中国图象图形学学会
CHINA SOCIETY OF IMAGE AND GRAPHICS

2024 中国图象图形大会
CHINESE CONGRESS ON IMAGE AND GRAPHICS

AIGC视觉内容生成与溯源研究进展

王岚君

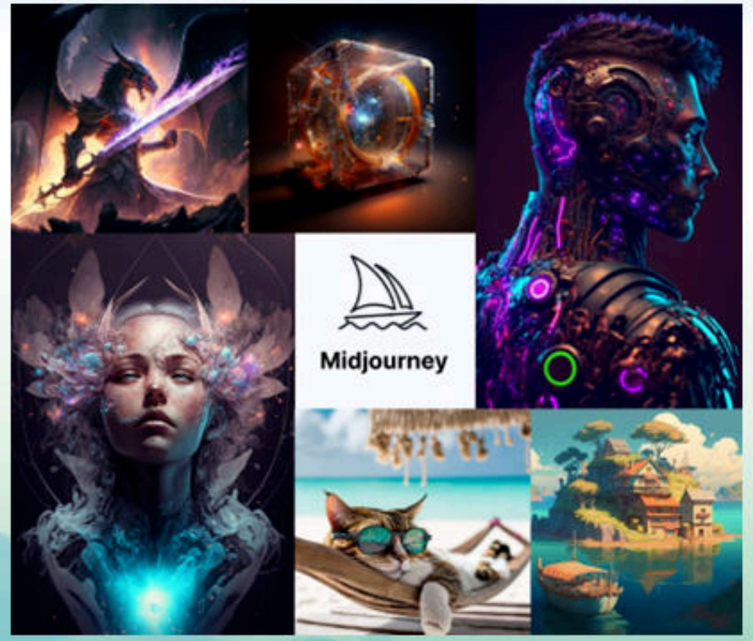
2024年5月

论文引用格式：Liu Anan, Su Yuting, Wang Lanjun, Li Bin, Qian Zhenxing, Zhang Weiming, Zhou Linna, Zhang Xinpeng, Zhang Yongdong, Huang Jiwu, Yu Nenghai. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(06):1535-1554(刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(06):1535-1554)[DOI:10.11834/jig.240003]

报告内容

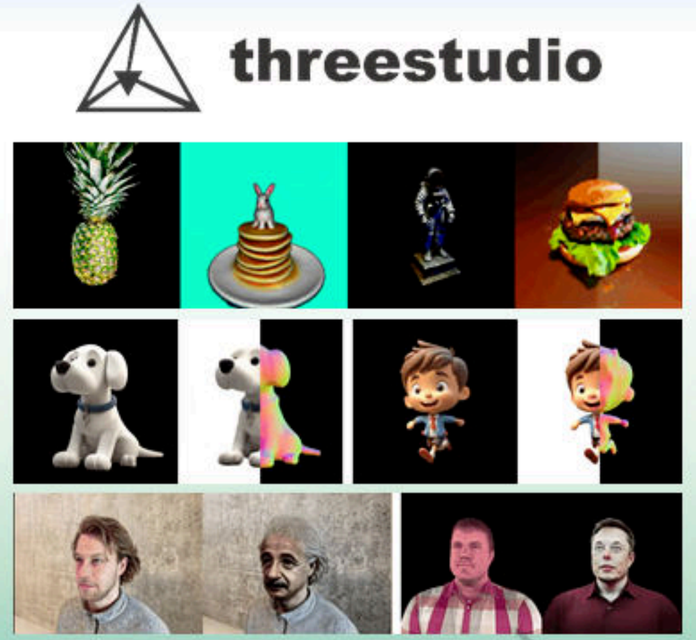
研究背景

- 视觉内容生成技术发展迅速，为多媒体内容创作提供有力工具。



图像内容生成
Midjourney [1]

视频内容生成
OpenAI Sora [2]



3D内容生成
threestudio [3]

[1] https://openai.com/sora/
[2] https://midjourney.com
[3] https://github.com/threestudio-project/threestudio

研究背景

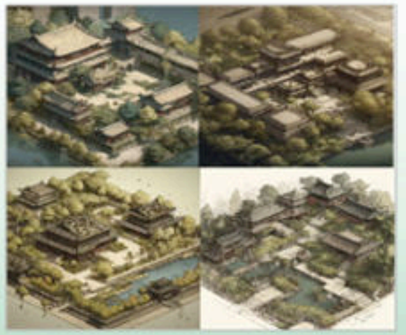
- 图像生成技术发展迈入成熟，应用领域广泛
  - 提高生产效率、变革生产方式

图像生成服务

GOOGLE IMAGEN AI
DALL·E 2
文心一格
Midjourney
通义万相
runway
Adobe Firefly
Clipdrop by stability.ai

应用领域广泛

推文内容生成



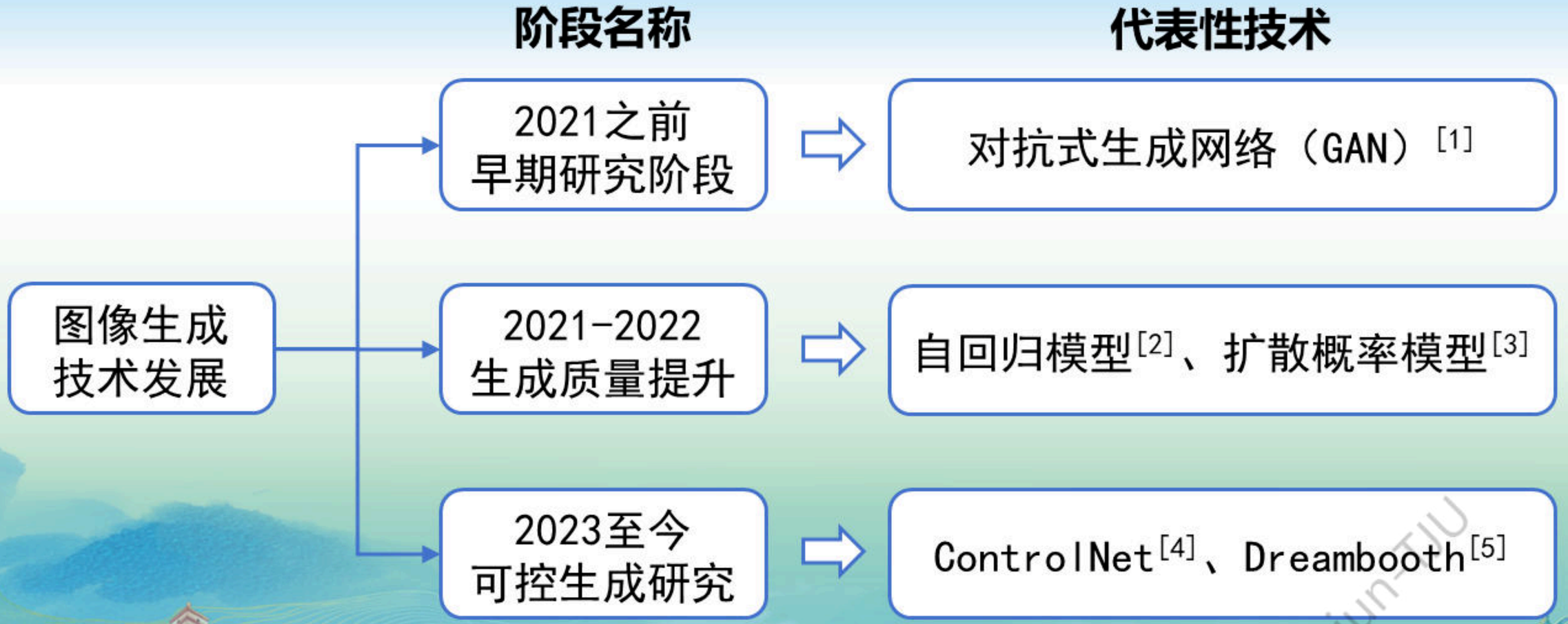
建筑设计

艺术创作

AI模特

研究背景

■ 图像生成技术发展阶段



[1] Goodfellow, Ian, et al. "Generative adversarial nets." *NeurIPS*. 2014.
[2] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *ICML*. 2021.
[3] Ho J, Jain A, Abbeel P. " Denoising diffusion probabilistic models." *NeurIPS*. 2020.
[4] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *ICCV*. 2023.
[5] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *CVPR*. 2023.

研究背景

- 图像生成技术的成熟与快速发展
  - 2020-2022年，图像生成技术研究出现多个跨越式突破。

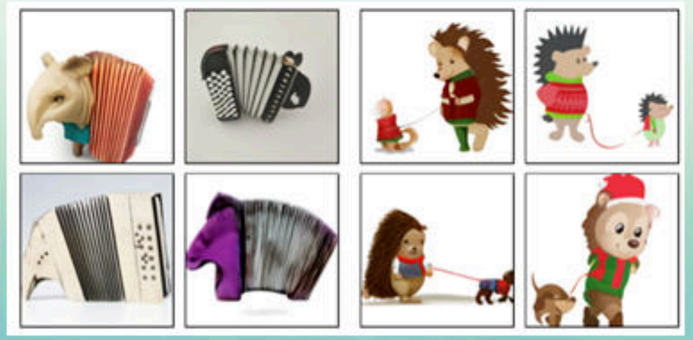
DF-GAN[1]（2020）

对抗式生成网络（GAN）
参数量：19M
训练数据：
20万图文对（MS-COCO）

受训练数据集限制，仅能生成训练数据相似类型图像

DALL-E[2]（2021）

自回归模型
参数量：**12B**（~600倍）
训练数据：
2.5亿图文对（网络收集）



支持**零样本生成**
根据文本描述生成任意内容

LDM[3]（2022）

扩散概率模型
参数量：**约1B**
训练数据：
14.5亿图文对（LAION）

生成**质量显著提升**
支持更加复杂场景的零样本生成

[1] Tao, Ming, et al. "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis." *arXiv*. 2020.
[2] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *ICML*. 2021.
[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *CVPR*. 2022.

研究背景

- 图像生成技术的成熟与快速发展
  - 伴随着成熟商业模型的发布（2022年），生成图像已经具备极高视觉质量。
  - 生成模型“可控性不足”限制了技术的实际应用。

艺术创作（MidJourney）

动漫形象（MidJourney）

商品海报设计（MidJourney）

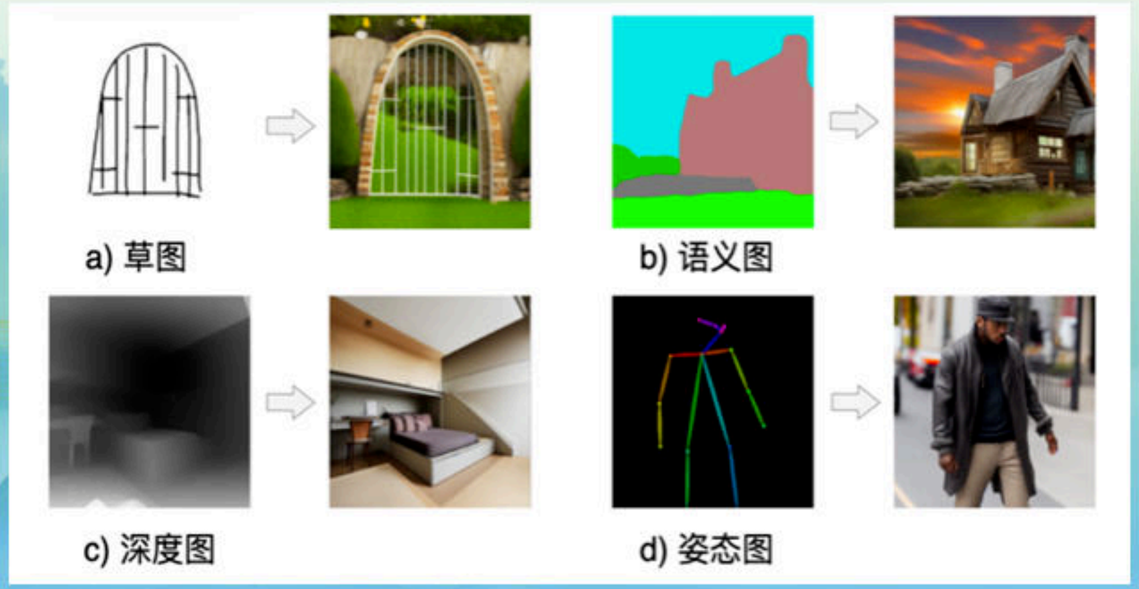
研究背景

- “可控生成”技术成为研究热点
  - 可控生成：根据创作者的特定需求和条件生成图像。

画面布局可控需求

⇩

基于布局条件的可控图像生成



视觉主体可控需求

⇩

基于视觉主体的可控图像生成

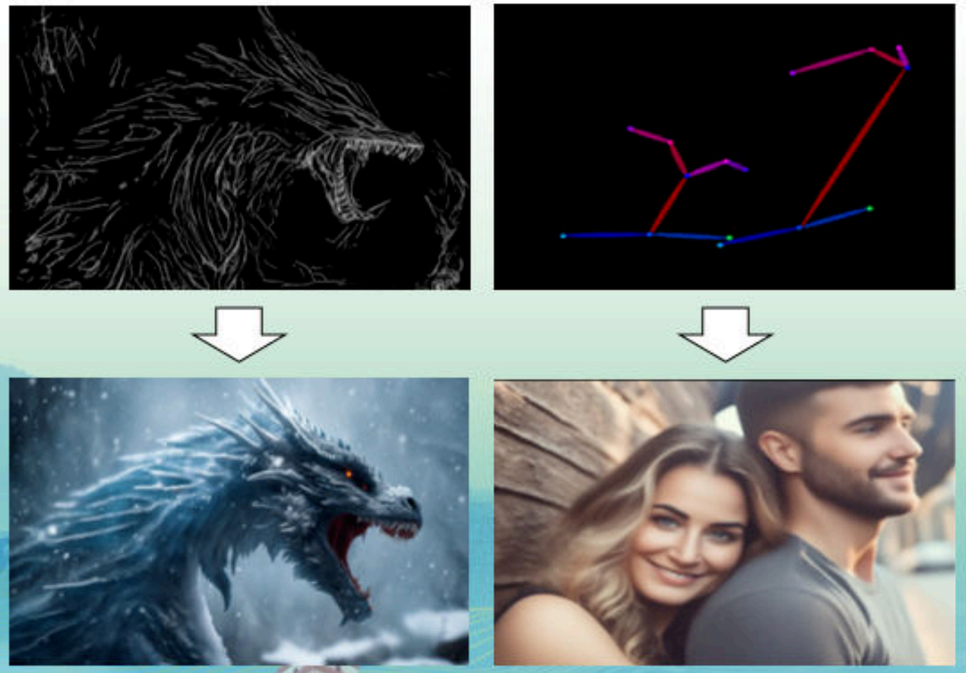
参考图片（3~5张）

围绕视觉主体进行场景生成

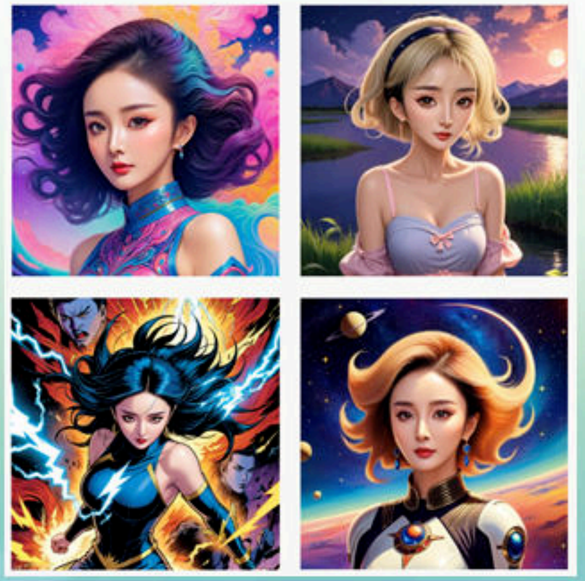
研究背景

■ “可控生成”为图像生成带来广泛实际应用

画面布局可控
（艺术创作）



视觉主体可控
（创意人像、商品海报定制）



[1] Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models.”*AAAI.* 2024.
[2] Li, Zhen, et al. "Photomaker: Customizing realistic human photos via stacked id embedding.” *CVPR.* 2024.
[3] Chen, Wang, et al. ”AnyScene: Customized Image Synthesis with Composited Foreground” *CVPR.* 2024.

