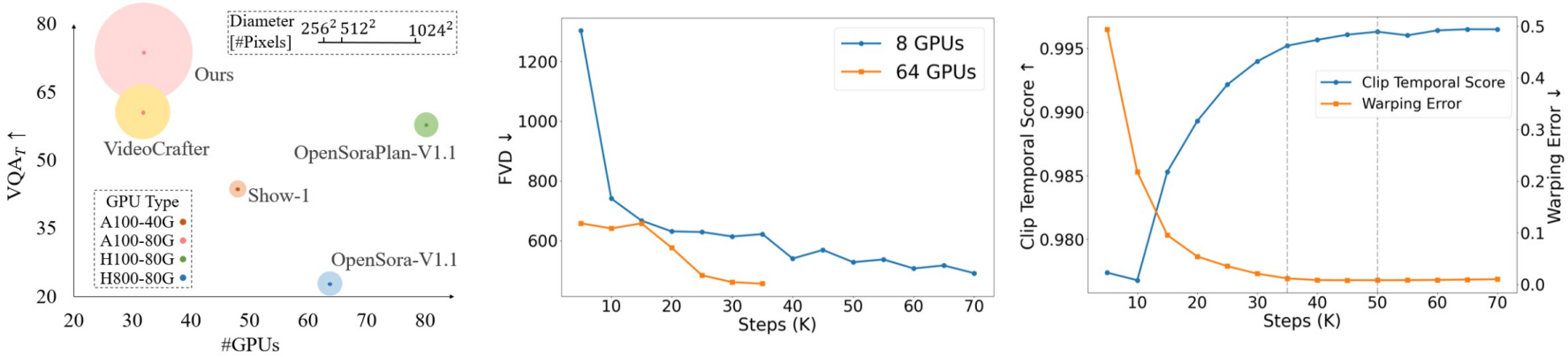


Figure 5: **Left:** Comparison with SotA T2V models on VQA_T, GPU type and resolution. The color of the middle dot in each circle indicates GPU type, and circle diameter represents video resolution. **Middle:** Curves of FVD with different number of GPUs. More GPUs accelerates the model's convergence. **Right:** Curves of clip temporal score and warping error. Our T2V model typically starts to stabilize at 35K steps and achieves temporal consistency around 50K steps.

与WebVid-10M和Panda-50M的定量指标对比

- 在大分辨率上，基于OpenVid-1M训练的模型在**美学、清晰度方面有明显优势**
- 在小分辨率上，基于OpenVid-1M训练的模型在美学方面有明显优势，文本-视频对齐指标有一定优势
- WebVid不支持1024px训练；Panda本身未筛选高质量数据；因此结合4x视频超分模型获得1024px效果
- 所有模型均**使用32卡算力，经过至少14天的充分训练**。其中openvid-1m最快收敛，panda最慢收敛

Table 4: Comparisons with previous representative text-to-video training datasets. The STDiT model used in OpenSora is adopted and kept the same for all of the cases. For fair comparison, training iterations are selected at the same step (50K) for fair comparison. All models with 256×256 resolution are adequately trained on 32 A100 GPUs for at least 14 days to reach 50K iterations.

Resolution	Training Data	VQA_A↑	VQA_T↑	Blip_bleu↑	SD_score↑	Clip_temp_score↑	Warping_error↓
256×256	WebVid-10M (Bain et al., 2021)	13.40	13.34	23.45	67.64	99.62	0.0138
256×256	Panda-50M (Chen et al., 2024c)	17.08	9.60	24.06	67.47	99.60	0.0200
256×256	`OpenVid-1M` (Ours)	17.78	12.98	**24.93**	67.77	99.75	0.0134
1024×1024	WebVid-10M (4× Super-resolution)	69.26	65.74	23.15	67.60	99.64	0.0137
1024×1024	Panda-50M (4× Super-resolution)	63.25	53.21	23.60	67.44	99.57	0.0163
1024×1024	Panda-50M-HD	13.48	42.89	21.78	**68.43**	99.84	0.0136
1024×1024	`OpenVidHD-0.4M` (Ours)	**73.46**	**68.58**	23.45	68.04	**99.87**	**0.0052**

定量指标对比：在清晰度和美学上与CogvideoX-5B可比

- 高质量的美学、清晰度、分辨率生成表现
- 不错的文本-视频对齐、时序一致表现
- 仅使用1M的训练集，1B左右的参数量

Table 3: Comparison with state-of-the-art text-to-video generation methods. The best results are marked in **bold**, while the second best ones are underscored.

Method	Resolution	Training Data	VQA_A↑	VQA_T↑	Blip_bleu↑	SD_score↑	Clip_temp_score↑	Warping_error↓
Lavie (Wang et al., 2023c)	512×320	Vimeo25M	63.77	42.59	22.38	68.18	99.57	0.0089
Show-1 (Zhang et al., 2023)	576×320	WebVid-10M	23.19	44.24	23.24	68.42	99.77	0.0067
OpenSora-V1.1	512×512	Self collected-10M	22.04	23.62	23.60	67.66	99.66	0.0170
Latte (Ma et al., 2024a)	512×512	Self collected-330K	55.46	48.93	22.39	68.06	99.59	0.0203
VideoCrafter (Chen et al., 2023a)	1024×576	WebVid-10M; Laion-600M	66.18	58.93	22.17	68.73	99.78	0.0295
Modelscope (Wang et al., 2023b)	1280×720	Self collected-Billions	40.06	32.93	22.54	67.93	99.74	0.0162
Pika	1088×612	Unknown	59.09	64.96	21.14	68.57	**99.97**	**0.0006**
OpenSoraPlan-V1.2 (Lab & etc., 2024)	640×480	Self collected-7.1M	23.25	65.86	19.93	**69.21**	**99.97**	0.001
CogVideoX-5B (Yang et al., 2024)	720×480	Self collected-35M	35.12	**76.86**	**24.21**	68.91	99.79	0.0077
Ours	1024×1024	`OpenVid-1M`	**73.46**	68.58	23.45	68.04	99.87	0.0052

Ablation: MVDiT vs. STDiT

- 在相同训练集下（如Ours-0.4M），**MVDiT在视频质量方面（美学、清晰度）有显著优势**
- 基于OpenVid-1M训练的模型，在视频质量方面，效果相比子集进一步提升
- 更高的分辨率，视频的稳定性指标更好，MVDiT相比STDiT优势明显

Table 4: Ablations on different resolutions, architectures and training data. For models trained with 256×256 resolution, training iterations are selected at the similar steps for fair comparison. 'Pre-trained Weight' means initializing with a corresponding pretrained model, , 'MVDiT-256' indicates that the MVDiT model with 256×256 resolution is used as the pretrained weight.

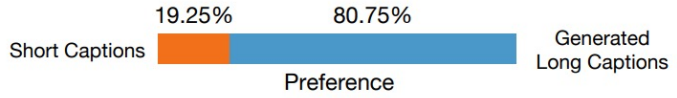
Model	Resolution	Training Data	Pretrained Weight	VQA_A↑	VQA_T↑	Blip_bleu↑	SD_score↑	Clip_temp_score↑	Warping_error↓
STDiT	256×256	Ours-0.4M	PixArt-α	11.11	12.46	**24.55**	67.96	99.81	0.0105
STDiT	512×512	Ours-0.4M	STDiT-256	65.15	59.57	23.73	68.24	99.80	0.0089
MVDiT	256×256	Ours-0.4M	PixArt-α	22.39	14.15	23.72	67.73	99.71	0.0091
MVDiT	256×256	OpenVid-1M	PixArt-α	24.87	14.57	24.01	67.64	99.75	0.0081
MVDiT	512×512	OpenVid-1M	MVDiT-256	**66.65**	**63.96**	24.14	**68.31**	**99.83**	**0.0008**

主观评价实验: 不同长度描述的对比（短：Panda；长：Ours）

- 目前的长文本内容更丰富，**在主观评价中优势显著**；但同时存在**相对更多的幻觉现象**

Panda: A man in a suit gives a speech at an event.

Ours: The video is a news segment featuring a man in a suit sitting in an office chair. The man is bald and wearing a suit with a red tie. He is looking to the side with a serious expression. The background is blurred, but it appears to be an office setting. The video has a news ticker at the bottom with the text "WHAT TO EXPECT FROM BEZOS' TESTIMONY" and "ON THE PHONE MATT DAY BLOOMBERG TECHNOLOGY". The style of the video is a standard news segment with a focus on the man in the suit. The video is likely discussing the testimony of Jeff Bezos, the CEO of Amazon.