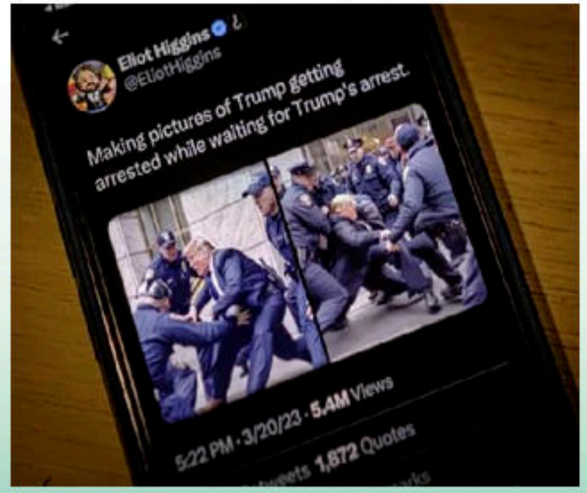
研究背景

■ 视觉生成模型引发的争议

产权争议

由于涉及版权侵犯，艺术家们对Stability AI、Midjourney等提起集体诉讼

新闻伪造

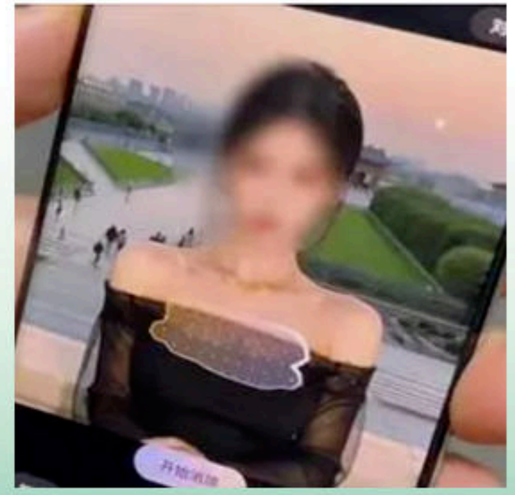


特朗普的合成“被捕照”在社交媒体谣传，误导民众的舆论和决策

肖像侵犯

泰勒·斯威夫特大量虚假“不雅照片”，浏览量高达上千万

隐私争议



华为小艺AI修图导致照片人物的衣物意外消失

如何实现视觉生成模型创作内容的**溯源**是亟需解决的问题

研究背景

■ 国内外相关政策

国内政策

互联网信息服务深度合成管理规定

发布时间: 2022-12-31 12:25 浏览次数: 1070次 来源: 公安部

国家互联网信息办公室

中华人民共和国工业和信息化部令

中华人民共和国公安部

第12号

《互联网信息服务深度合成管理规定》已经2022年11月3日国家互联网信息办公室2022年第21次室务会议审议通过，并经工业和信息化部、公安部同意，现予公布，自2023年1月10日起施行。

国家互联网信息办公室主任 庄荣文

工业和信息化部部长 金壮龙

公安部部长 王小洪

2022年11月25日

互联网信息服务深度合成管理规定

生成式人工智能服务管理暂行办法

国家互联网信息办公室

中华人民共和国国家发展和改革委员会

中华人民共和国教育部

中华人民共和国科学技术部

中华人民共和国工业和信息化部

中华人民共和国公安部

国家广播电视总局

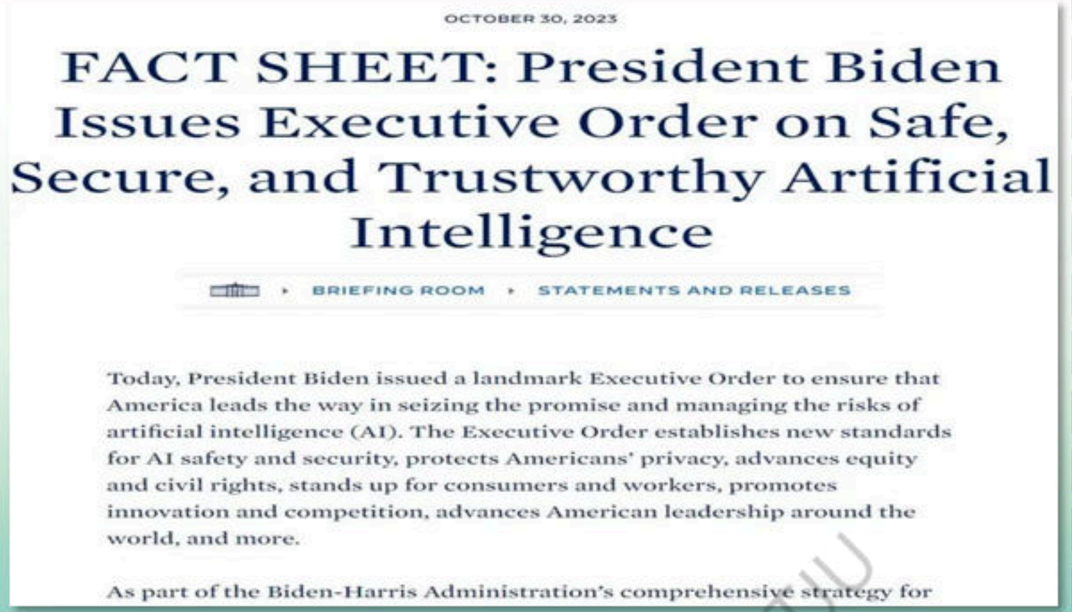
令

第15号

《生成式人工智能服务管理暂行办法》已经2023年5月23日国家互联网信息办公室2023年第12次室务会会议审议通过，并经国家发展和改革委员会、教育部、科学技术部、工业和信息化部、公安部、国家广播电视总局同意，现予公布，自2023年8月15日起施行。

2023年1月10日，网信办实施《互联网信息服务深度合成管理规定》，随后在2023年7月13日公布《生成式人工智能服务管理暂行办法》， 以 促进**生成式人工智能健康发展**和规范应用。

国外政策



OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM • STATEMENTS AND RELEASES

Today, President Biden issued a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for

2023年10月30日，拜登签署关于生成式人工智能监管的首套行政令，要求对**人工智能进行新的安全评估，制定新安全标准**。

研究背景

■ 应对措施

□OpenAI、微软等七大科技巨头公司联合承诺开发针对生成内容的水印工具

OpenAI, Google, others pledge to watermark AI content for safety, White House says

□DeepMind推出SynthID工具以标记图像是否由Imagen生成，但无法应用于其他生成模型

研究背景

■ 溯源对象的演变

□传统的溯源方法是难以适应当前生成式大模型

2022年之前（空域、频域）

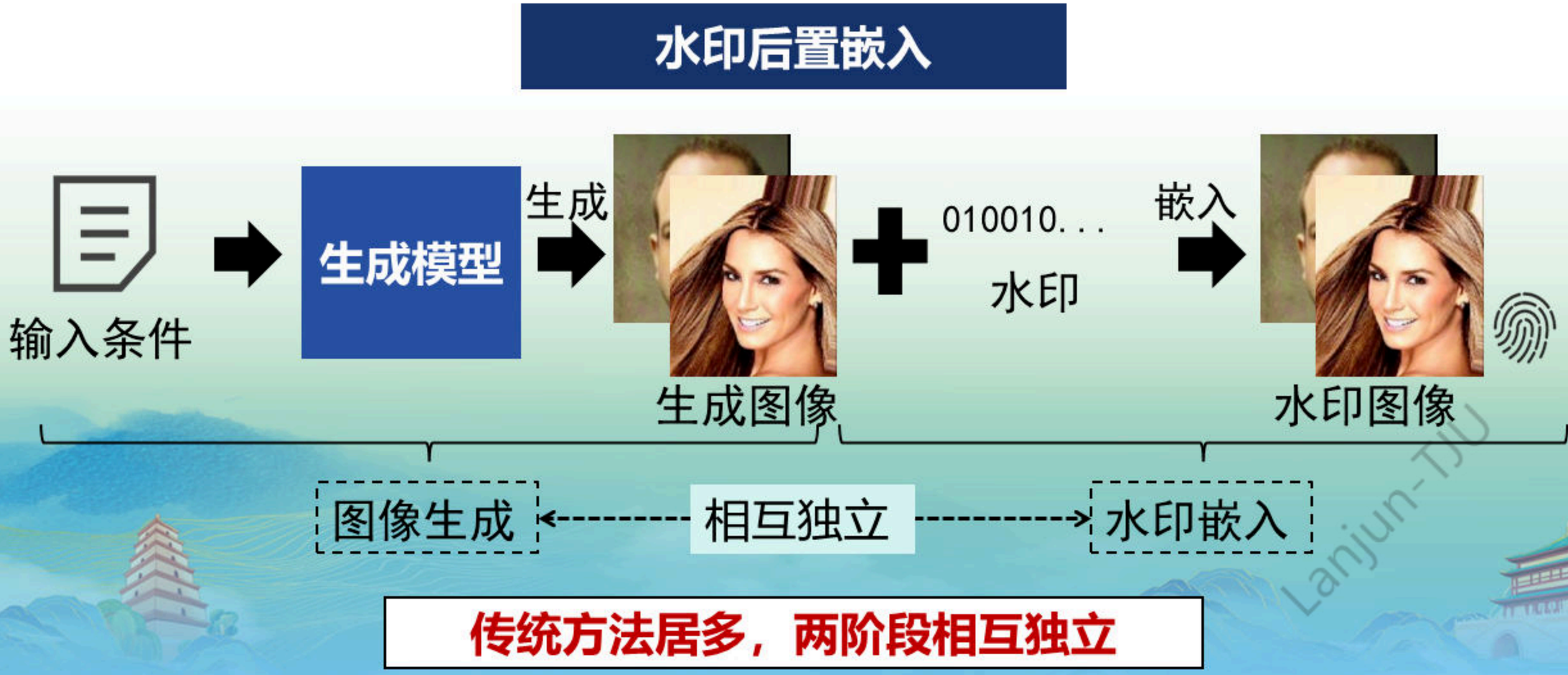
模式固定　　风险单一　　数据稀少

自2022年起（特征域）

模式多样　　风险增加　　数据庞杂

研究背景

■依据溯源信息嵌入的节点，将现有技术划分为水印后置嵌入、水印前置嵌入以及联合生成的溯源技术。



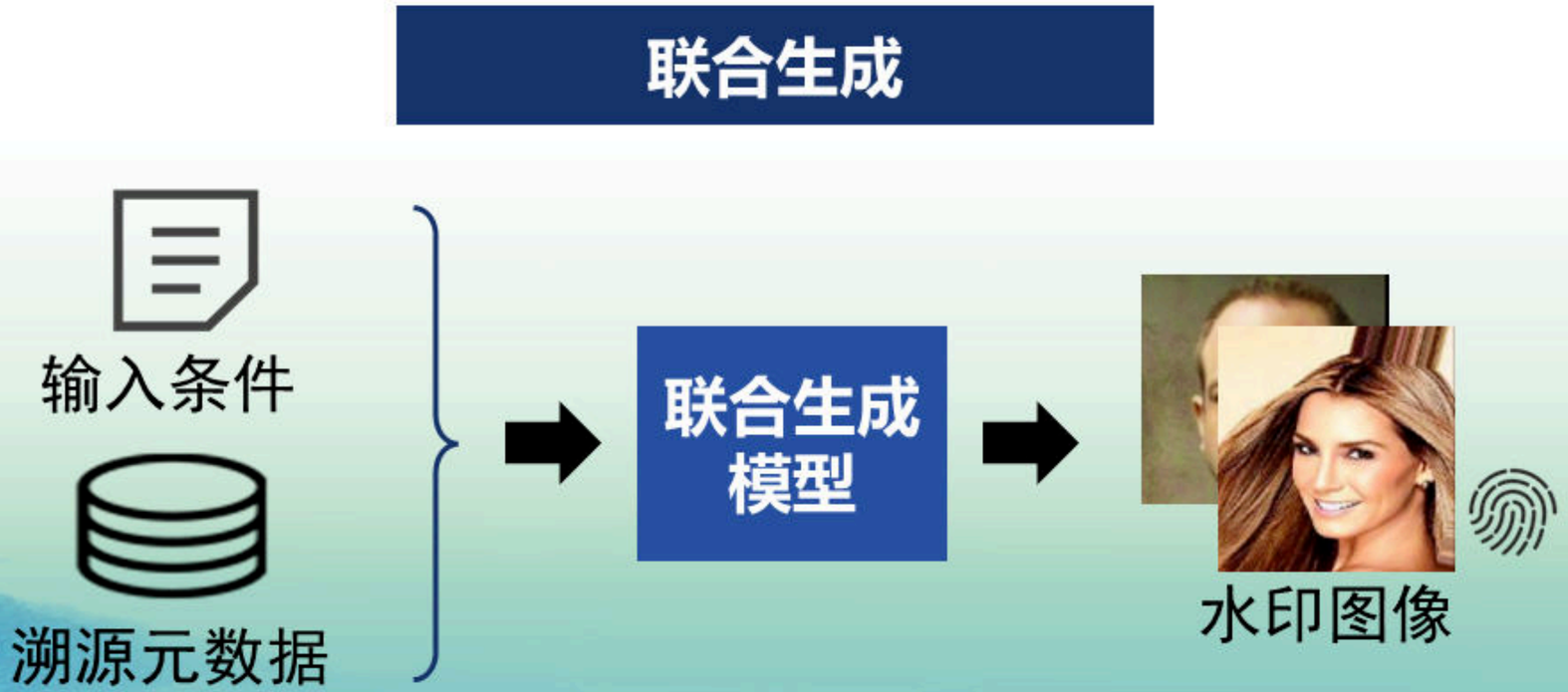
研究背景

■依据溯源信息嵌入的节点，将现有技术划分为<u>水印后置嵌入、水印前置嵌入以及联合生成</u>的溯源技术。

预先将水印信息嵌入到训练图像中

研究背景

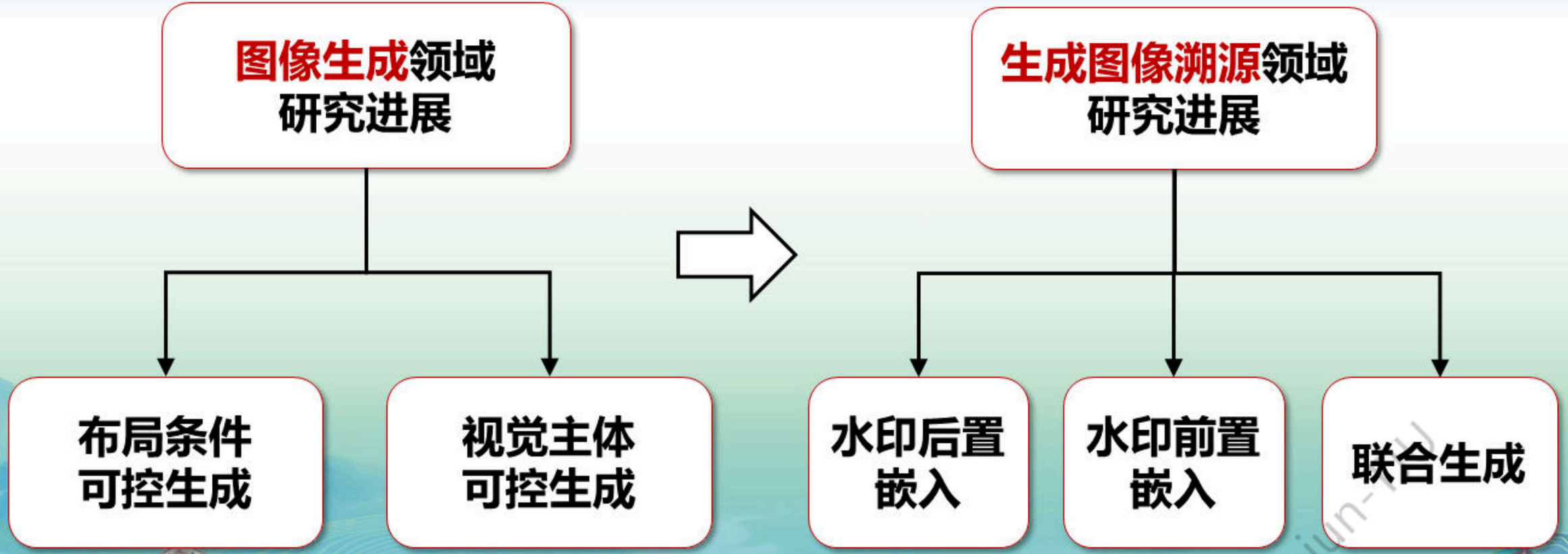
■依据溯源信息嵌入的节点，将现有技术划分为水印后置嵌入、水印前置嵌入以及联合生成的溯源技术。



联合生成模型同时处理输入条件和溯源信息

研究背景

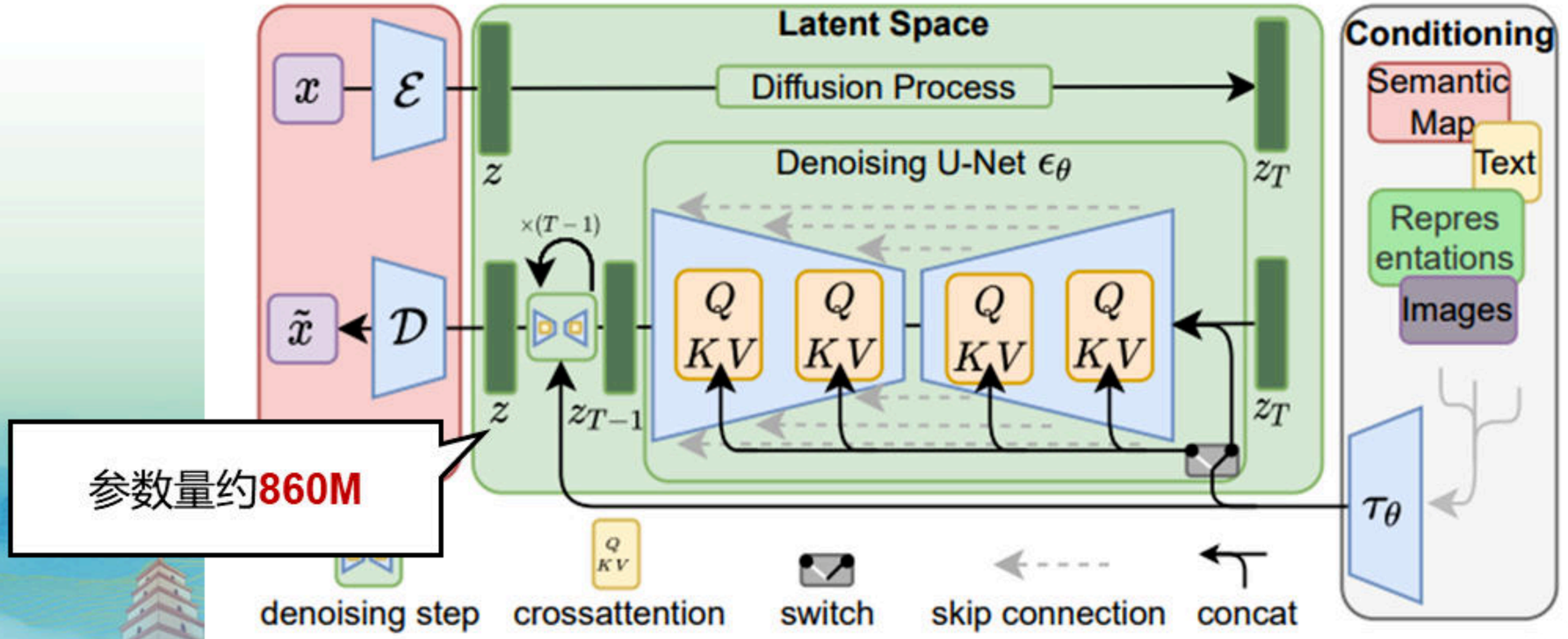
■报告框架



报告内容

图像生成领域研究进展

- 生成模型：扩散概率模型（Diffusion Model）
  - LDM：在**潜在特征空间**内进行扩散和去噪训练，提升训练稳定性。
  - 基于**交叉注意力**机制为图像生成引入各种生成条件。

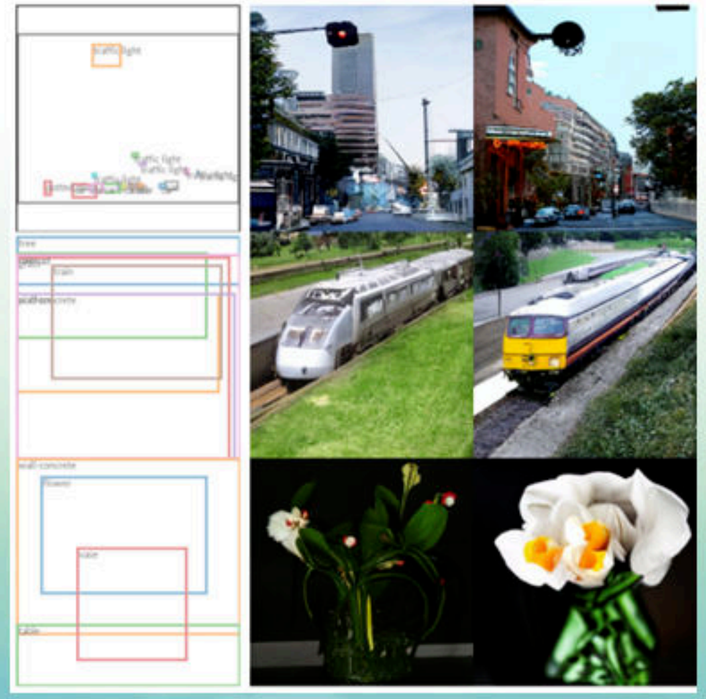


[1] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *NeurIPS*. 2021.

图像生成领域研究进展

- 生成模型：扩散概率模型（Diffusion Model）
  - LDM支持超分辨率、布局引导、图像修复（Inpaint）等多种任务。

超分辨率生成样例



布局引导生成样例

图像修复生成样例

[1] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *NeurIPS*. 2021.

图像生成领域研究进展

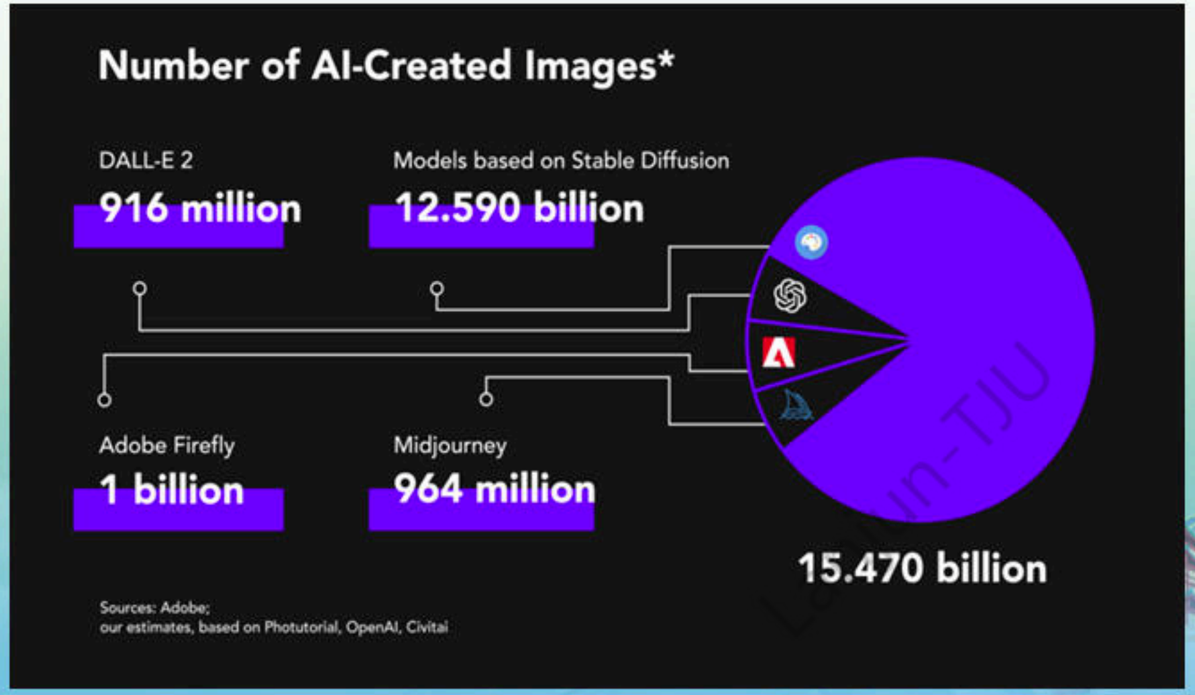
- 生成模型：扩散概率模型（Diffusion Model）
  - 基于LDM开发的Stable Diffusion成为最受欢迎的开源图像生成模型

由Stability AI发布开源预训练版本SD模型，并频繁进行版本更新。

Stable Diffusion

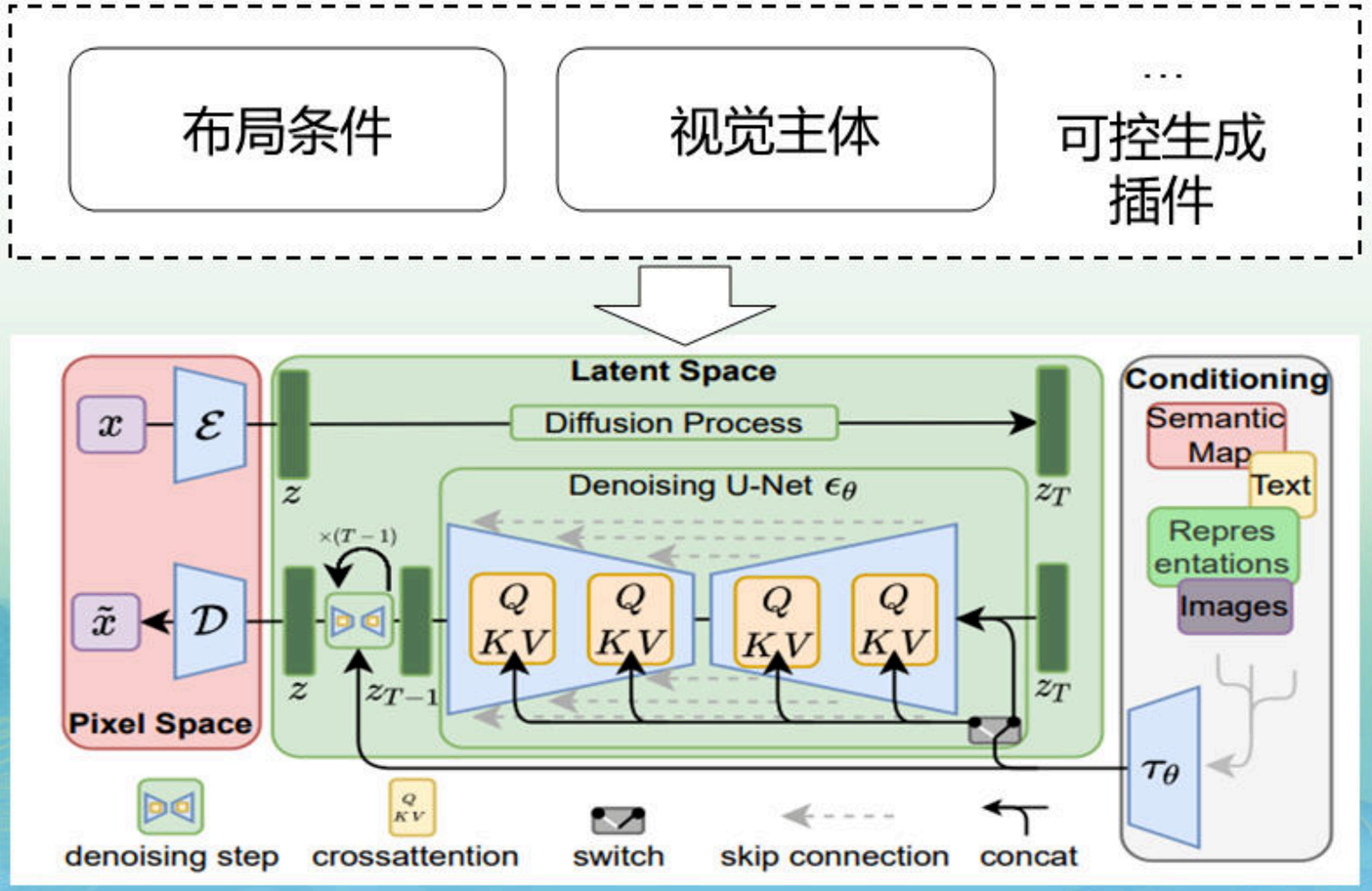
- Stable Diffusion V1.4——2022年8月
- Stable Diffusion V1.5——2022年10月
- Stable Diffusion V2.0——2022年11月
- Stable Diffusion V2.1——2022年12月
- Stable Diffusion XL 1.0——2023年7月
- Stable Diffusion v3.0——2024年4月

基于Stable Diffusion模型创造的图像达1259亿，占所有 AI 图像的80%。



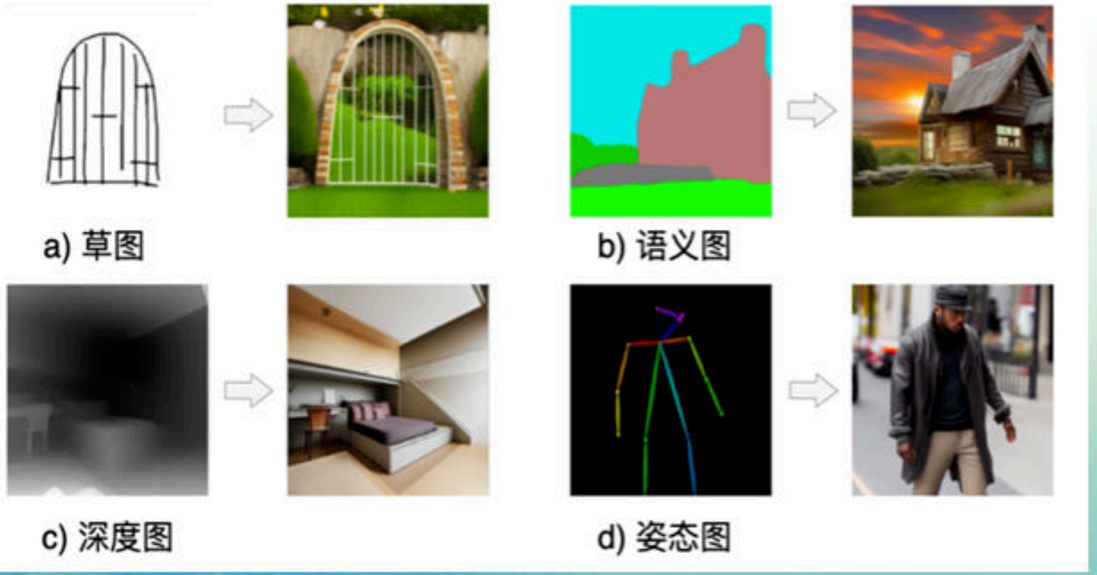
图像生成领域研究进展

- 生成模型：扩散概率模型（Diffusion Model）
  - 训练成本低、扩展性强的特点使SD成为可控生成研究的主流基座模型。



可控图像生成

- 可控图像生成技术分类
  - 基于布局条件：引入布局控制条件，实现画面空间布局可控。
  - 基于视觉主体：引导模型针对特定视觉主体进行内容创作。

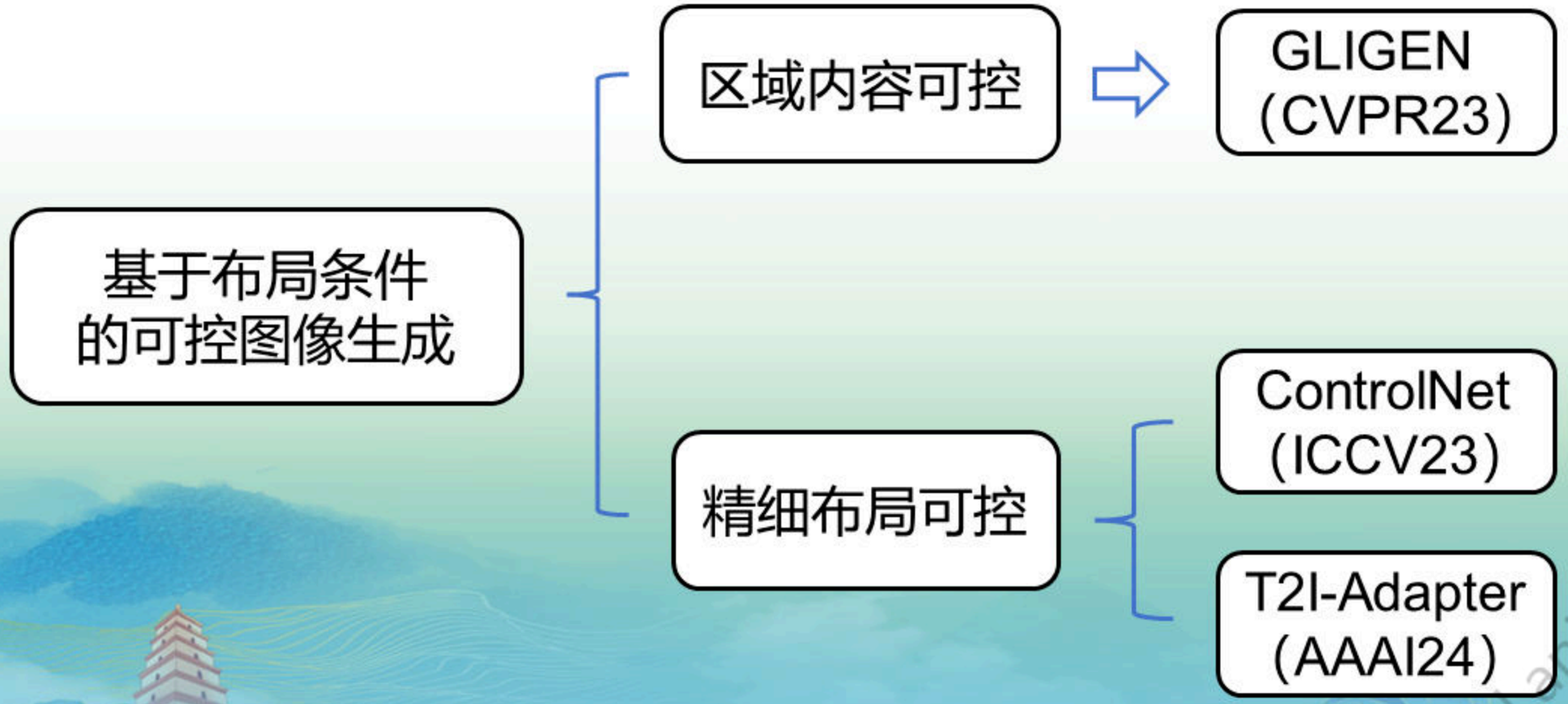


基于**布局条件**的可控图像生成

基于**视觉主体**的可控图像生成

可控图像生成

■ 基于布局条件的可控图像生成



基于布局条件的可控图像生成

- GLIGEN：区域内容可控生成
  - 在不修改原始模型参数的条件下，扩展文生图模型的区域内容可控能力
  - 区域内容可控：根据指定区域坐标和对应内容生成图像