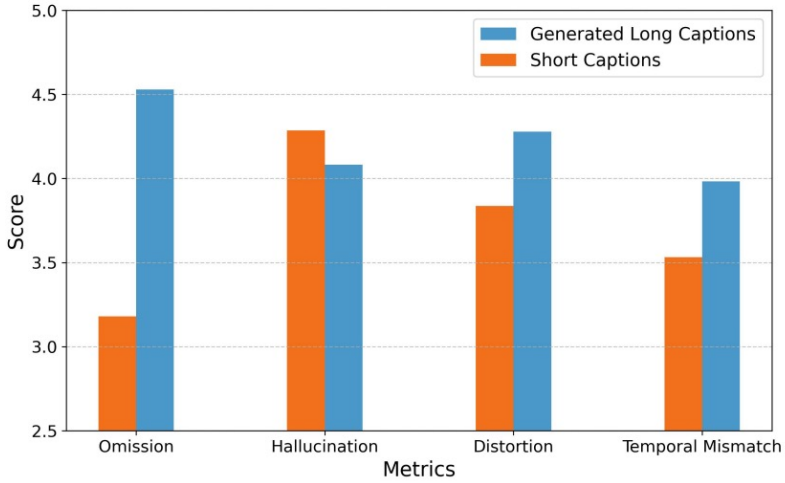
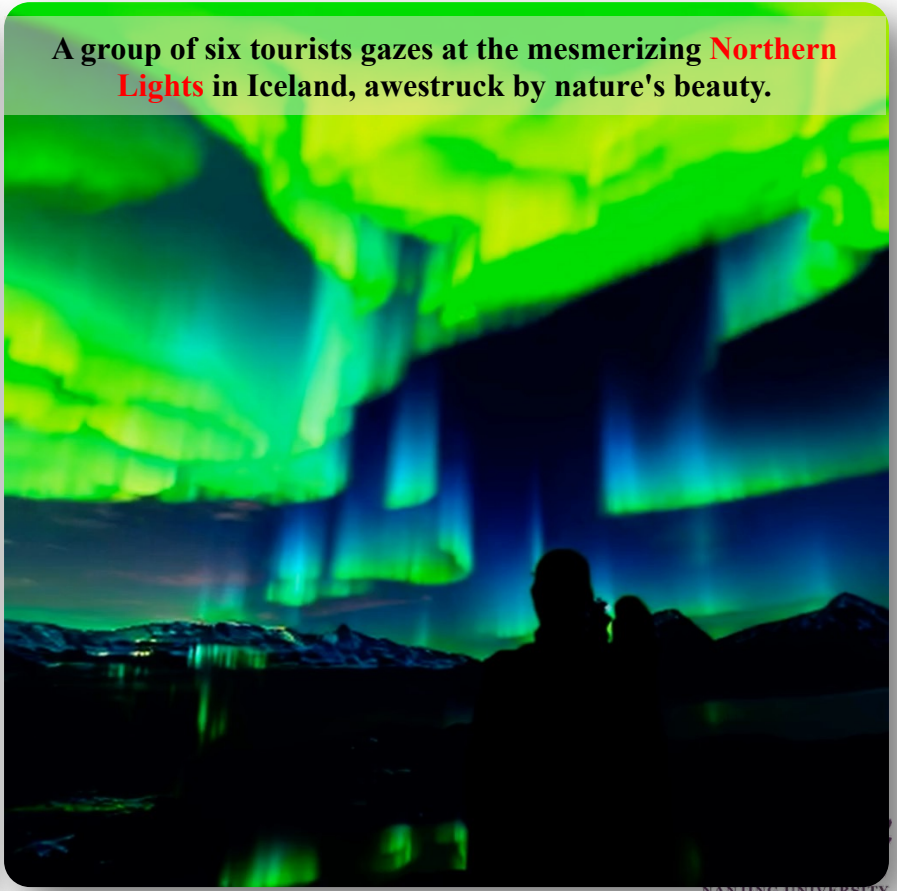


Figure 1: Human Preference on short captions vs. our generated long captions over 1,117 validation samples and 10 qualified volunteers.

核心能力：高分辨率，最高支持1080P分辨率



核心能力：高美学、高清晰度

核心能力：文本-视频对齐

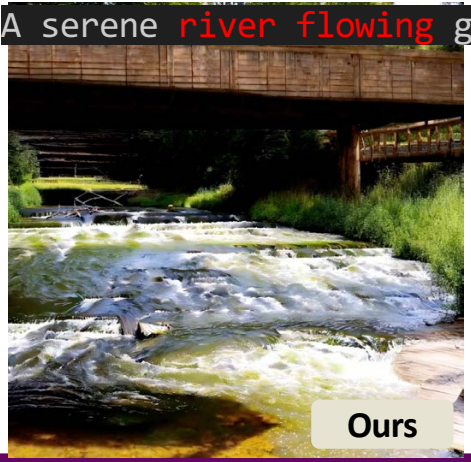

Two fish eating spaghetti on a subway

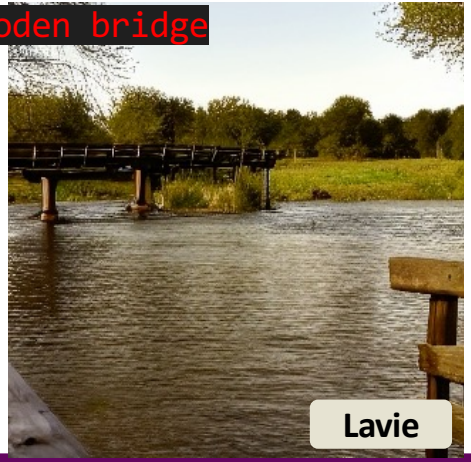

a pumpkin car with a candle in it

视觉对比 [OpenSoraPlan-V1.1, Lavie, VideoCrafter]

In goldfish glass

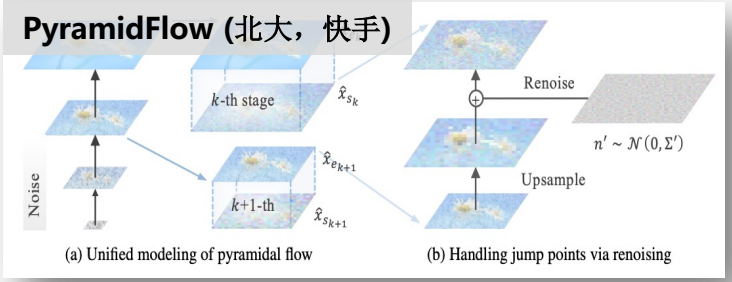
A serene river flowing gently under a rustic wooden bridge



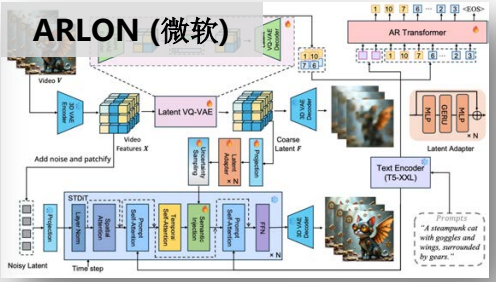


引用分析：覆盖视频生成、理解、3D/4D生成、视频复原等

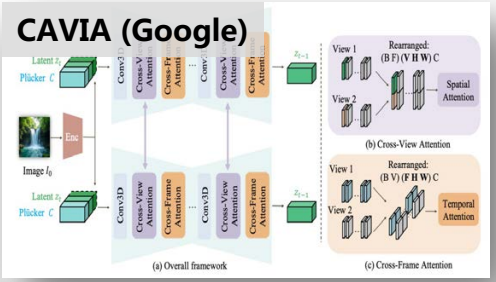
PyramidFlow (北大，快手)



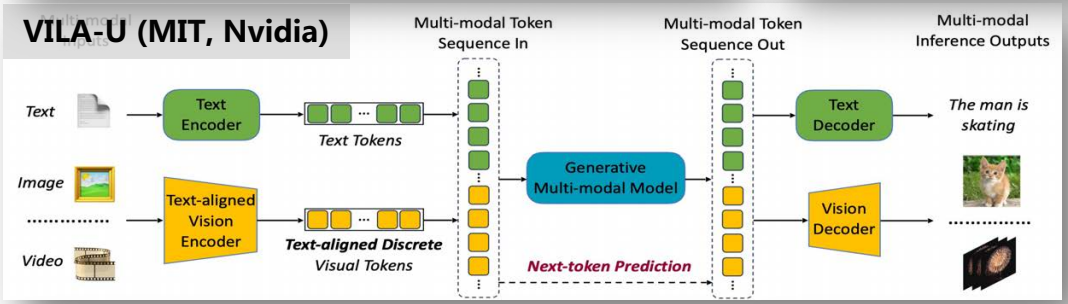
ARLON (微软)



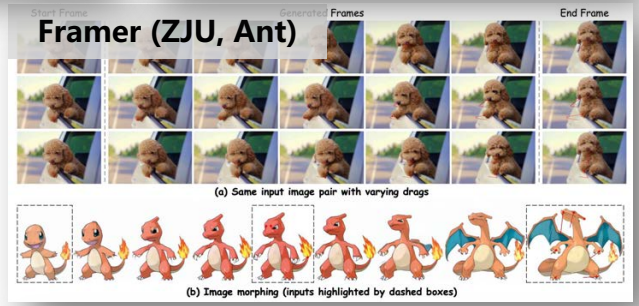
CAVIA (Google)



VILA-U (MIT, Nvidia)



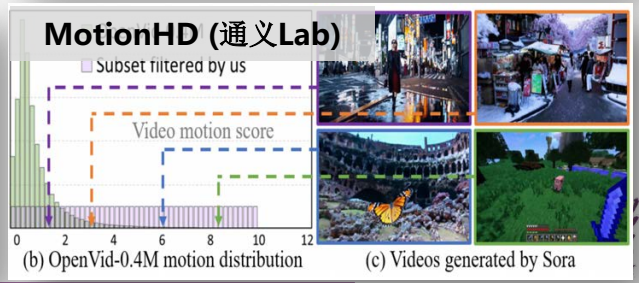
Framer (ZJU, Ant)