输入描述：

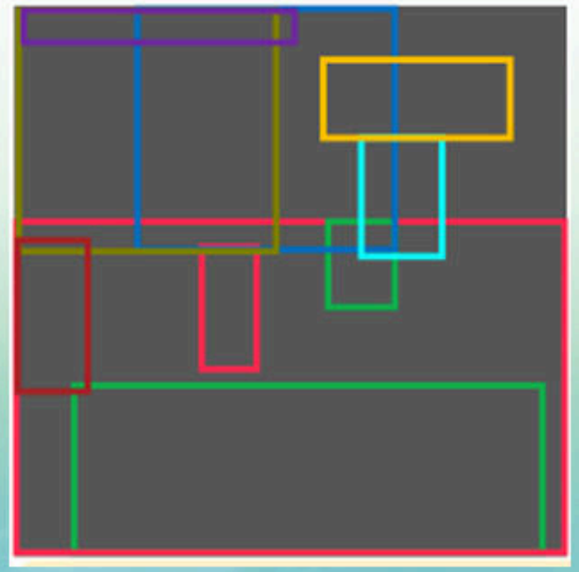
“一个女人在餐厅吃披萨”

控制条件：

红色框-桌子；绿色框-披萨；

蓝色框-人；紫色框-窗户；

黄色框-汽车……



布局框控制

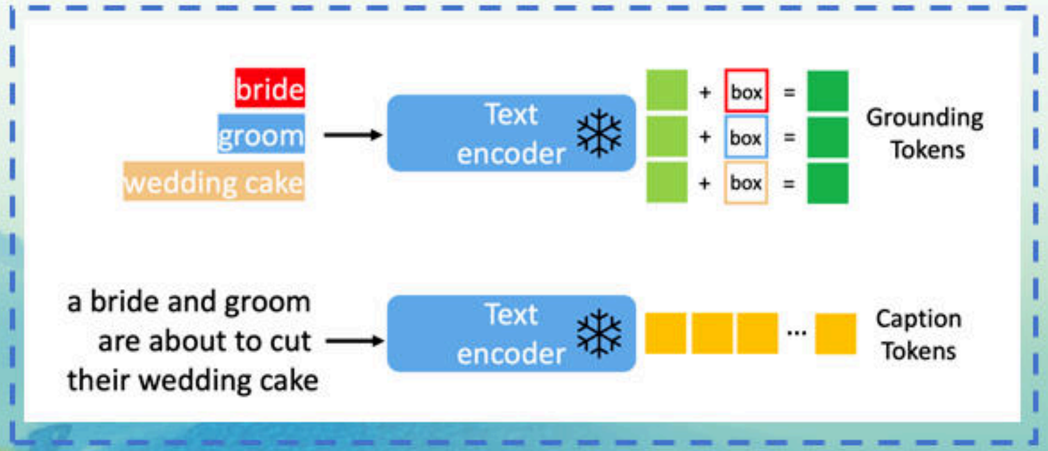
生成结果

[1] Li, Yuheng, et al. "Gligen: Open-set grounded text-to-image generation." *CVPR*. 2023.

基于布局条件的可控图像生成

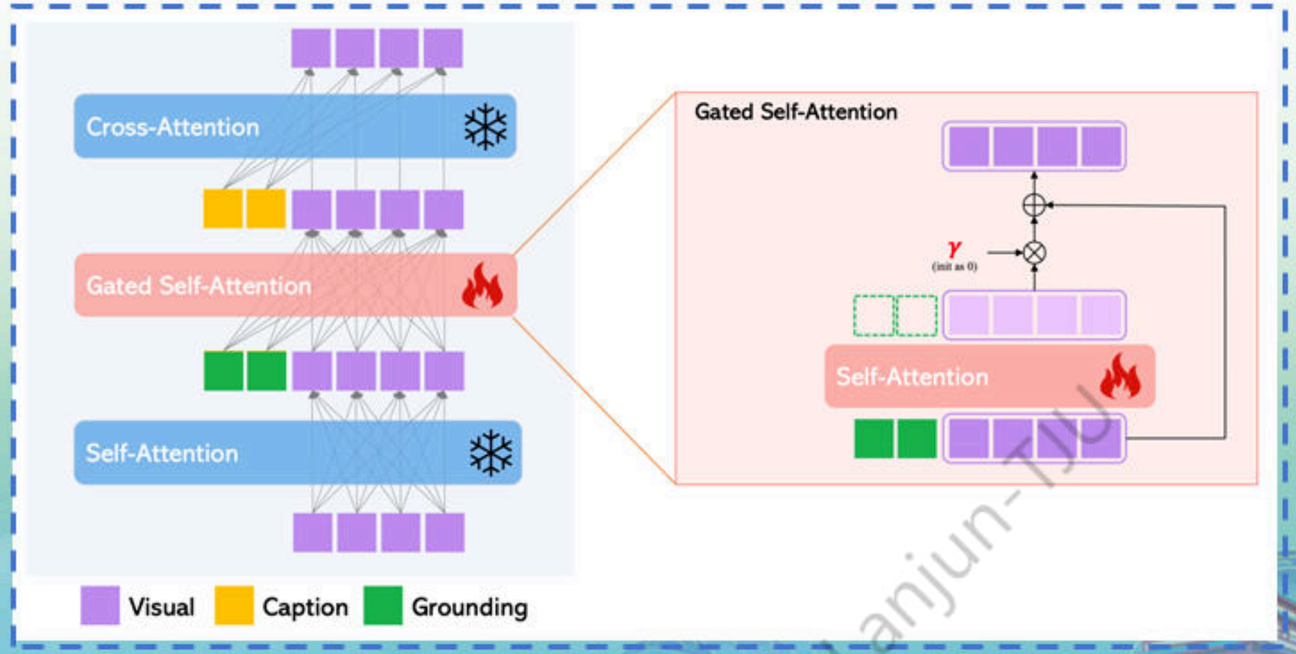
- GLIGEN：区域内容可控生成
  - 在不修改原始模型参数的条件下，扩展文生图模型的区域内容可控能力
  - 在预训练扩散模型（SD）中添加条件注入层，引入额外控制条件

1）生成额外条件编码



其中，控制布局的布局框（box）用MLP处理得到相应token

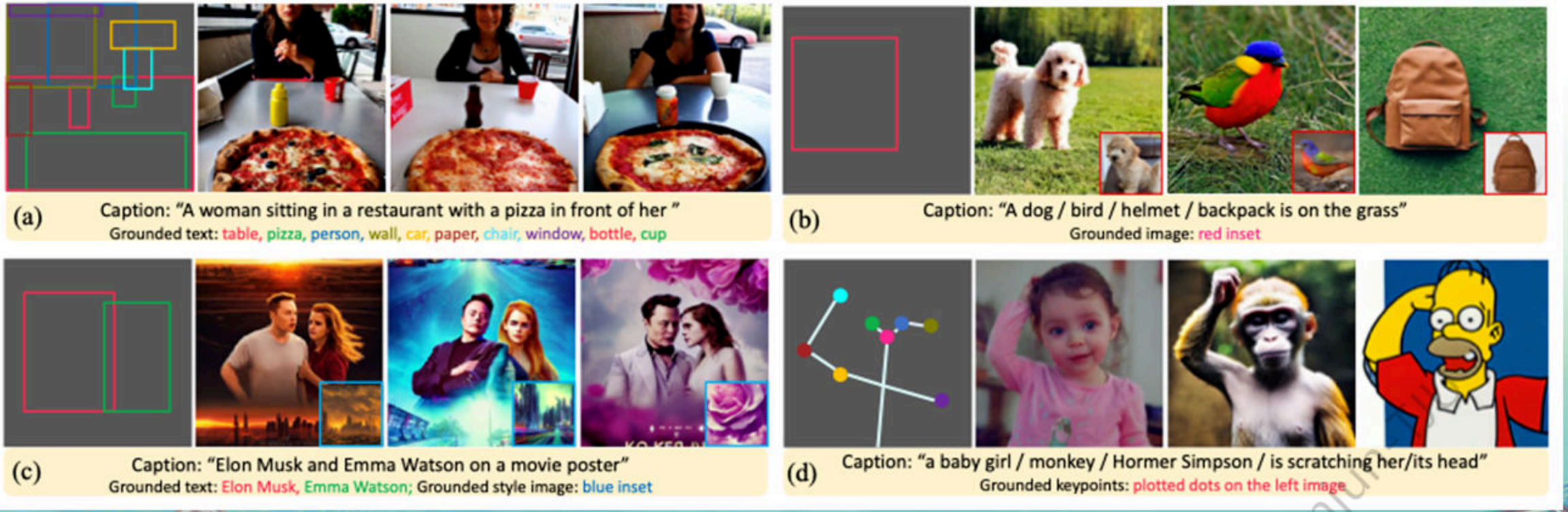
2）添加自注意力条件注入层



[1] Li, Yuheng, et al. "Gligen: Open-set grounded text-to-image generation." *CVPR*. 2023.

基于布局条件的可控图像生成

- GLIGEN：区域内容可控生成
  - 支持使用布局框、关键点、参考图像等条件来控制文本到图像的生成。

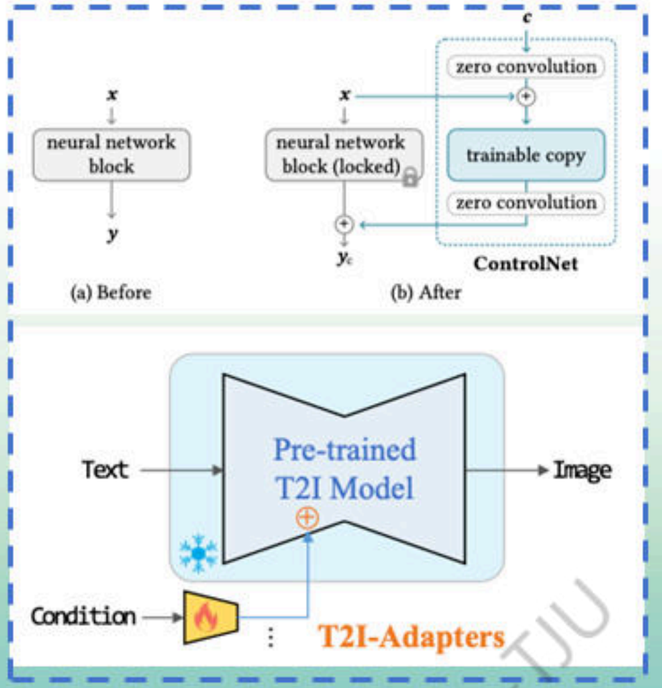


[1] Li, Yuheng, et al. "Gligen: Open-set grounded text-to-image generation." *CVPR*. 2023.

基于布局条件的可控图像生成

- ControlNet、T2I-Adapter：精细化布局控制
  - 通过额外的**布局控制网络**，提供**即插即用**的**精细化布局控制**。

实现引入各种额外附加控制条件的图像生成



ControlNet与T2I-Adapter
通过额外的网络引入新控制条件

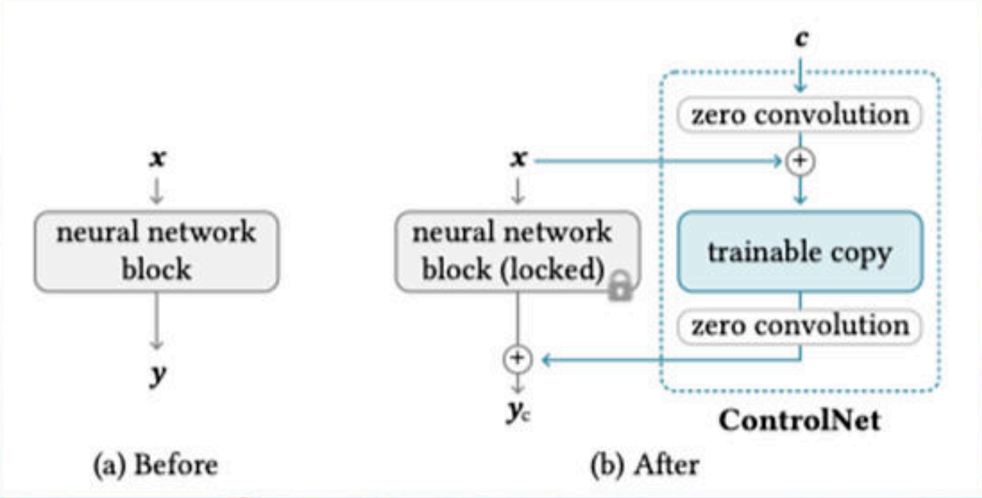
[1] Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models." *AAAI*. 2024.
[2] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *ICCV*. 2023.

基于布局条件的可控图像生成

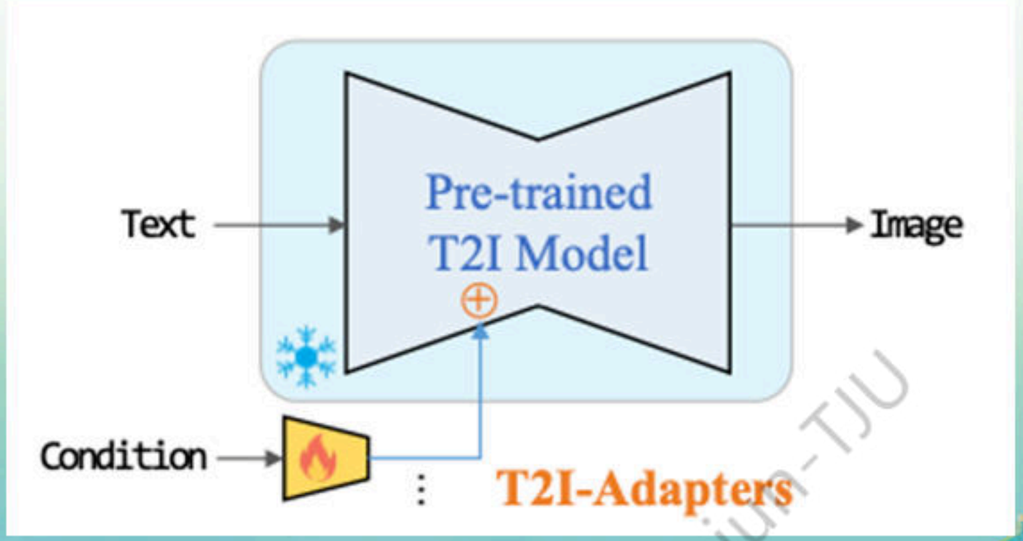
■ControlNet：使用去噪U-Net**编码器部分的拷贝**引入新控制条件（利用预训练模型对视觉特征空间的理解能力，提升训练稳定性）。

■T2I-Adapter：训练额外的**小型Adapter网络**引入新控制条件（将空间布局条件与预训练特征空间对应，从而实现布局控制）

ControlNet模型示意图



T2I-Adapter模型示意图



参数量：视去噪U-Net网络规模（~7亿参数）

参数量：77~79M

[1] Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models." *AAAI*. 2024.
[2] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *ICCV*. 2023.

基于布局条件的可控图像生成

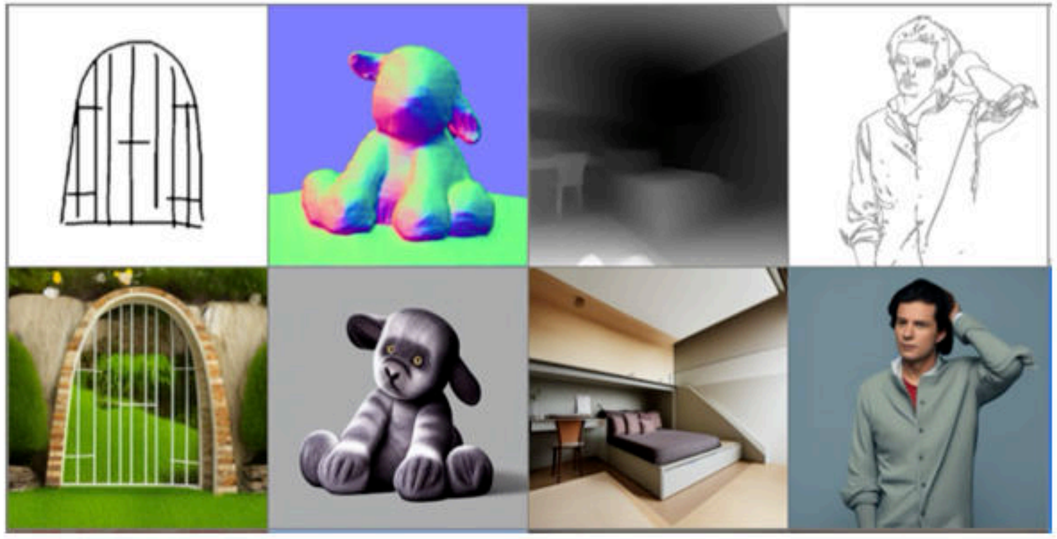
■ControlNet：使用去噪U-Net**编码器部分的拷贝**引入新控制条件（利用预训练模型对视觉特征空间的理解能力，提升训练稳定性）。

■T2I-Adapter：训练额外的**小型Adapter网络**引入新控制条件（将空间布局条件与预训练特征空间对应，从而实现布局控制）

ControlNet模型生成样例

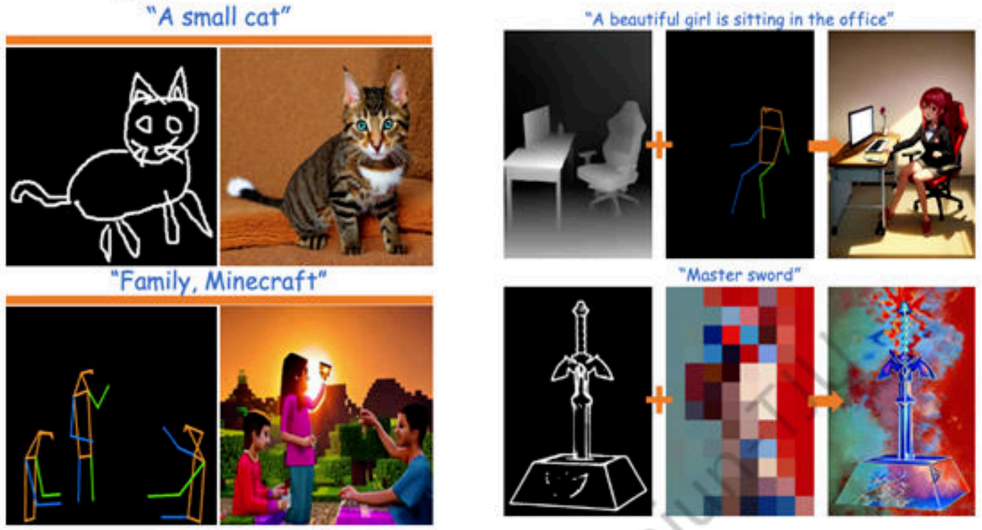
控制条件

生成结果



画面布局**控制更加精细**

T2I-Adapter模型生成样例

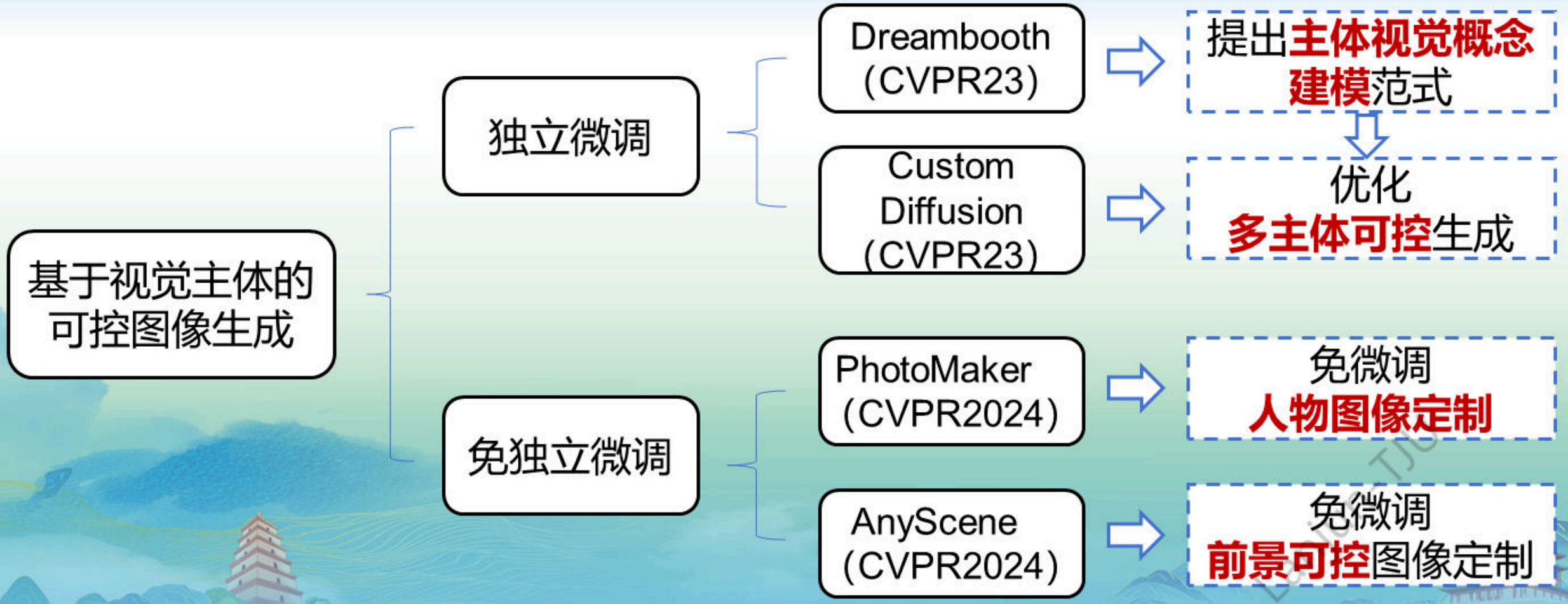


单控制条件　　组合控制条件

网络**轻量化**，支持**多条件组合**

可控图像生成

■ 基于视觉主体的可控图像生成



基于视觉主体的可控图像生成

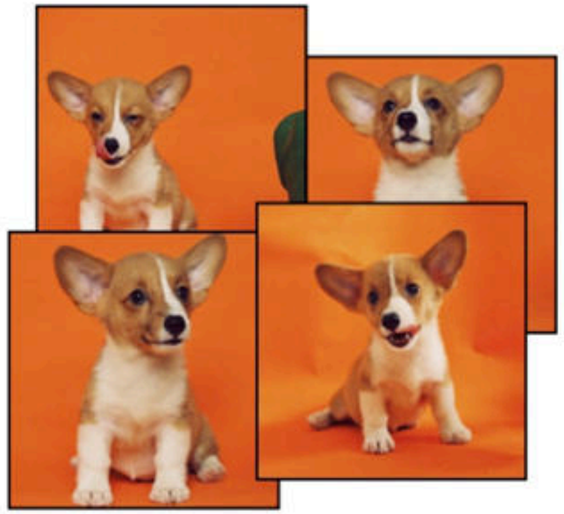
- Dreambooth：围绕特定视觉主体进行定制图像生成
  - 根据特定物体学习视觉概念，结合文本条件生成定制化场景。

多张主体图片 → 学习视觉概念S*（微调模型） → 用视觉概念生成多样的场景

[1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *CVPR*. 2023.

基于视觉主体的可控图像生成

- Dreambooth：围绕特定视觉主体进行定制图像生成
  - 根据特定物体学习视觉概念，结合文本条件生成定制化场景。



【主体】（S*）图片

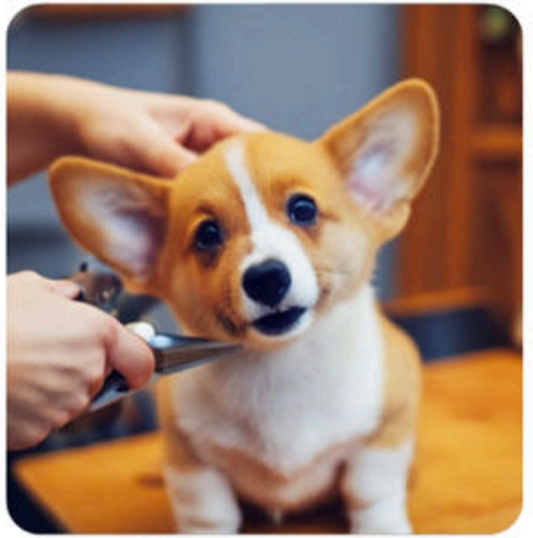
S* 在雅典卫城

S* 在游泳

S* 在睡觉

S* 在狗屋中

S* 在水桶中



S* 正在剪发

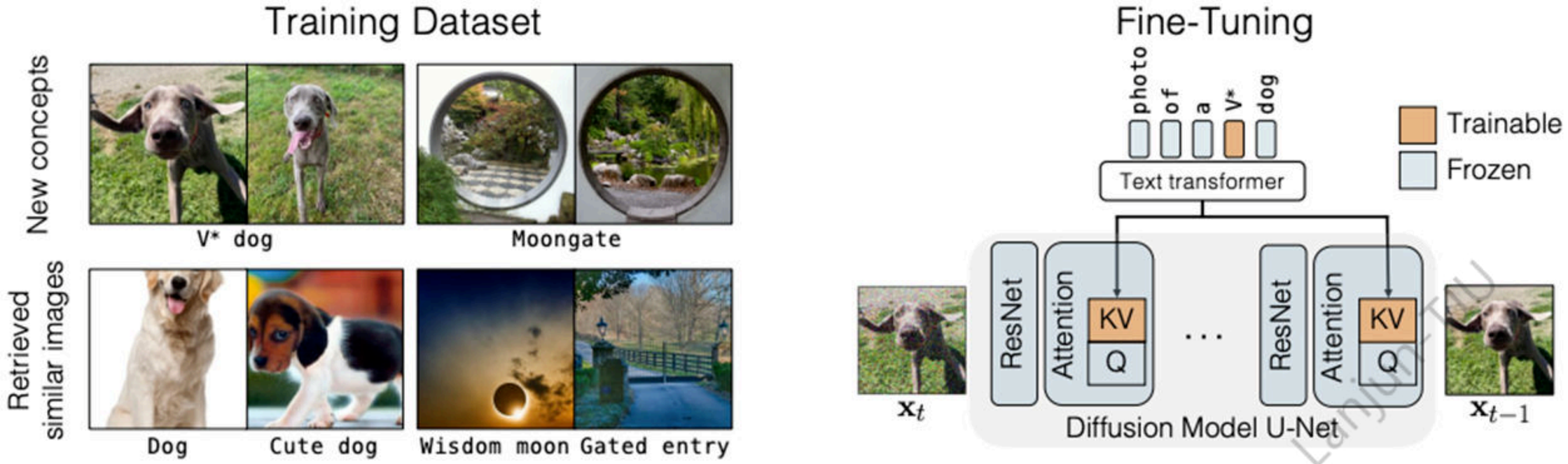
应用限制

需要针对每个视觉主体进行独立的模型微调，**微调时间长**（约30分钟）

不同的视觉概念会相互影响，**无法支持多实体组合**生成。

基于视觉主体的可控图像生成

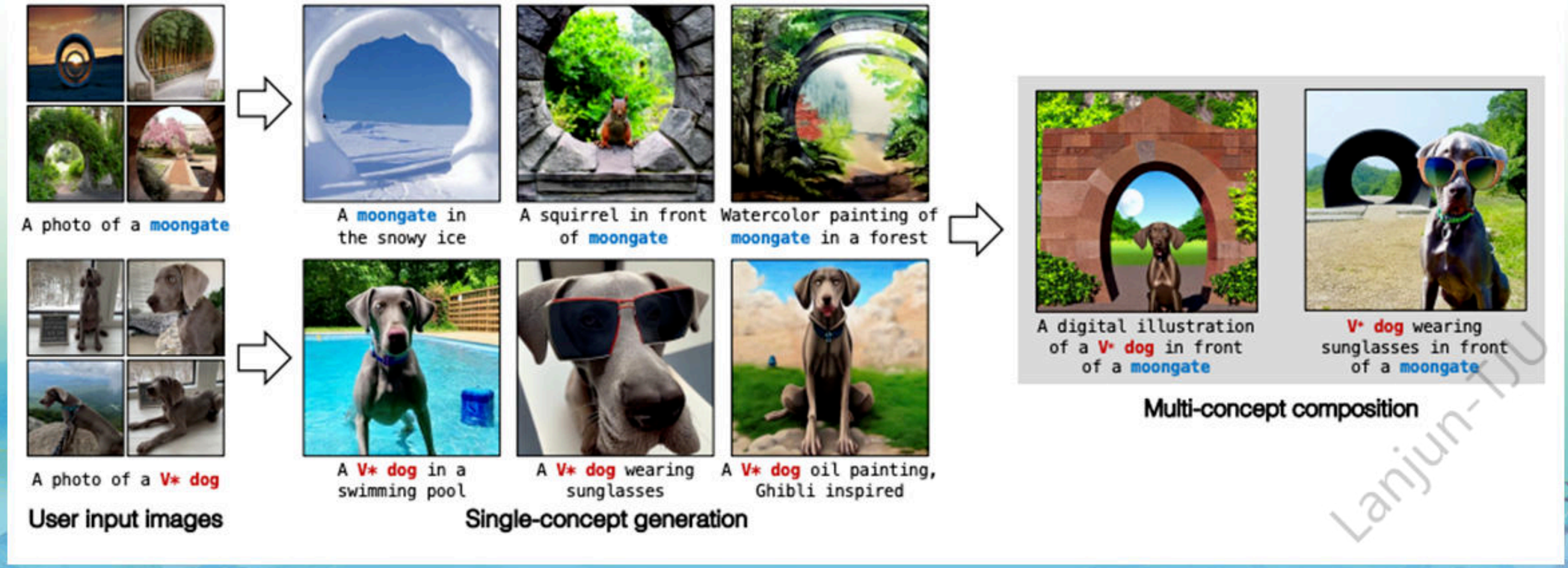
- CustomDiffusion：优化多主体组合能力的主体驱动生成
  - 根据给出图片素材，检索相关图像作为训练数据（提升训练稳定性）。
  - 微调去噪网络中部分“键”-“值”映射矩阵进行视觉概念学习（降低微调参数量）。



[1] Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." *CVPR*. 2023.

基于视觉主体的可控图像生成

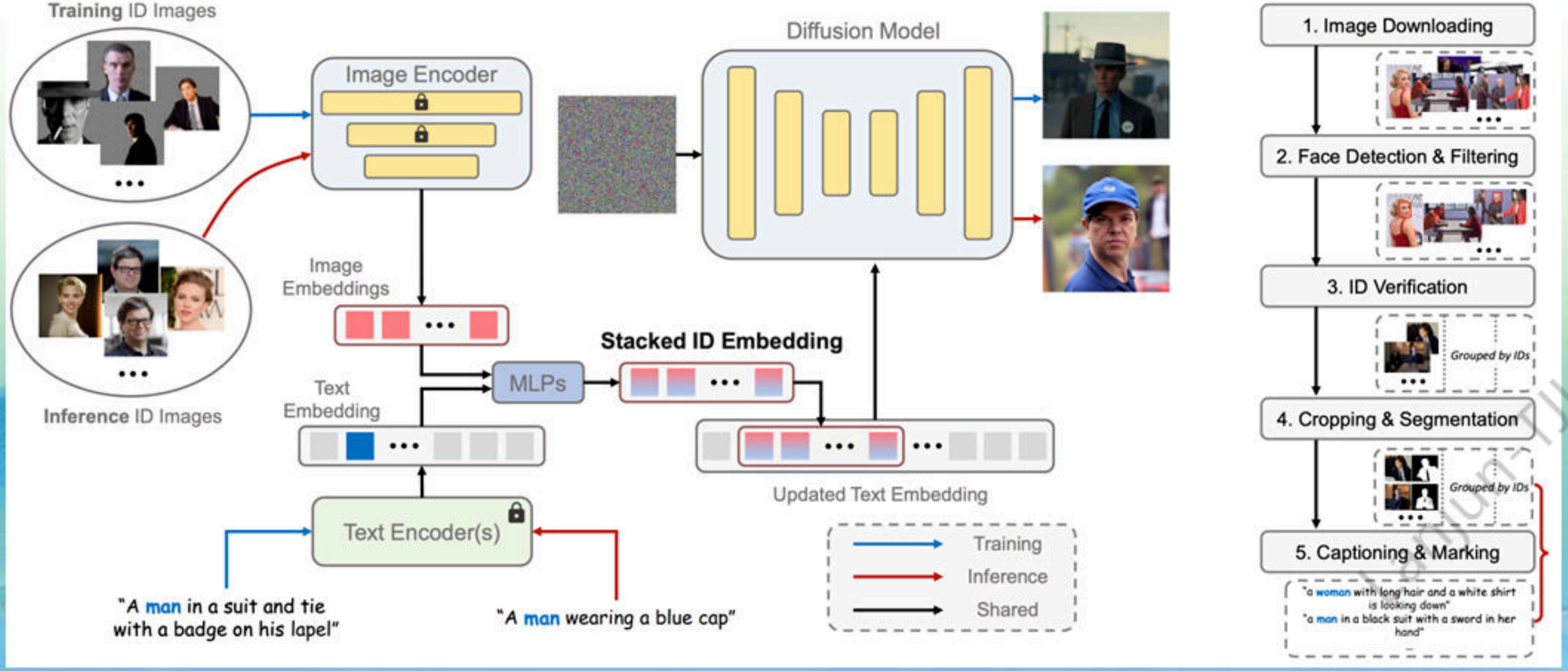
- CustomDiffusion：优化多主体组合能力的主体驱动生成
  - 通过**对齐训练**合并单主体微调的“键”-“值”映射矩阵，实现**多主体组合生成**。
  - 将模型微调时间缩短至约10分钟（相比Dreambooth提升三倍）



[1] Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." *CVPR*. 2023.