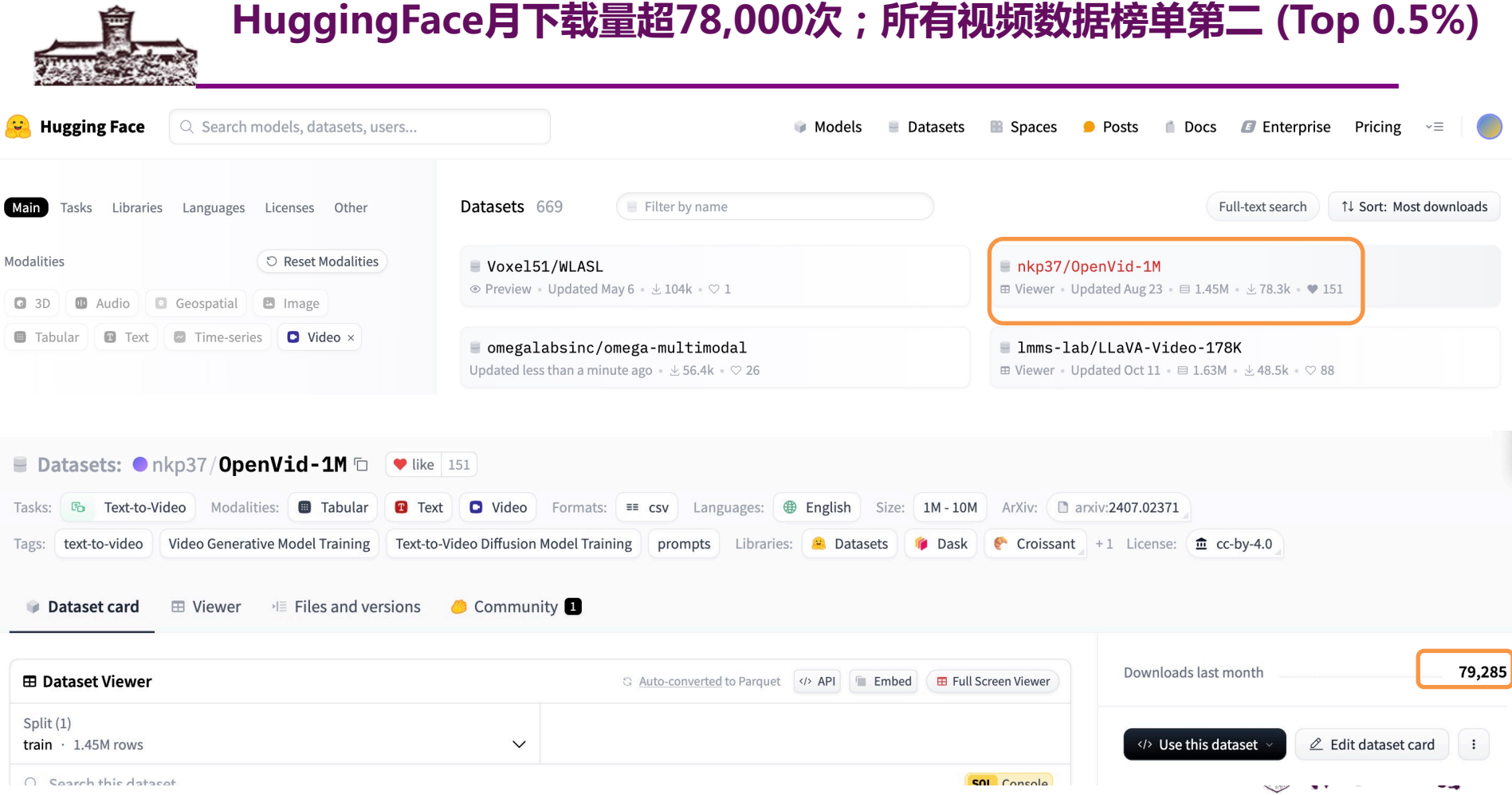
DimensionX (生数, 清华)

GenXD (NUS，微软)

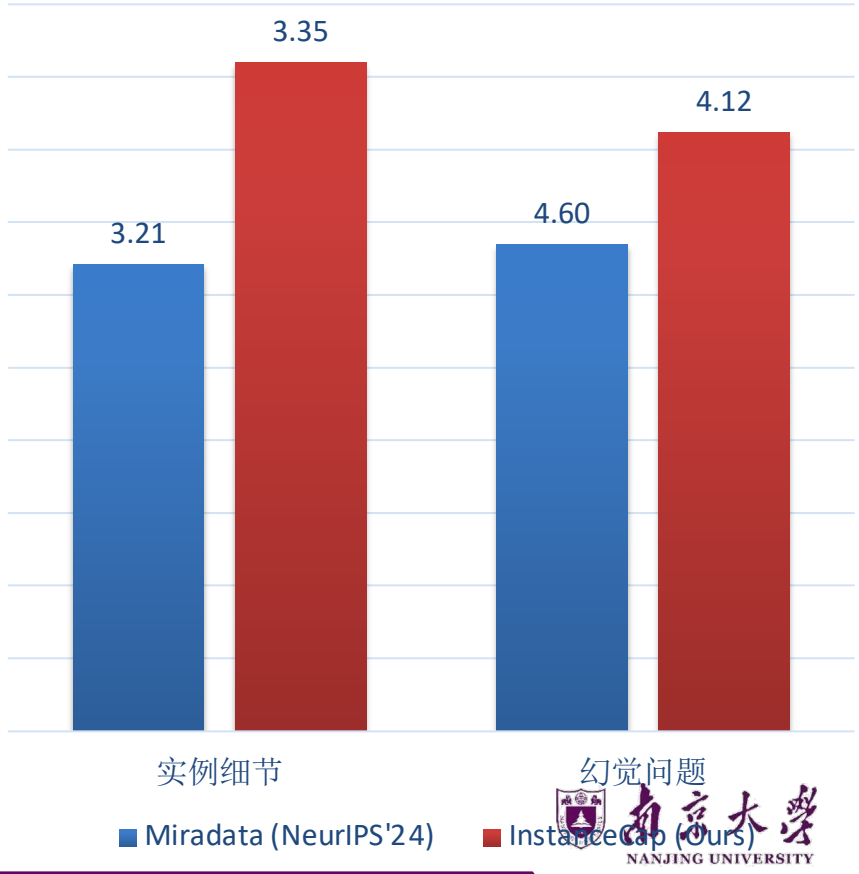
MotionHD (通义Lab)



HuggingFace月下载量超78,000次；所有视频数据榜单第二 (Top 0.5%)



拓展工作InstanceCap：细节更丰富、文本更准确的视频描述



南京大学 NANJING UNIVERSITY

拓展方法InstanceCap：流程图

Prior Information

Positive/Negative Lexicon

Chain-of-Thought Prompt

Human-Designed Class Hints

Auxiliary Model Cluster

Camera Motion Detector

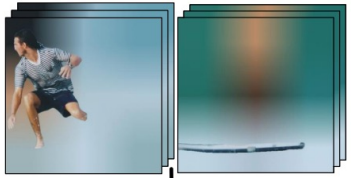
Object Detection

SAM2

Position

Class

From Global videos to Local instances



MLLMs

From Dense prompts to Structured phrases

LLaVA LLaVA-Video Qwen-VL GPT-4o

...

Appearance

Actions and Motion

Background And Camera

Results

{
"Global Description": "A man in a striped shirt and shorts skillfully surfs on a wave.",
"Structured Description": {
"Main Instance": {
"No.0": {
"Class": "person",
"Appearance": "The person in the video is wearing a horizontally striped shirt with alternating shades of white and dark colors, paired with navy blue shorts. The individual appears to be male, with short hair, and is not wearing any visible accessories such as necklaces or watches.",
"Actions and Motion": "The individual appears to be engaged in an action that involves balancing and movement, consistent with surfing. Their body is angled in a way that suggests they are riding a wave, with one arm extended for balance and the other bent at the elbow. The person's legs are bent, and their feet are positioned wide apart, indicating a stance that is both stable and ready to adjust to the motion of the water.",
"Position": "person moves from bottom-left to bottom-middle."
},
"No.1": {
"Class": "surfboard",
"Appearance": "The surfboard in the video is white with a black stripe running down the middle. The text on the surfboard reads \"SOFTOPS\" in green letters.",
"Actions and Motion": "The surfboard is in motion, gliding smoothly over the water. The board's nose occasionally pierces through the wave, and its tail follows the curve of the wave, maintaining balance. The surfboard's white and black stripes are clearly visible, and it appears to be a longboard style, suitable for the waves present in the video.",
"Position": "surfboard moves from bottom-left to bottom-middle."
}
},
"Background Detail": "The environment in the video is a vibrant and dynamic ocean scene. The water is a clear, turquoise blue-green color, with white, frothy waves crashing and splashing around. The sunlight reflects off the surface of the water, creating a sparkling effect. The waves are in constant motion, with some cresting and others breaking, indicating a strong current or wind. The overall atmosphere is energetic and refreshing, with the sound of the waves implied by their appearance.",
"Camera Movement": "The camera movement in the video is a zoom-in, which occurs gradually and smoothly. The shooting angle appears to be at medium distance, capturing the surfer and the wave from a side perspective. The camera does not move rapidly or sharply; instead, it maintains a steady focus on the surfer as they ride the wave."
}