基于视觉主体的可控图像生成

- PhotoMaker：免微调人物定制图像生成
  - 收集人物图像训练数据（13000名人物，12万张图片）进行模型微调，同时微调CLIP图像编码器从多张人物图像提取面部细节特征。
  - 推理阶段：融合人物ID表征与输入文本特征，形成堆栈人物表征引导生成。



[1] Zhen Li, et al. "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding." *CVPR*. 2024.

基于视觉主体的可控图像生成

■ PhotoMaker：免微调人物定制图像生成

□实现免微调人脸定制图像生成（10秒左右，相比Dreambooth提速130倍）

[1] Zhen Li, et al. "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding." *CVPR*. 2024.

基于视觉主体的可控图像生成

■ AnyScene：前景可控的定制图像生成

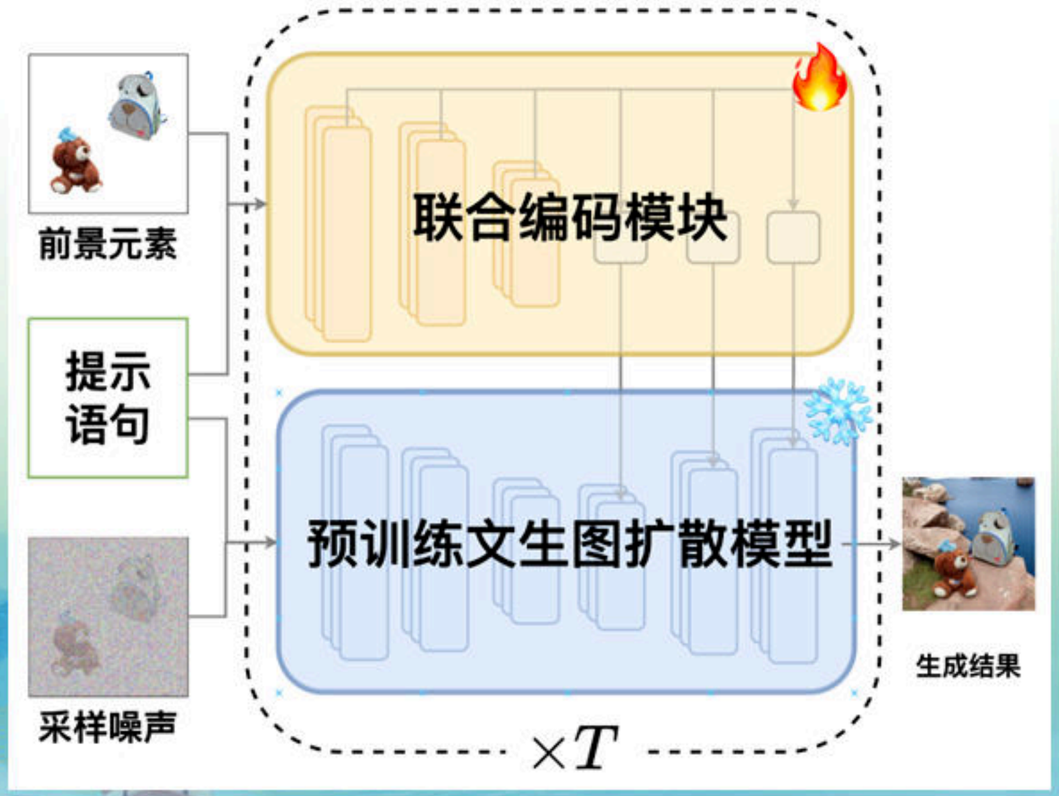
□基于前景元素和定制提示语句进行多样化定制图像生成

[1] Ruidong Chen, et al. ”AnyScene: Customized Image Synthesize with composited foreground.” *CVPR*. 2024.

基于视觉主体的可控图像生成

■ AnyScene：前景可控的定制图像生成

□模型训练：训练联合编码模块，使生成模型具备**理解前景信息**的能力，**从而进行整体场景生成**。



训练目标

■ 扩散损失：保障全局视觉和谐性

$$L_{denoising} = ||\epsilon - \epsilon_\theta(z_t, t, y, c_f)||_2^2$$

■ 内容损失：像素级对齐，提升前景色彩/纹理保真度

$$L_{content} = ||x_0 - \hat{x}_t||_2^2 \odot M$$

■ 边缘损失：模拟梯度变化，提升前景边缘视觉质量

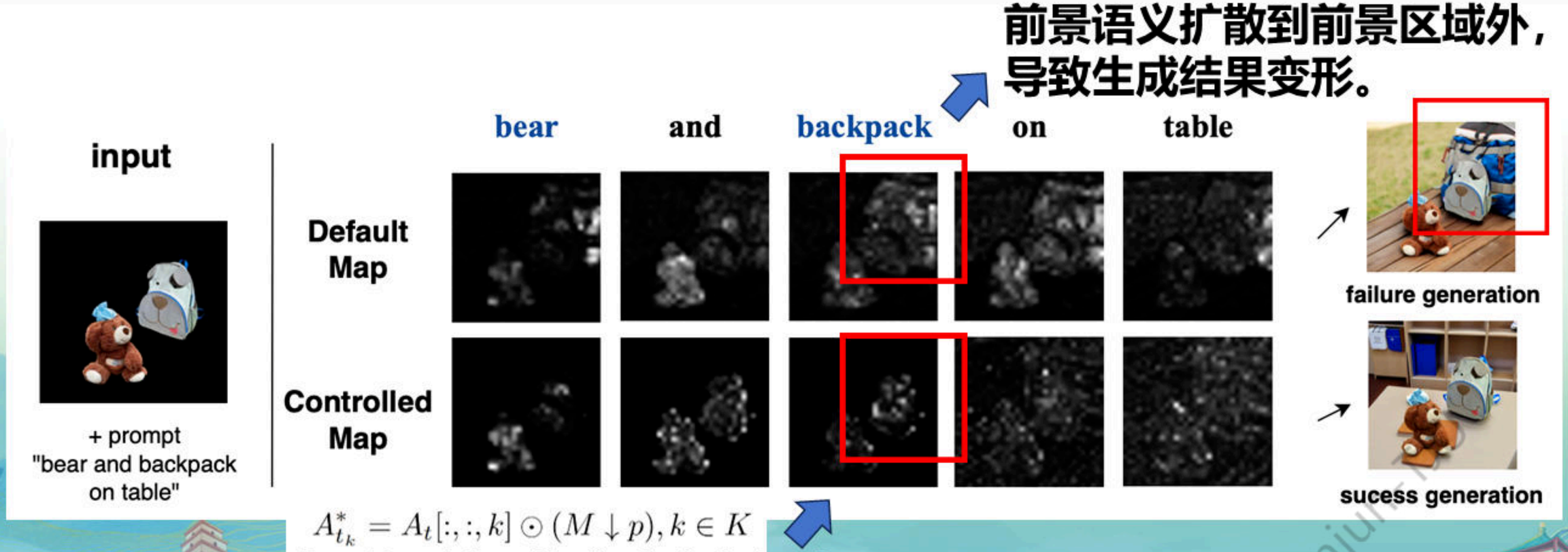
$$L_{gradient} = ||\nabla x_0 - \nabla \hat{x}_t||_2^2 \odot M_{edge}$$

[1] Ruidong Chen, et al. ”AnyScene: Customized Image Synthesize with composited foreground.” *CVPR*. 2024.

基于视觉主体的可控图像生成

■ AnyScene：前景可控的定制图像生成

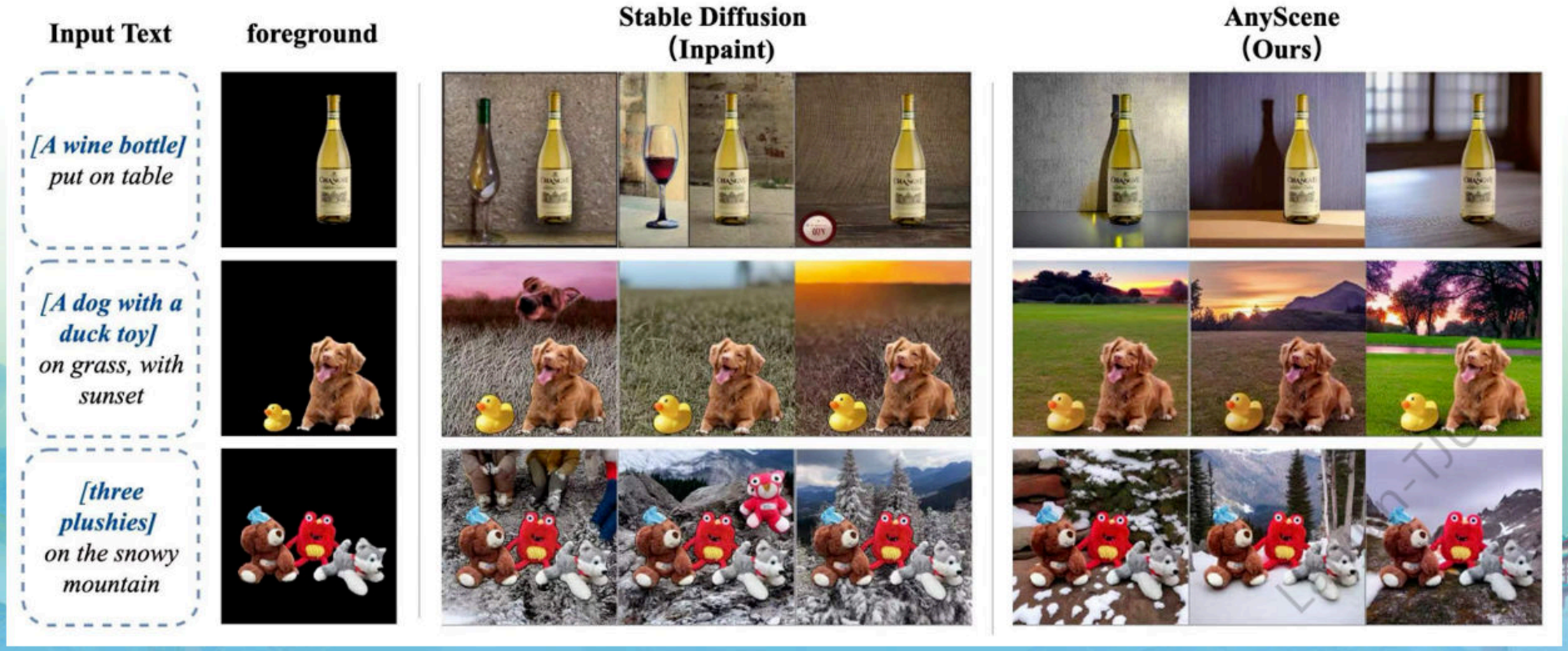
□模型推理：调节注意力矩阵，避免前景元素发生形变。



[1] Ruidong Chen, et al. "AnyScene: Customized Image Synthesize with composited foreground." *CVPR*. 2024.

基于视觉主体的可控图像生成

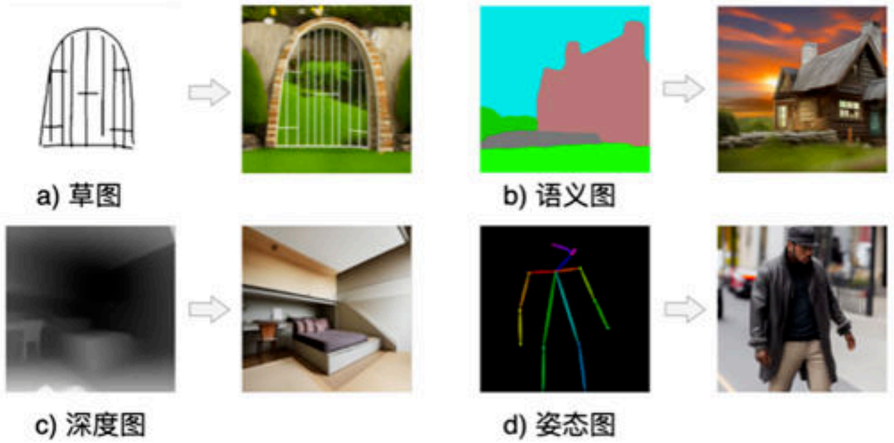
- AnyScene：前景可控的定制图像生成
  - 生成场景和谐性、准确性相比SD（Inpaint）优势明显。

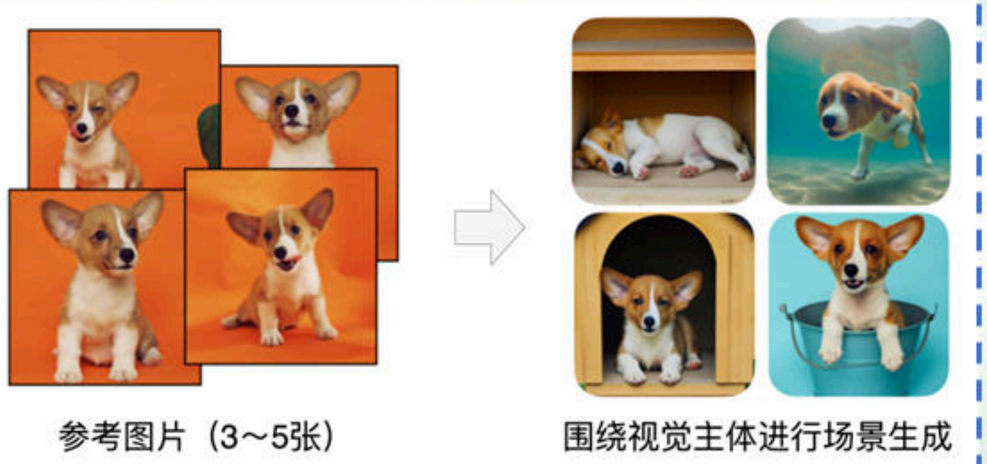


[1] Ruidong Chen, et al. ”AnyScene: Customized Image Synthesize with composited foreground.” *CVPR*. 2024.

图像生成领域研究进展

■ 总结





基于**布局条件**的可控图像生成

- 区域内容控制
- 精细布局控制

基于**视觉主体**的可控图像生成

- 独立微调（任何视觉概念）
- 无需微调（特定生成任务）

图像生成领域研究进展

- 展望
  - “布局”、“主体内容”等较为直观的“可控”研究已经趋向成熟。
  - 仍待研究适用于**具体任务、具体需求**的“可控生成”

2024—未来可控生成研究

- 生成模型能力边界的进一步探索
  - 开放主题+无需微调的主体驱动生成
  - 丰富视觉上下文的连续场景生成
  - …
- 适用于具体需求的可控性研究
  - AI设计稿分图层生成
  - 3D信息引导的高精度连续动作生成
  - …

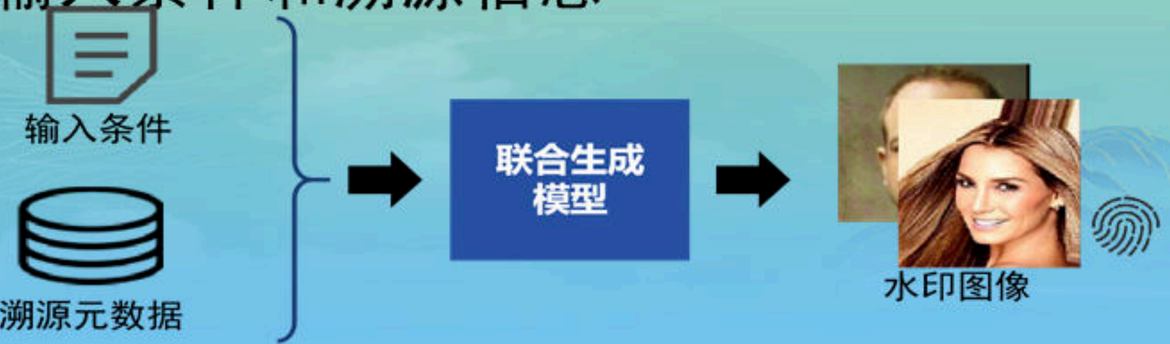
报告内容

溯源方法

■水印后置嵌入：传统方法居多，两阶段相互独立

■水印前置嵌入：预先将水印信息嵌入到训练图像中

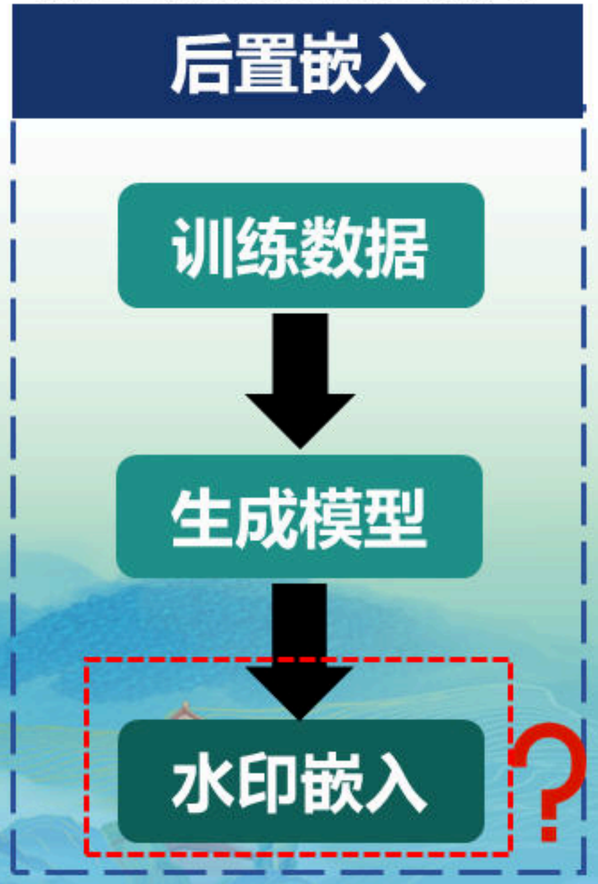
■联合生成：同时处理输入条件和溯源信息

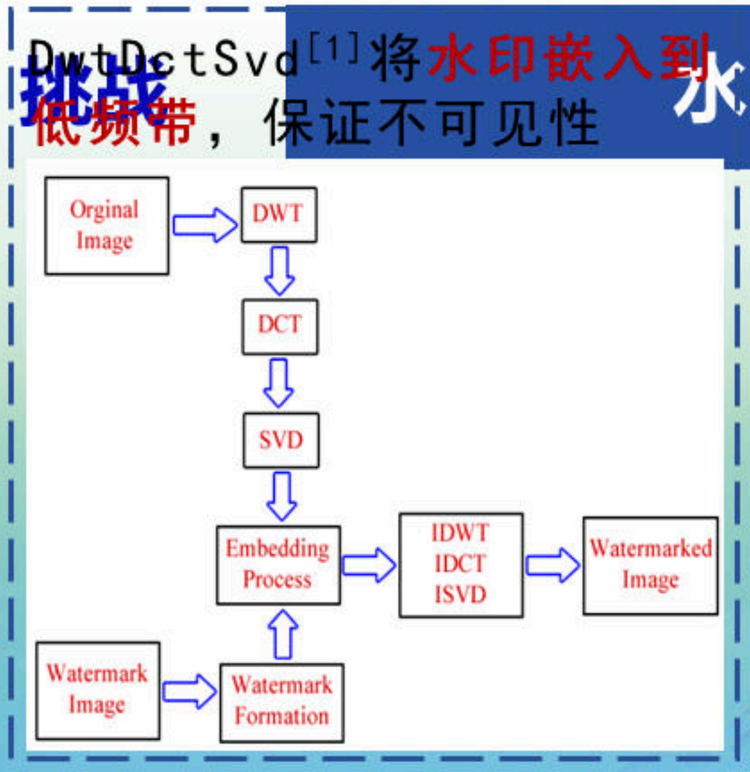


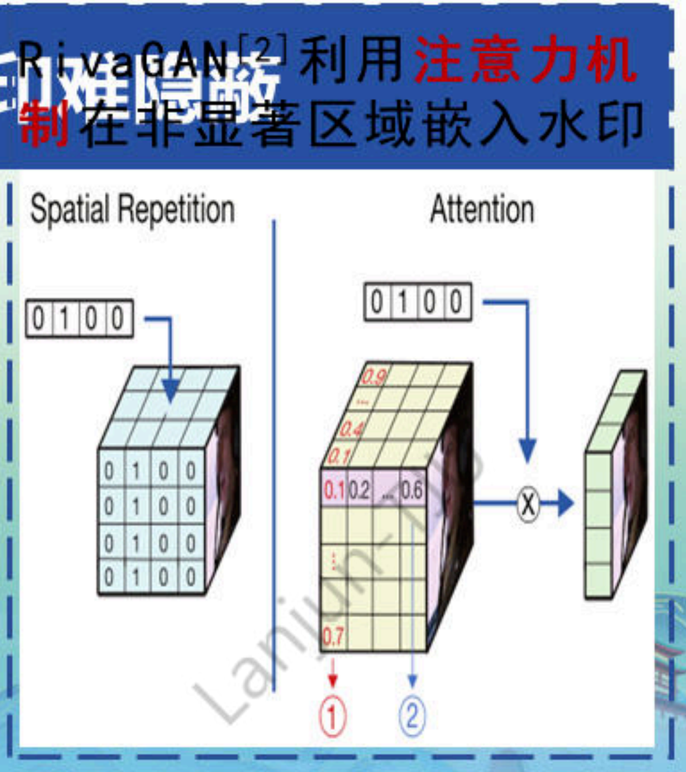
水印后置嵌入的生成图像溯源

■任务挑战

□水印后置嵌入方法，分为图像生成和水印嵌入两阶段，水印嵌入是在图像生成之后进行。





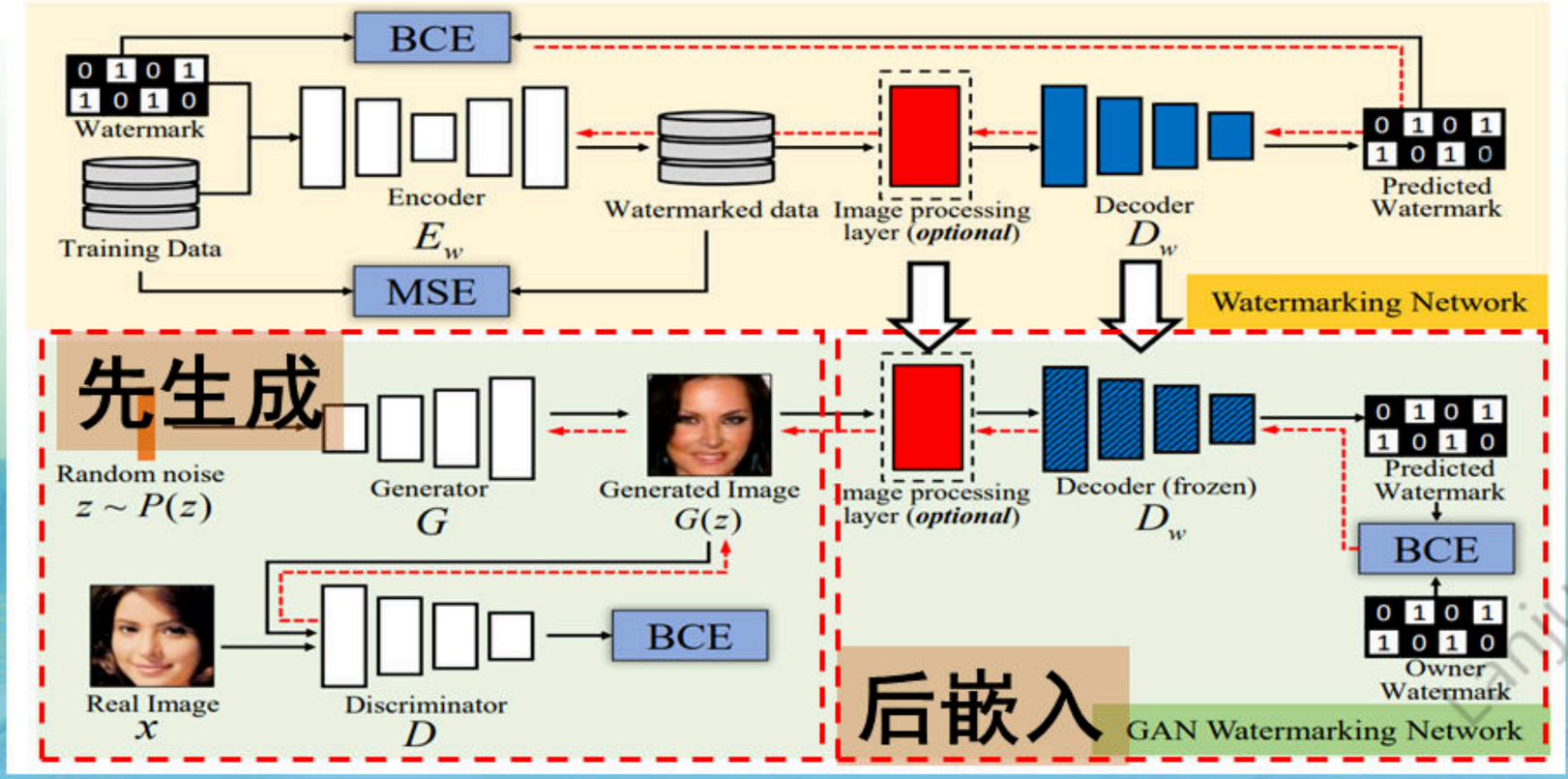


[1] Ingemar Cox, et al. ”Digital watermarking and steganography.” *Morgan kaufmann*. 2007.
[2] Kevin Alex Zhang, et al. ”Robust invisible video watermarking with attention.” *arXiv*. 2019

水印后置嵌入的生成图像溯源

■研究方法

□[1]将水印图像的生成过程分为图像生成和水印嵌入两阶段，利用**预训练水印解码器**为图像加水印，用以溯源生成图像归属。



[1] Jianwei Fei, et al. "Supervised GAN Watermarking for Intellectual Property Protection." *WIFS*. 2022.

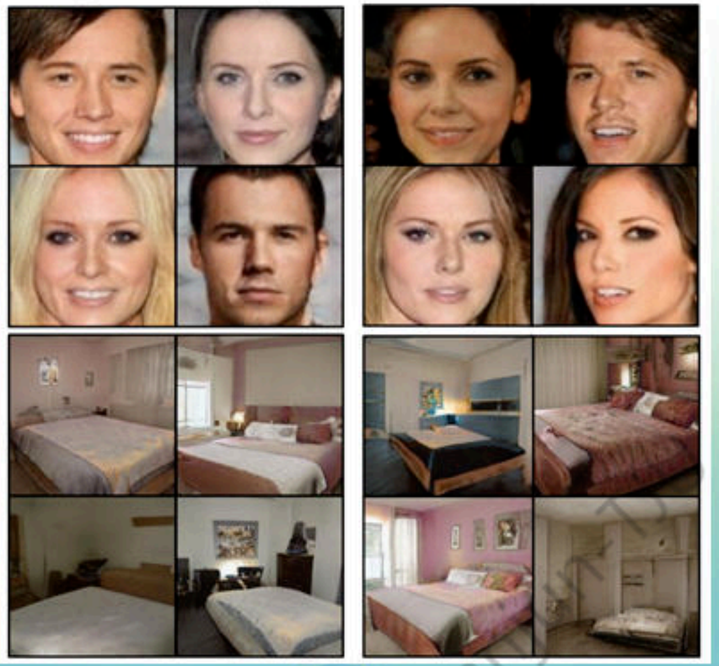
水印后置嵌入的生成图像溯源

■实验结果

- 在多个数据集达到100%的水印恢复准确率，且对图像质量影响较小

Dataset	MSE	Bit Acc	PSNR	SSIM(%)
CelebA	1.2e-5	100.00	45.31	99.48
LSUN-bedroom	5.0e-6	100.00	48.59	99.69
Flowers	5.0e-6	100.00	47.60	99.53

- 可视化结果表明，本方法能够确保生成图像嵌入水印后具有高质量



无水印图像　水印图像

[1] Jianwei Fei, et al. ”Supervised GAN Watermarking for Intellectual Property Protection.” *WIFS*. 2022.