拓展应用STAR：结合视频大模型先验的真实场景视频复原

欢迎下载OpenVid-1M、InstanceCap并使用

OpenVid - Code

OpenVid - Dataset & Models

InstanceCap - Code

分享内容简介

- 面向高质量视频生成的数据集OpenVid-1M

- 一个高质量的文生视频数据集，**最高支持1080p视频生成**
- 提出了一种**多模态视频DiT模型结构(MVDiT)**
- 在视频生成、视频复原、视频插帧、3D/4D生成等任务中被使用

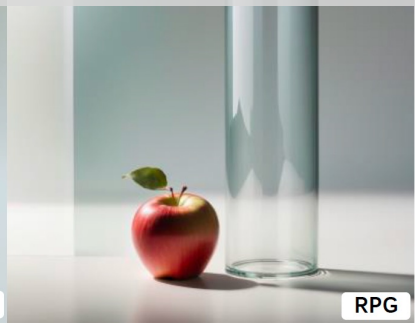
- 面向高质量图像分区生成的方法**RAG-Diffusion**

- 无需微调训练：**无缝兼容现有DiT框架**（如Flux, SD3）
- 精确区域控制：相比以往Flux-1.dev实现**准确生成复杂布局**
- 重绘功能：实现修改特定区域而不影响其他部分

以往文生图方法的问题：目标数量、交互、行为、属性控制

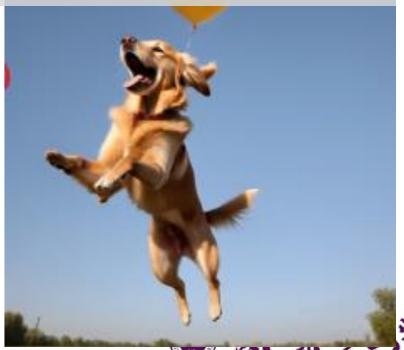
*Seven ceramic mugs in different colors are placed on a wooden table, with numbers from **1 to 7 written on the cups**, and a bunch of white roses on the left.*

*A cylindrical glass, **obscuring the right half** of the apple behind it.*



*A woman **walks toward** us in the forest, holding a dog on a leash*

*A balloon **on the bottom** of a dog.*



NANJING UNIVERSITY

RAG-Diffusion动机：高可控、高质量、低成本

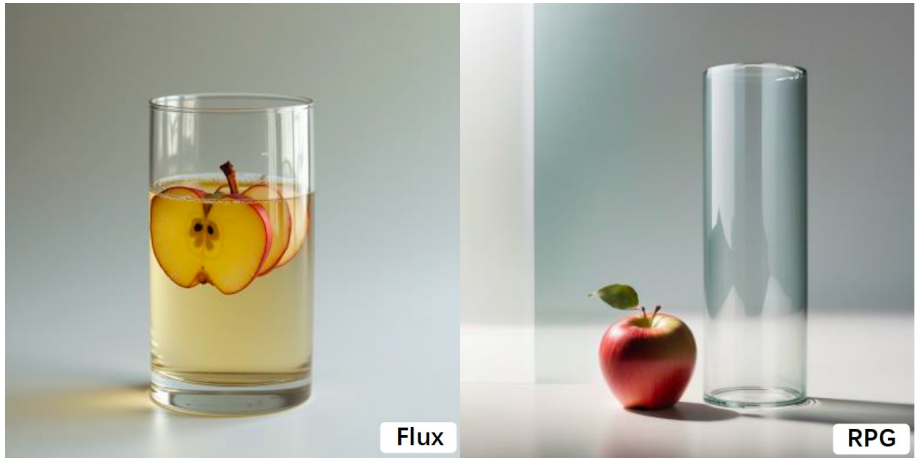
*On the left, Einstein is **painting** the Mona Lisa; in the center, Elon Reeve Musk is participating in the **U.S. presidential election**; on the right, Trump is hosting a **Tesla product launch**.*

RAG-Diffusion：更好的多目标生成和属性绑定

*Seven ceramic mugs in different colors are placed on a wooden table, with numbers from **1 to 7 written on the cups**, and a bunch of white roses on the left.*

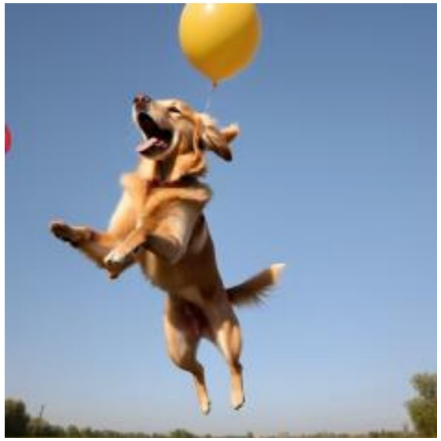
RAG-Diffusion：更好的交互关系（遮挡、材质）

*A cylindrical glass, **obscuring the right half** of the apple behind it.*



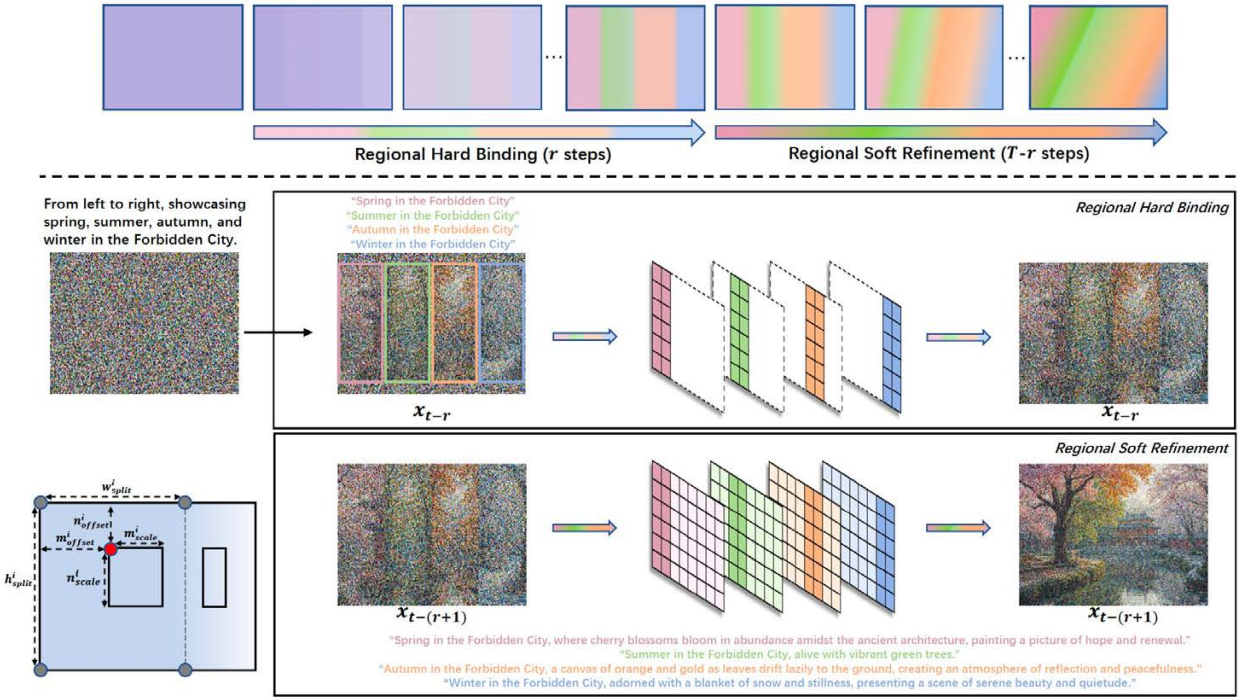
RAG-Diffusion：更好的空间关系

*A balloon **on the bottom** of a dog.*



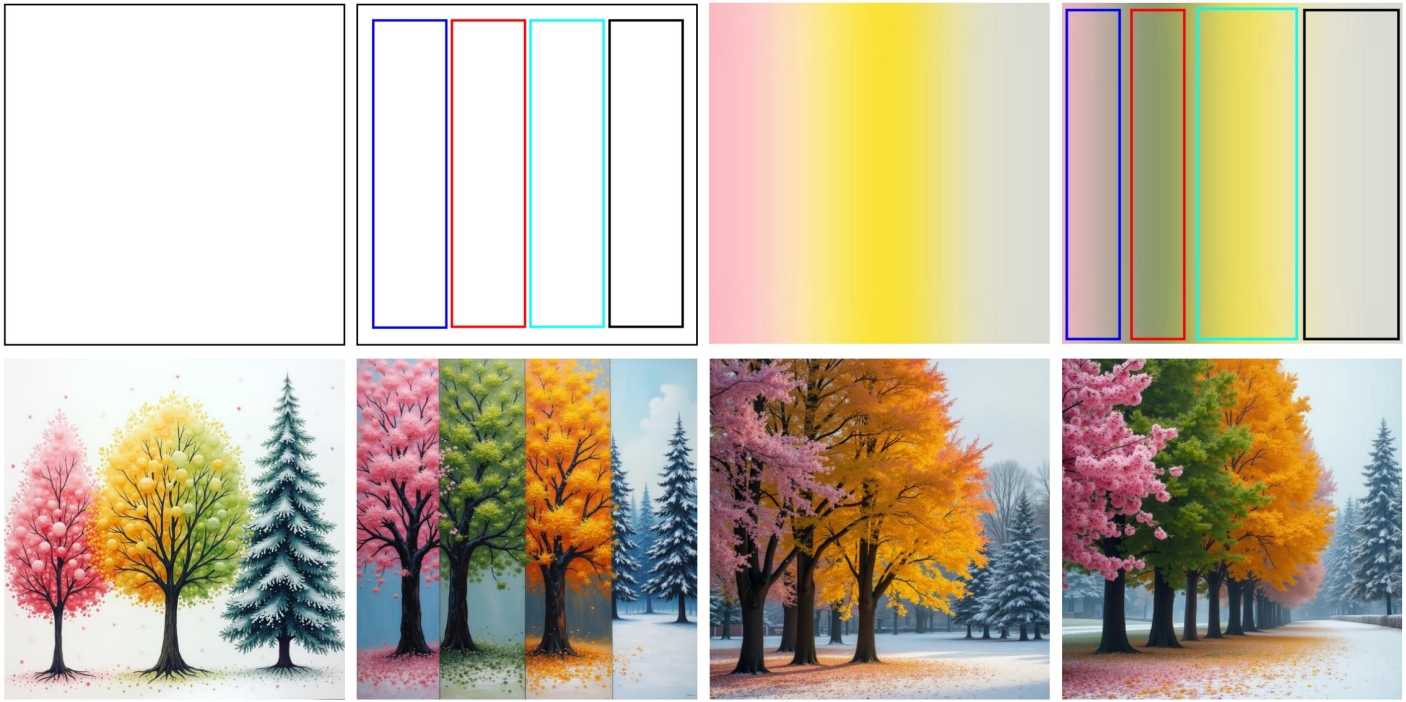
RAG-Diffusion：关注2大研究问题

- **研究问题1：如何更好地控制区域**
  - **Regional Hard Binding：** 将prompt根据目标拆解，每个目标局部去噪后再合并
  - **Regional Soft Refinement：** 对每个区域进行丰富描述，全局去噪，在CA特征层crop出目标区域后再合并
- **研究问题2：如何更好地实现指令跟随**
  - 与最好的DiT模型结合，如FLUX
  - ‘分而治之’的整体策略



RAG-Diffusion的设计优化了2类组合生成问题

- 优化目标遗漏问题
- 优化邻近区域分界问题



w/o Hard Binding
w/o Soft Refinement

w/ Hard Binding
w/o Soft Refinement

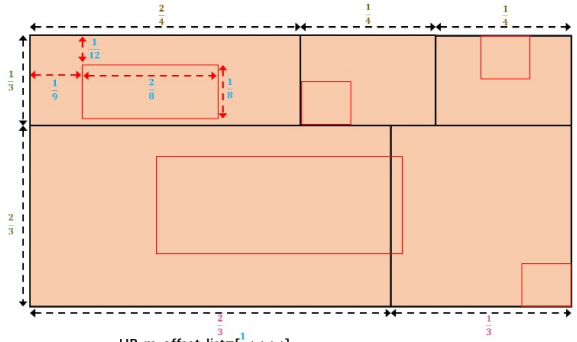
w/o Hard Binding
w/ Soft Refinement

w/ Hard Binding
w/ Soft Refinement

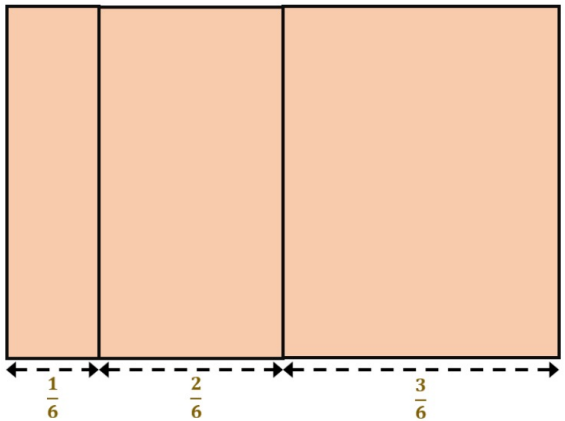
南京大學
NANJING UNIVERSITY

RAG-Diffusion：三类基本布局

▼ The following shows several schematic diagrams of `HB_m_offset_list`, `HB_n_offset_list`, `HB_m_scale_list`, `HB_n_scale_list`, `SR_hw_split_ratio`.



HB_m_offset_list=[$\frac{1}{9}$,*,*,*,*]
HB_n_offset_list=[$\frac{1}{12}$,*,*,*,*]
HB_m_scale_list=[$\frac{2}{8}$,*,*,*,*]
HB_n_scale_list=[$\frac{1}{8}$,*,*,*,*]
SR_hw_split_ratio=[$\frac{1}{3}$, $\frac{2}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{2}{3}$, $\frac{1}{3}$]

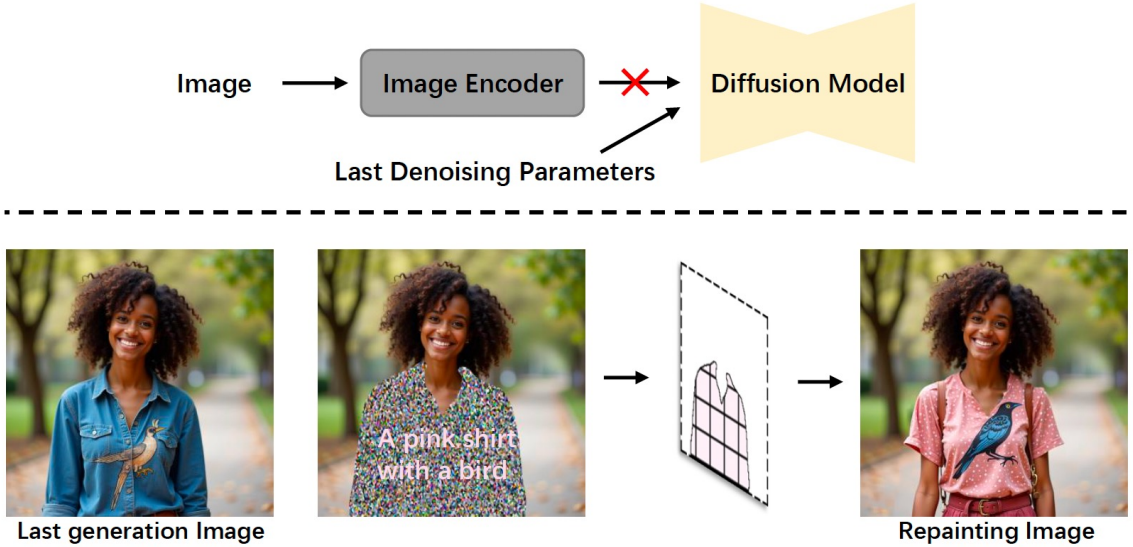


SR_hw_split_ratio=[$\frac{1}{6}$, $\frac{2}{6}$, $\frac{3}{6}$]

SR_hw_split_ratio=[$\frac{1}{2}$, $\frac{1}{2}$]

RAG-Diffusion：Repainting重绘功能

- **Repainting：**修正不满意的局部区域，同时保持其他区域不受影响



RAG-Diffusion：重绘效果展示

Text prompt: "A vase and an apple."

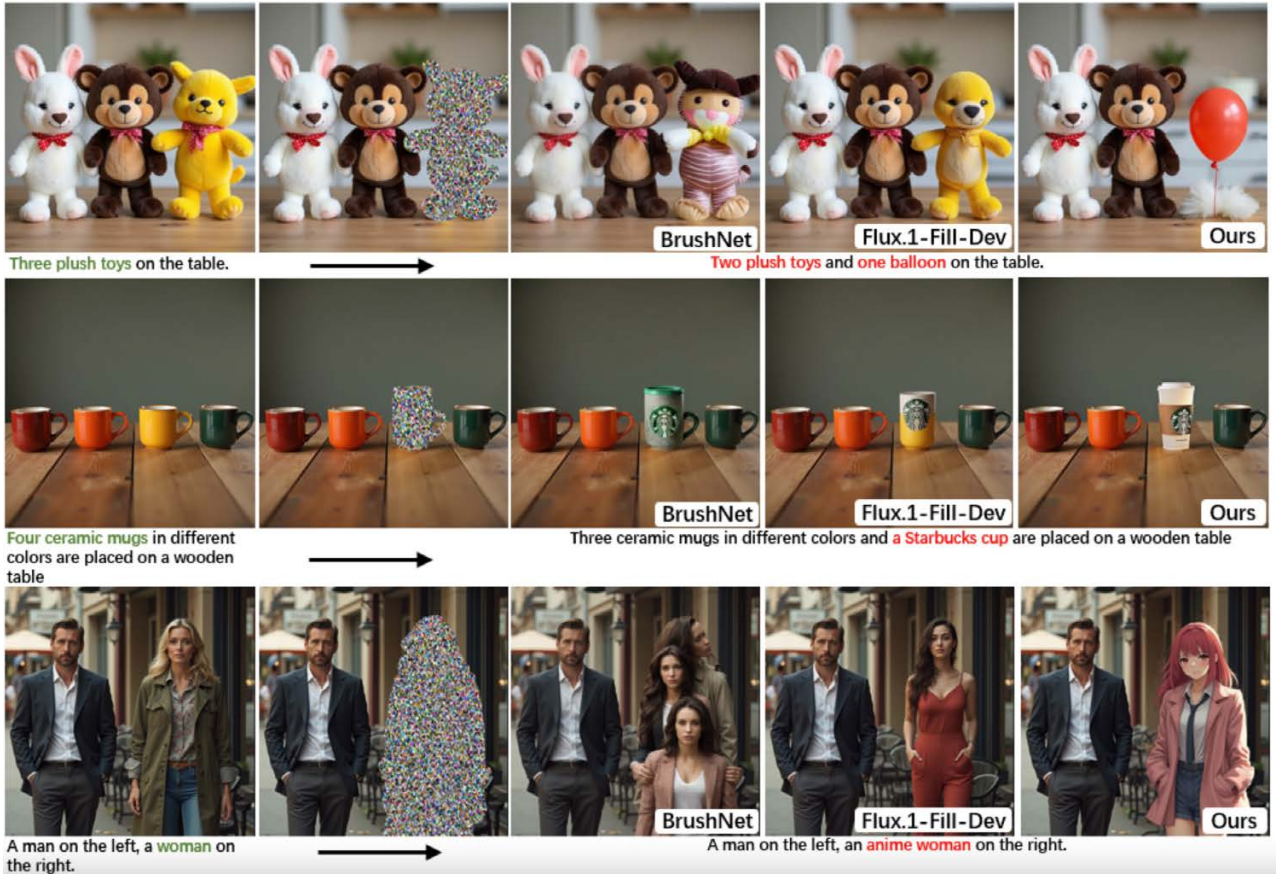
Repainting prompt: "A vase and a Rubik's Cube."

RAG-Diffusion：重绘效果展示

Text prompt: "A brown curly hair African woman in blue puffy skirt."

Repainting prompt: "A brown curly hair African woman in pink suit."

RAG-Diffusion：更多repainting对比结果



- 与inpainting领域sota方法Brushnet（ECCV 2024）的更多例子对比，效果有优势
- 与最近刚发布的Flux.1-Fill-Dev对比，在一些场景下效果可比

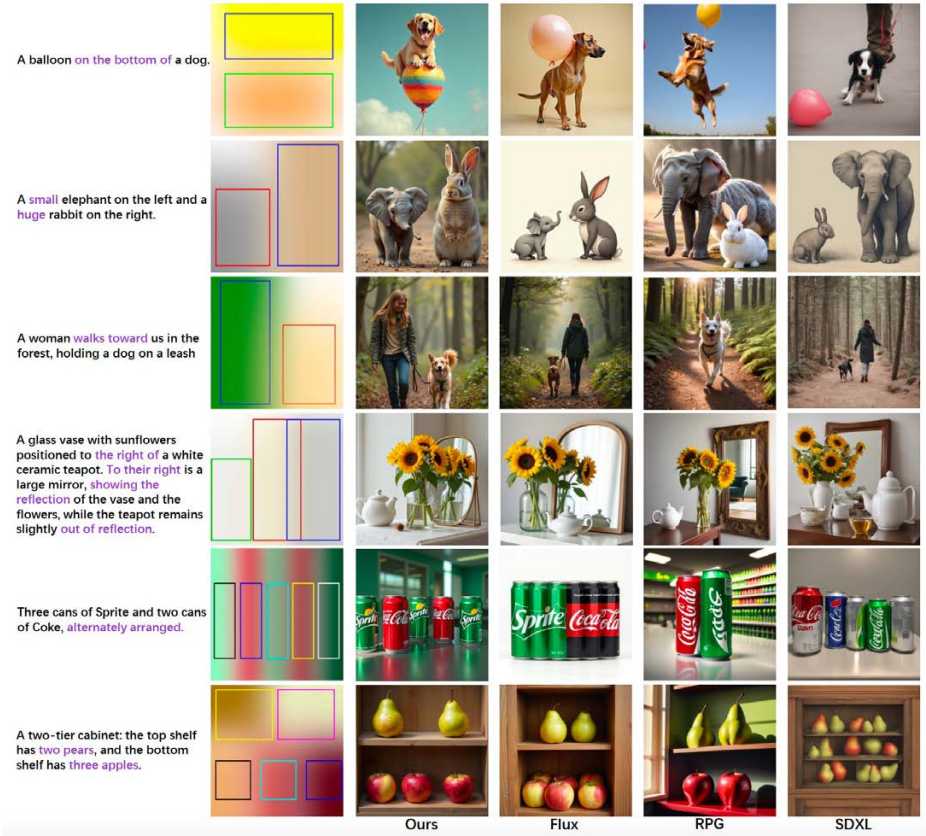
RAG-Diffusion：定量实验结果

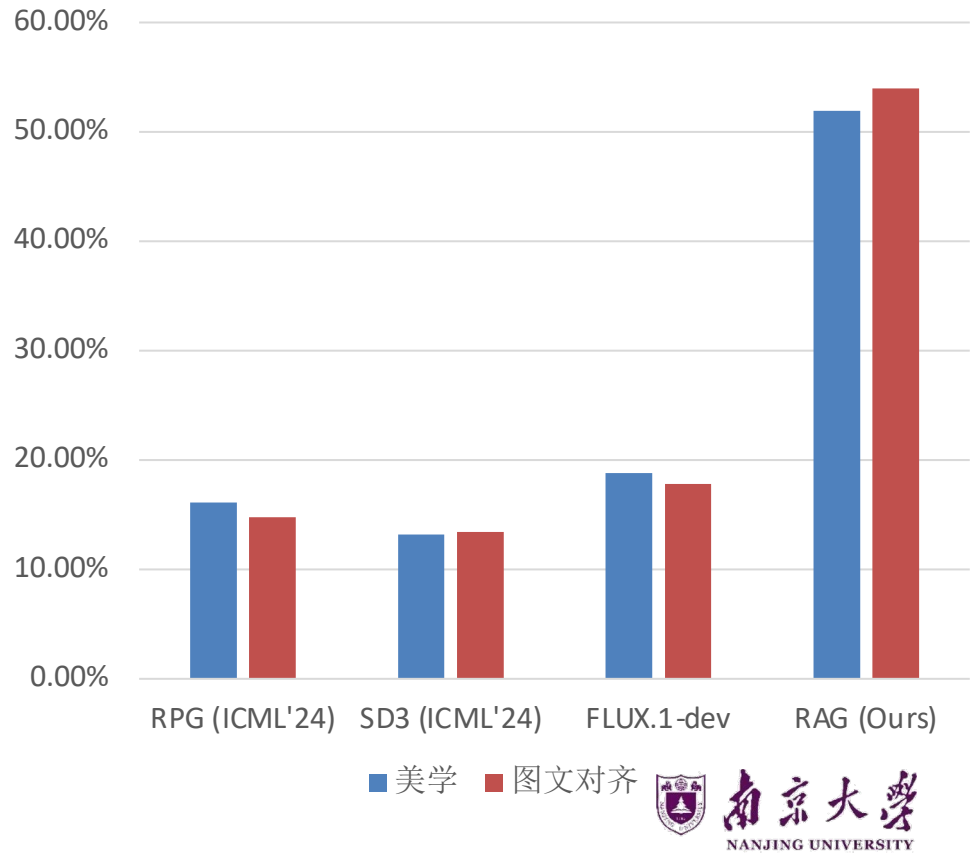
- 指标全面优于以往组合生成方法；相比**baseline Flux.1-dev**方法也有显著提升，尤其是空间关系方面

Model	Attribute Binding			Object Relationship		Complex ↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
Stable v1.4 [26] [CVPR 2022]	0.3765	0.3576	0.4156	0.1246	0.3079	0.3080
Composable v2 [19] [ECCV 2022]	0.4063	0.3299	0.3645	0.0800	0.2980	0.2898
Structured v2 [9] [ICLR 2023]	0.4990	0.4218	0.4900	0.1386	0.3111	0.3355
Stable v2 [26] [CVPR 2022]	0.5065	0.4221	0.4922	0.1342	0.3127	0.3386
Stable XL [2] [2022]	0.5879	0.4687	0.5299	0.2133	0.3119	0.3237
Attn-Exct v2 [4] [TOG 2023]	0.6400	0.4517	0.5963	0.1455	0.3109	0.3401
GORS [13] [Neurips 2023]	0.6603	0.4785	0.6287	0.1815	0.3193	0.3328
Pixart-α-ft [5] [ICLR 2024]	0.6690	0.4927	0.6477	0.2064	<u>0.3197</u>	0.3433
RPG* [39] [ICML 2024]	0.7476	<u>0.5640</u>	<u>0.6724</u>	<u>0.4017</u>	0.3032	<u>0.3702</u>
Flux.1-dev* [3] [2024]	<u>0.7680</u>	0.5078	0.6195	0.2606	0.3078	0.3650
Ours	**0.8039**	**0.6016**	**0.7085**	**0.5193**	**0.3263**	**0.4377**

Table 1. Comparison of alignment evaluation on T2ICompBench [13]. The best results are highlighted in **bold**, second-best in <u>underline</u>. The basic data is sourced from [13]. * indicates results we reproduced using the official open-source codes and configurations.

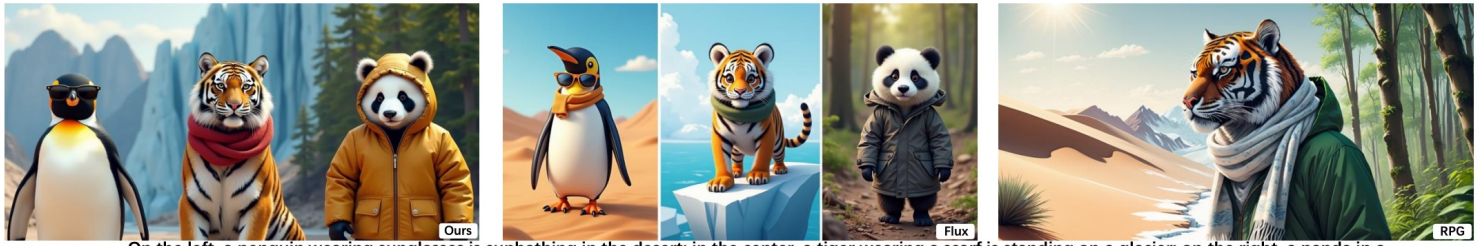
RAG-Diffusion：视觉对比效果、主观评测效果



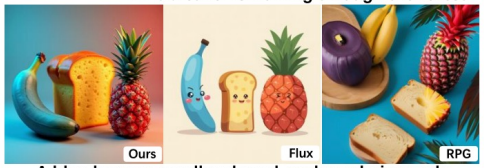


南京大學 NANJING UNIVERSITY

RAG-Diffusion：更多视觉对比效果

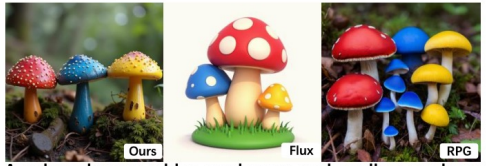


On the left, a penguin wearing sunglasses is sunbathing in the desert; in the center, a tiger wearing a scarf is standing on a glacier; on the right, a panda in a windbreaker is walking through the forest.



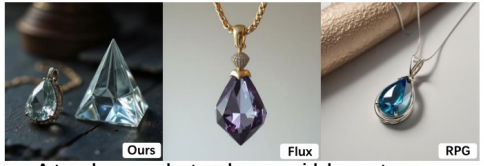
A blue banana, a yellow bread, and a red pineapple.

A purple sunflower, a pink newspaper, and a green cat.

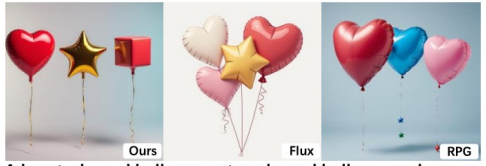


A red mushroom, a blue mushroom, and a yellow mushroom

A spherical Rubik's cube and a triangular iPad.



A teardrop pendant and a pyramidal gemstone.



A heart-shaped balloon, a star-shaped balloon, and a square balloon.

A straw hat on the bottom of a soccer.

A turtle on the bottom of a phone.

From top to bottom, a rabbit, a dog, and a cat stacked together.

RAG-Diffusion拓展：适配SD3

- RAG-Diffusion支持多个DiT类框架

Two green cans and two red cans, alternately arranged.

From left to right, an orange kitten wearing a yellow hat, a gray Schnauzer wearing a pink hat, and a white duck wearing a green hat.

RAG-Diffusion拓展：适配Flux-redux模型，支持图文输入

The left side is a skeleton with fire, and the right side is an ice dragon

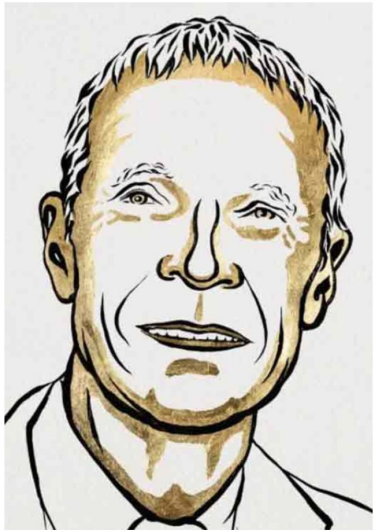
RAG-Diffusion拓展：适配Flux-redux模型，支持图文输入

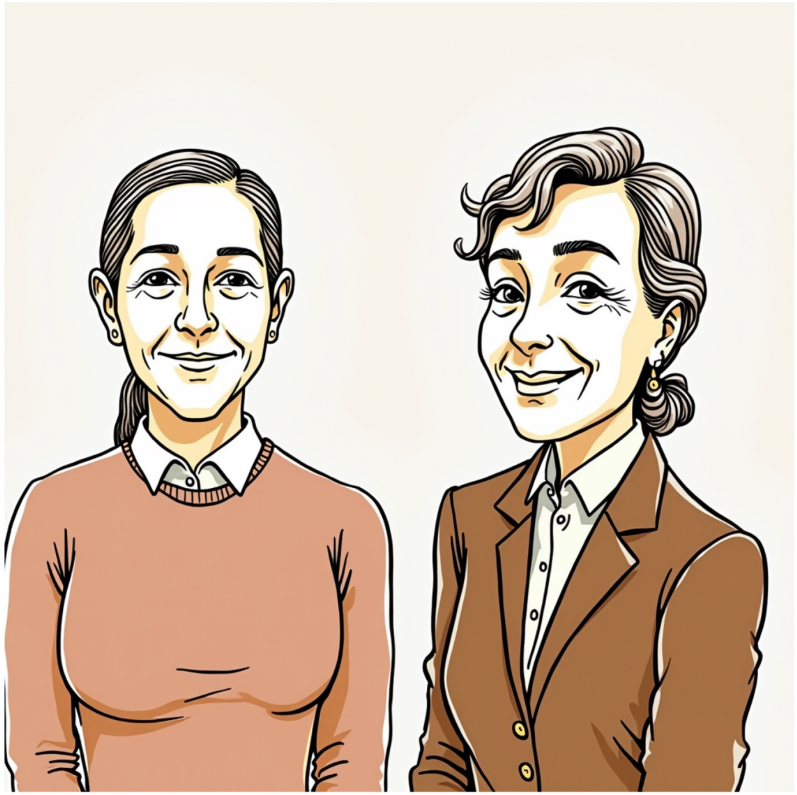




Four ceramic mugs are placed on a wooden table

RAG-Diffusion拓展：适配Flux-redux模型，支持图文输入



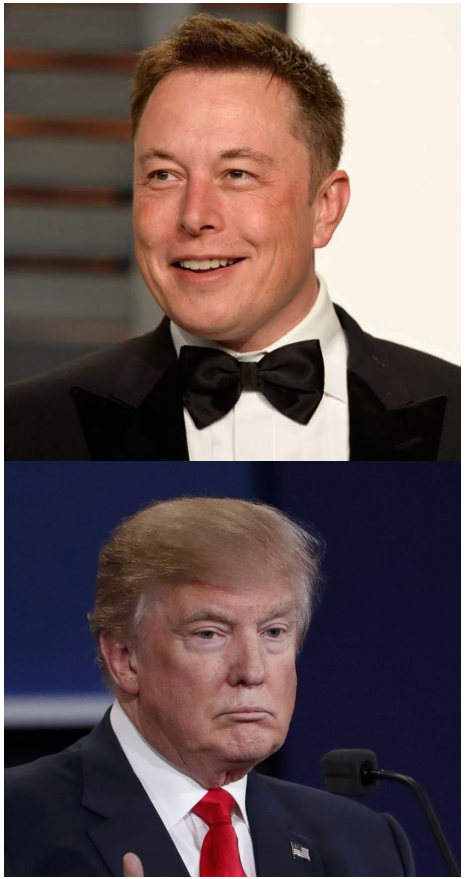


Two women in an illustration style.

RAG-Diffusion拓展：适配PuLID模型，人像ID保持更好

RAGD +
PuLID

RAG-Diffusion拓展：适配PuLID模型，人像ID保持更好



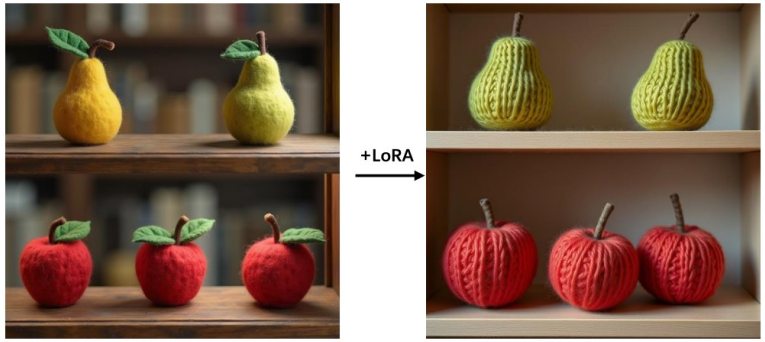
RAGD +
PuLID

RAG-Diffusion拓展：适配PuLID模型，人像ID保持更好

RAG-Diffusion拓展：适配社区Lora模型

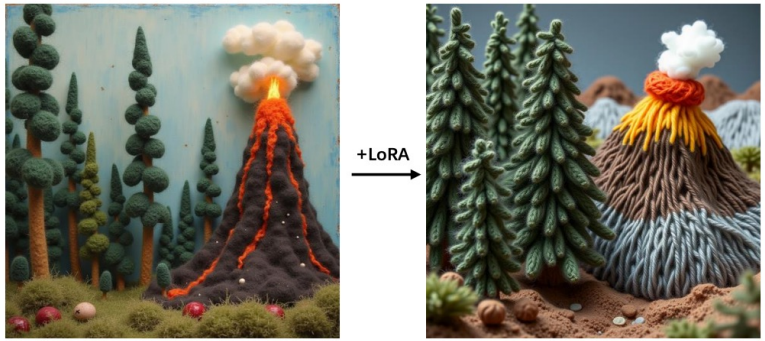
- 无缝适配黑神话悟空、Wool Yarn（羊毛纱）等Lora

A man on the left is holding a bag and a man on the right is holding a book



A two-tier cabinet: the top shelf has two pears made of wool, and the bottom shelf has three apples made of wool.

A mountain on the left, a crouching man in the middle, and an ancient architecture on the right.



On the left is a forest made of wool, and on the right is a volcano made of wool.

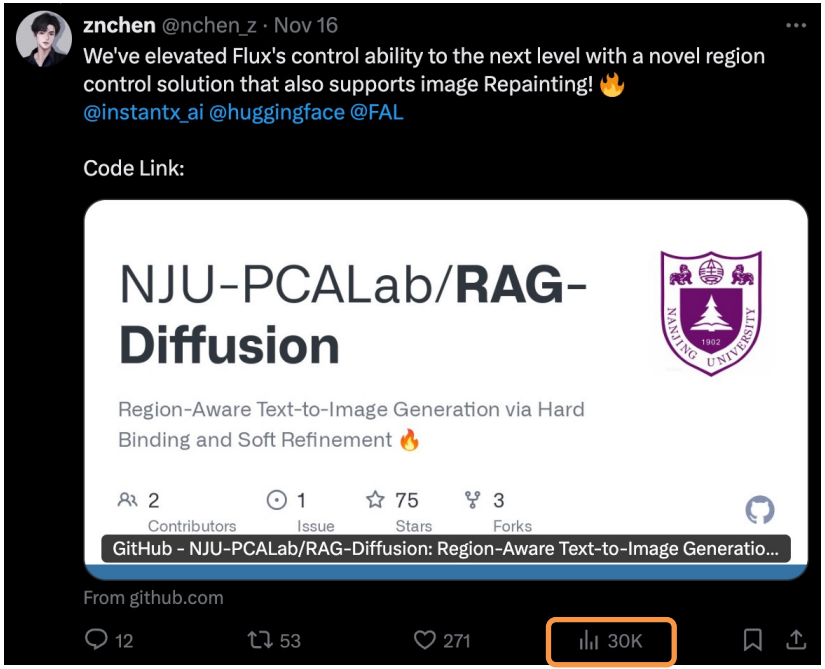
RAG-Diffusion：稳定性

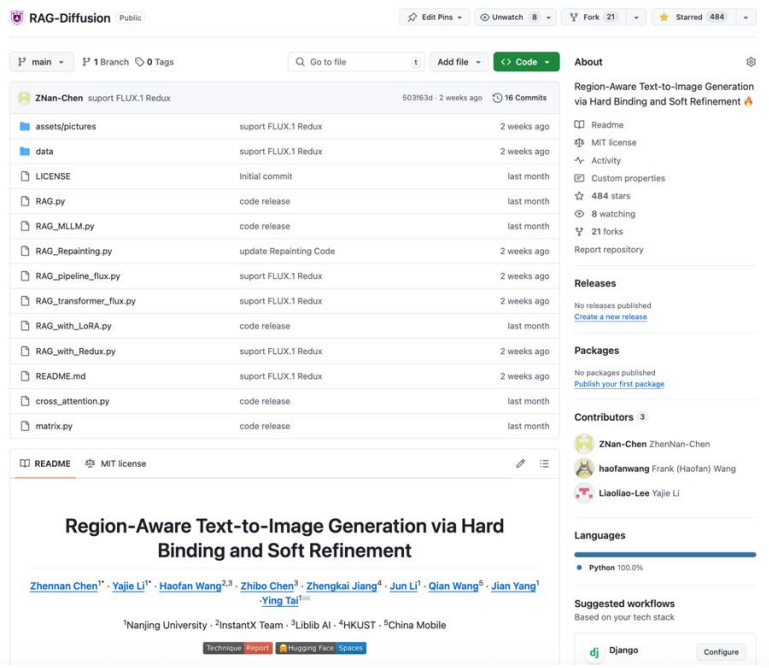
- RAG-Diffusion生成稳定性相比Flux，RPG更高

Seven ceramic mugs in different colors are placed on a wooden table, with numbers from 1 to 7 written on the cups, and a bunch of white roses on the left.

RAG-Diffusion：已开源代码、HF体验链接，欢迎下载使用

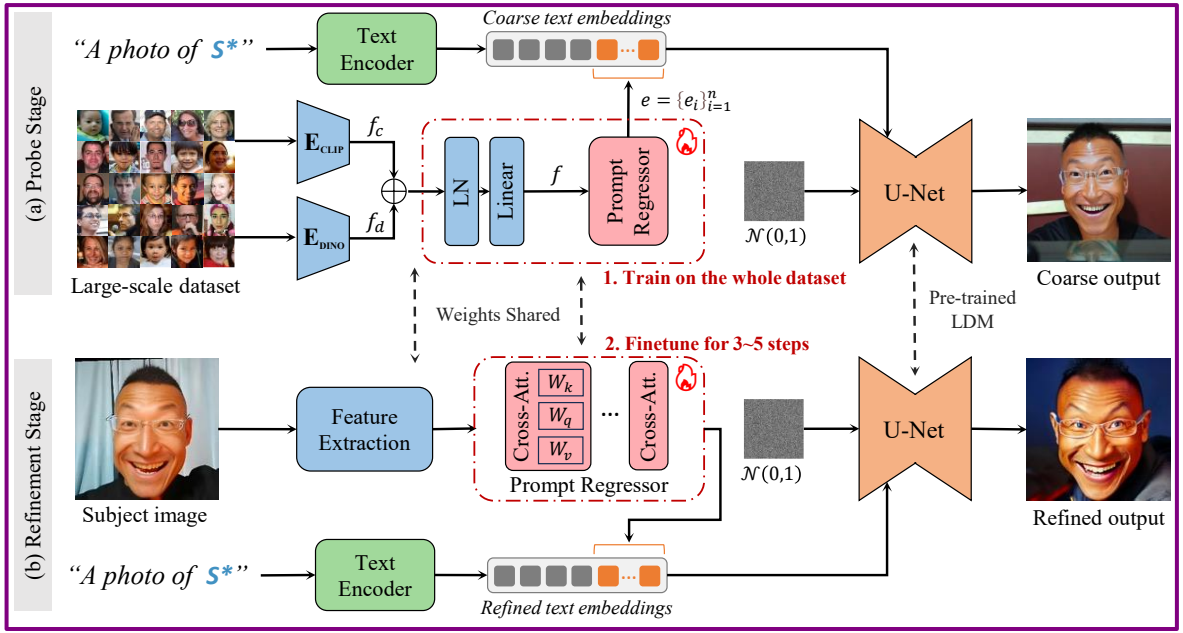
- 在X平台上引起一定热度的讨论，已开源代码（~500 Stars）、HuggingFace体验链接

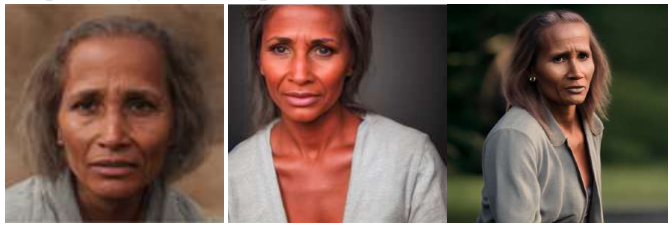




图像生成相关工作：HybridBooth（ECCV 2024）

- **Self-supervised correction framework for the subject-driven generation.**





南京大学 NANJING UNIVERSITY

结论

- OpenVid-1M是一个高质量的文生视频数据集，旨在提升视频质量，具有**高美学、清晰度和分辨率的特点，最高支持1080p视频生成**；【TODO：更高的动态性、更强的可控性？与AR、World model架构结合】

- RAG-Diffusion是一个无需训练的区域感知文生图方法，实现**更精确的多目标组合生成**；同时支持**局部重绘功能**，保持其他区域内容不变。【TODO：从图像到视频/3D/4D？】

OpenVid-1M Code

OpenVid-1M

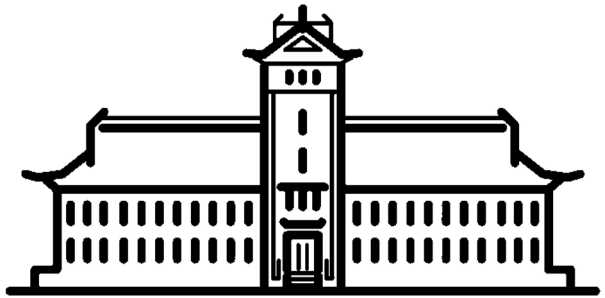
InstanceCap Code

RAG-Diffusion Code

Dataset & Models

南京大學 NANJING UNIVERSITY

谢谢



南京大學
NANJING UNIVERSITY