水印后置嵌入的生成图像溯源

■总结

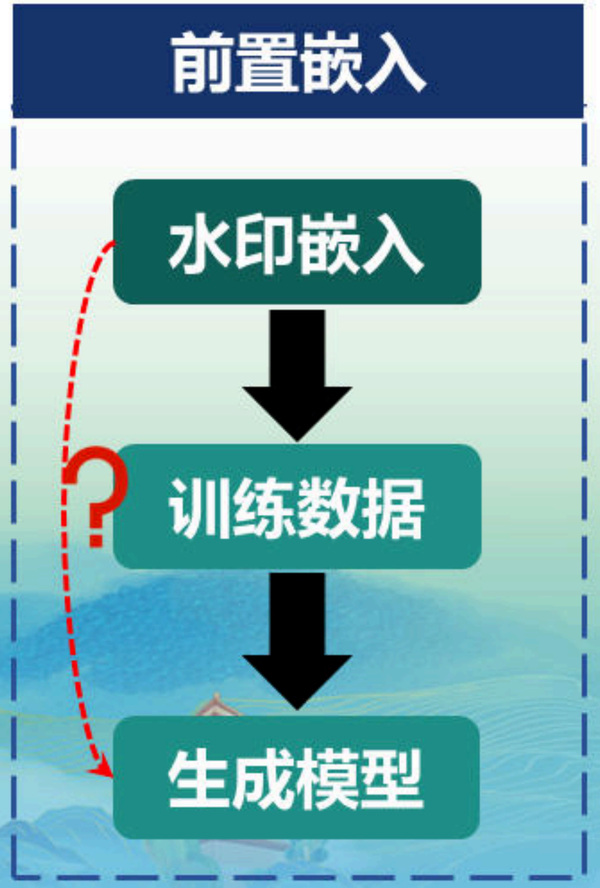
- 优势：生成图像和嵌入水印之间相互独立，**避免了两过程相互干扰**。

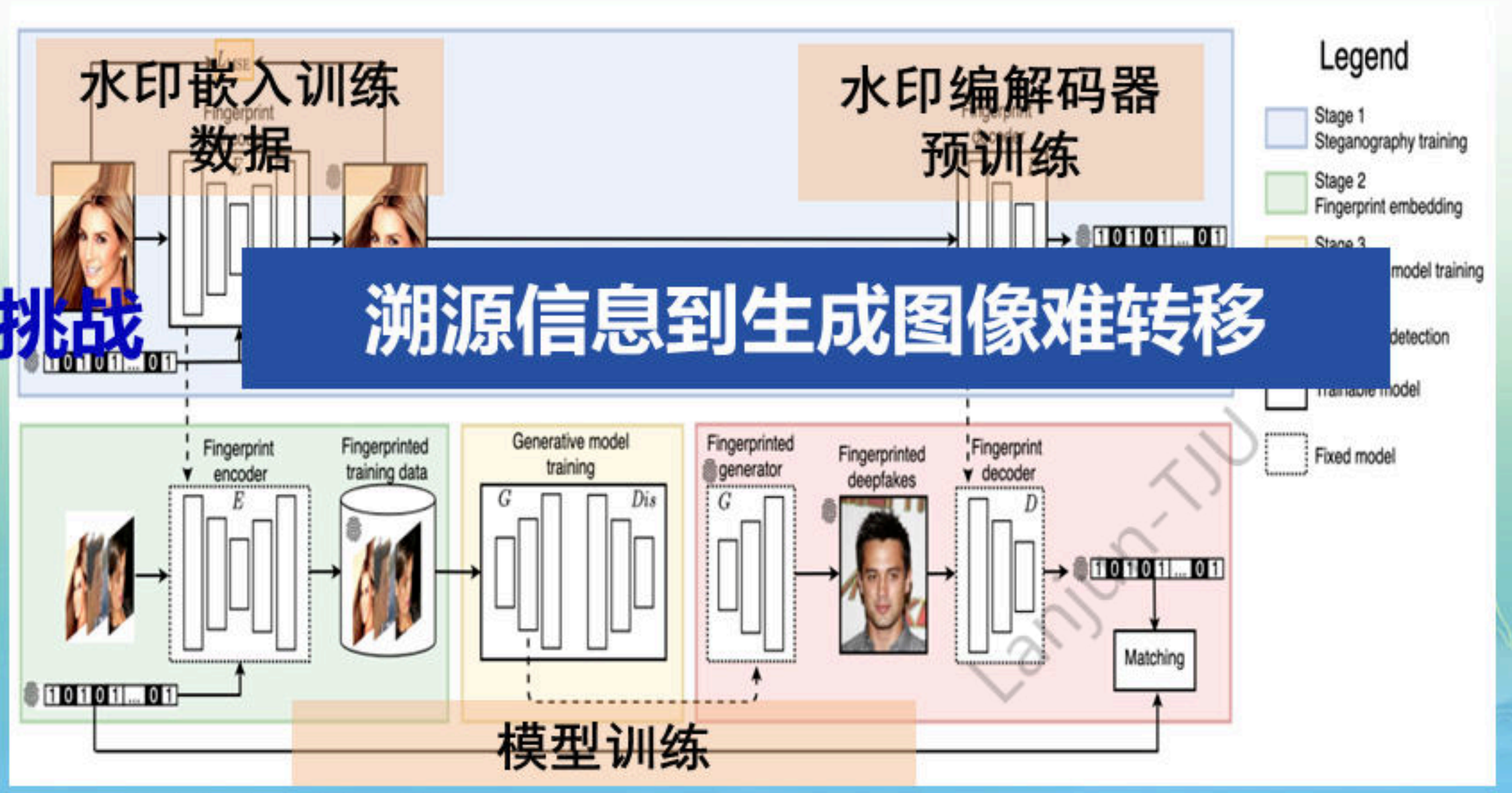
- 缺陷：在两阶段之间存在**信息篡改**和**逃逸攻击**等安全性问题。

水印前置嵌入的生成图像溯源

■任务挑战

□水印前置嵌入方法，首先将水印嵌入到训练数据中，再利用嵌入水印的数据来训练生成模型，使得生成图像也具备相同的水印信息。



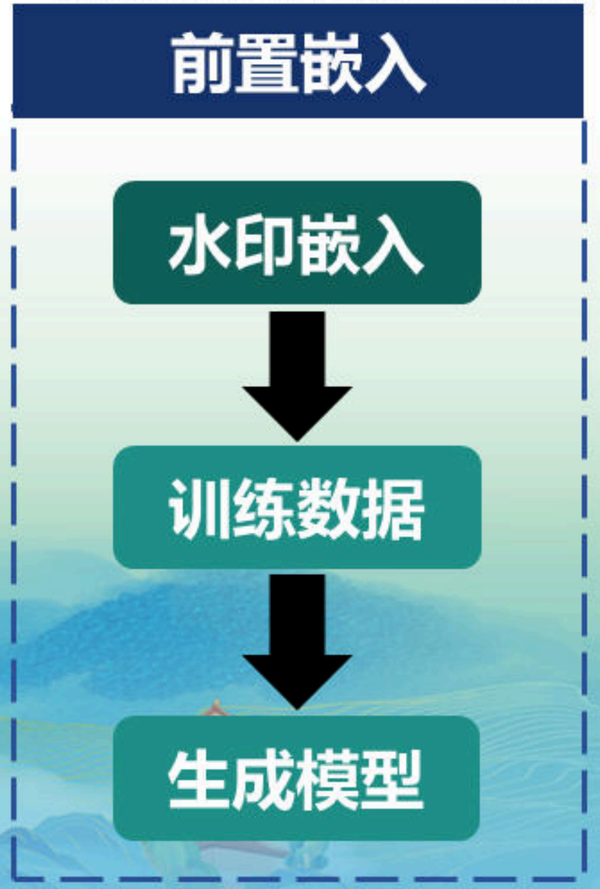


[1] Ning Yu, et al. ”Artificial Fingerprinting for Generative Models.” *ICCV*. 2021.

水印前置嵌入的生成图像溯源

■任务挑战

□水印前置嵌入方法，首先将水印嵌入到训练数据中，再利用嵌入水印的数据来训练生成模型，使得生成图像也具备相同的水印信息。



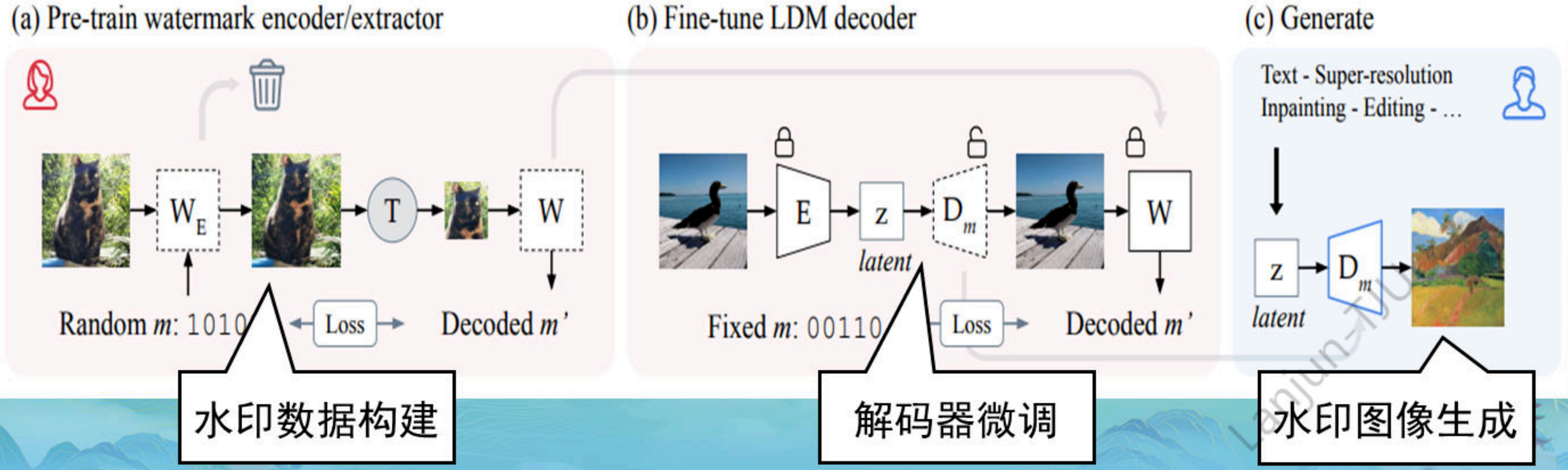
挑战 **水印信息到生成图像的转移性**

随着生成式大模型的发展，自2023年开始聚焦扩散模型

水印前置嵌入的生成图像溯源

■研究方法

□针对生成模型部署后难以判定生成图像来源，Meta预先将**水印嵌入训练数据**并训练水印编解码器，再通过**微调扩散模型解码器**生成水印图像，确保生成图像含有水印信息[1] 。



[1] Fernandez, P, et al. ”The stable signature: Rooting watermarks in latent diffusion models.” *ICCV*. 2023

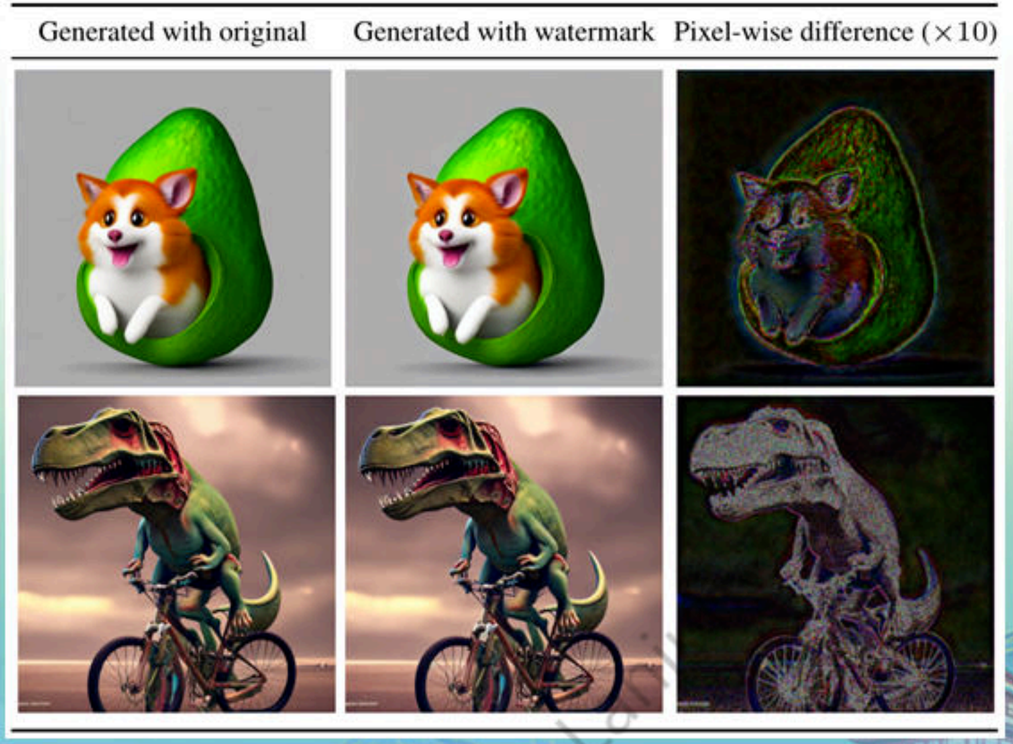
水印前置嵌入的生成图像溯源

■实验结果

- 在多种生成任务上针对SD模型进行测试，图像质量下降较少，且能实现**较高的水印恢复准确率**

			PSNR / SSIM ↑	FID ↓	Bit accuracy ↑ on: None	Crop	Brigh.	Comb.
Tasks	Text-to-Image	LDM [68]	30.0 / 0.89	19.6 (−0.3)	0.99	0.95	0.97	0.92
	Image Edition	DiffEdit [13]	31.2 / 0.92	15.0 (−0.3)	0.99	0.95	0.98	0.94
	Inpainting - Full	Glide [57]	31.1 / 0.91	16.8 (+0.6)	0.99	0.97	0.98	0.93
	Inpainting - Mask only	Glide [57]	37.8 / 0.98	9.0 (+0.1)	0.89	0.76	0.84	0.78
	Super-Resolution	LDM [68]	34.0 / 0.94	11.6 (+0.0)	0.98	0.93	0.96	0.92
WM Methods	*Post generation*							
	Dct-Dwt [14]	0.14 (s/img)	39.5 / 0.97	19.5 (−0.4)	0.86	0.52	0.51	0.51
	SSL Watermark [25]	0.45 (s/img)	31.1 / 0.86	20.6 (+0.7)	1.00	0.73	0.93	0.66
	FNNS [42]	0.28 (s/img)	32.1 / 0.90	19.0 (−0.9)	0.93	0.93	0.91	0.93
	HiDDeN [108]	0.11 (s/img)	32.0 / 0.88	19.7 (−0.2)	0.99	0.97	0.99	0.98
	Merged in generation							
	Stable Signature	0.00 (s/img)	30.0 / 0.89	19.6 (−0.3)	0.99	0.95	0.97	0.92

- 原图像和水印图像差分结果表明，**嵌入水印前后图像的视觉差异较小**



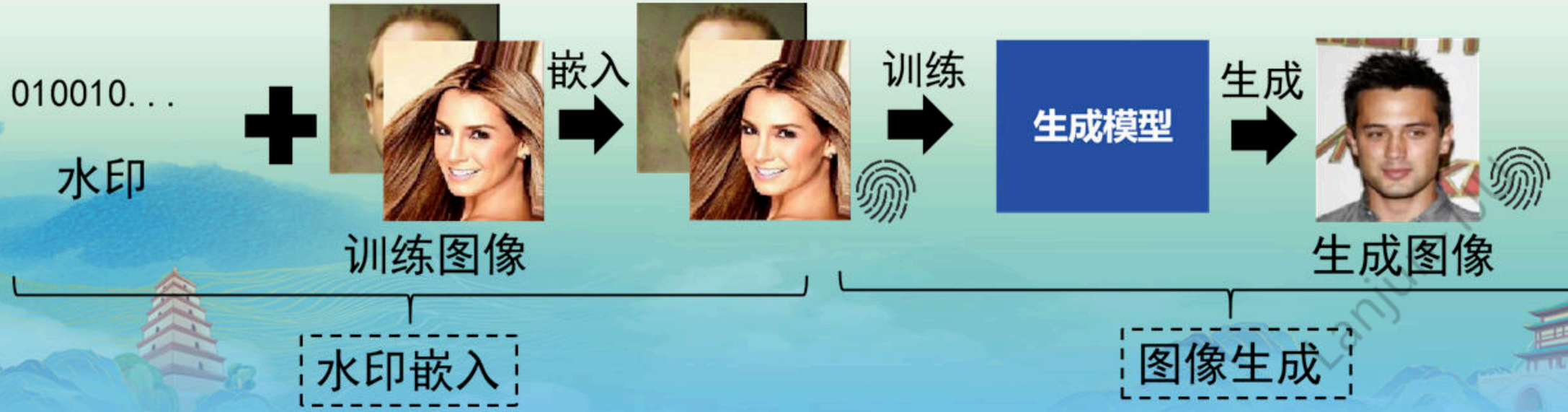
[1] Fernandez, P, et al. ”The stable signature: Rooting watermarks in latent diffusion models.” *ICCV*. 2023

水印前置嵌入的生成图像溯源

■总结

- 优势：对训练数据预先嵌入水印使得图像生成过程中不引入额外信息，因此对**生成图像质量损害较小**。

- 缺陷：需要对水印编解码器进行预训练或者水印嵌入训练数据进行预处理，在大规模数据集上的**构建代价较高，并且水印模式单一**



联合生成的生成图像溯源

■任务挑战

□联合生成方法，在图像生成过程中实现水印信息的自适应嵌入，同时显著减少对图像生成过程的影响，最终生成携带水印的图像。

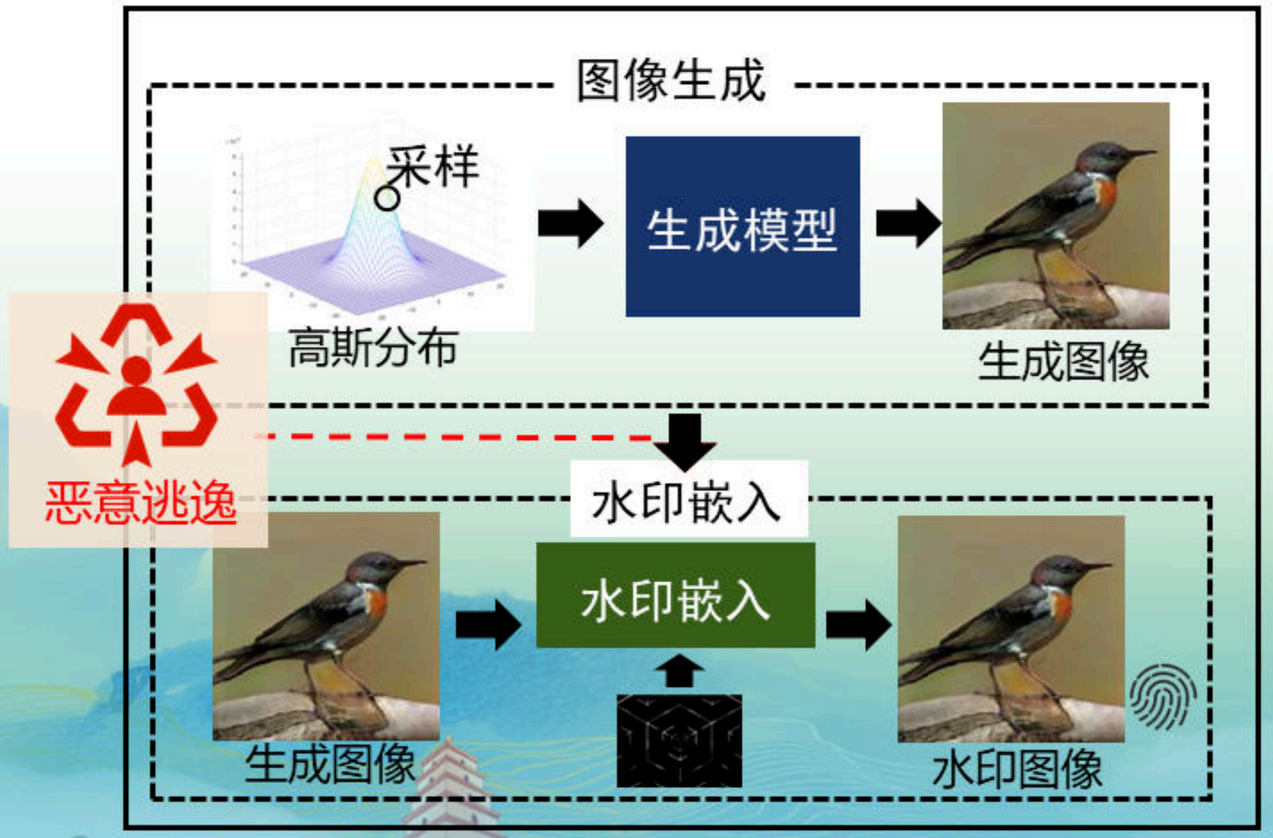
挑战1

水印和语义难耦合

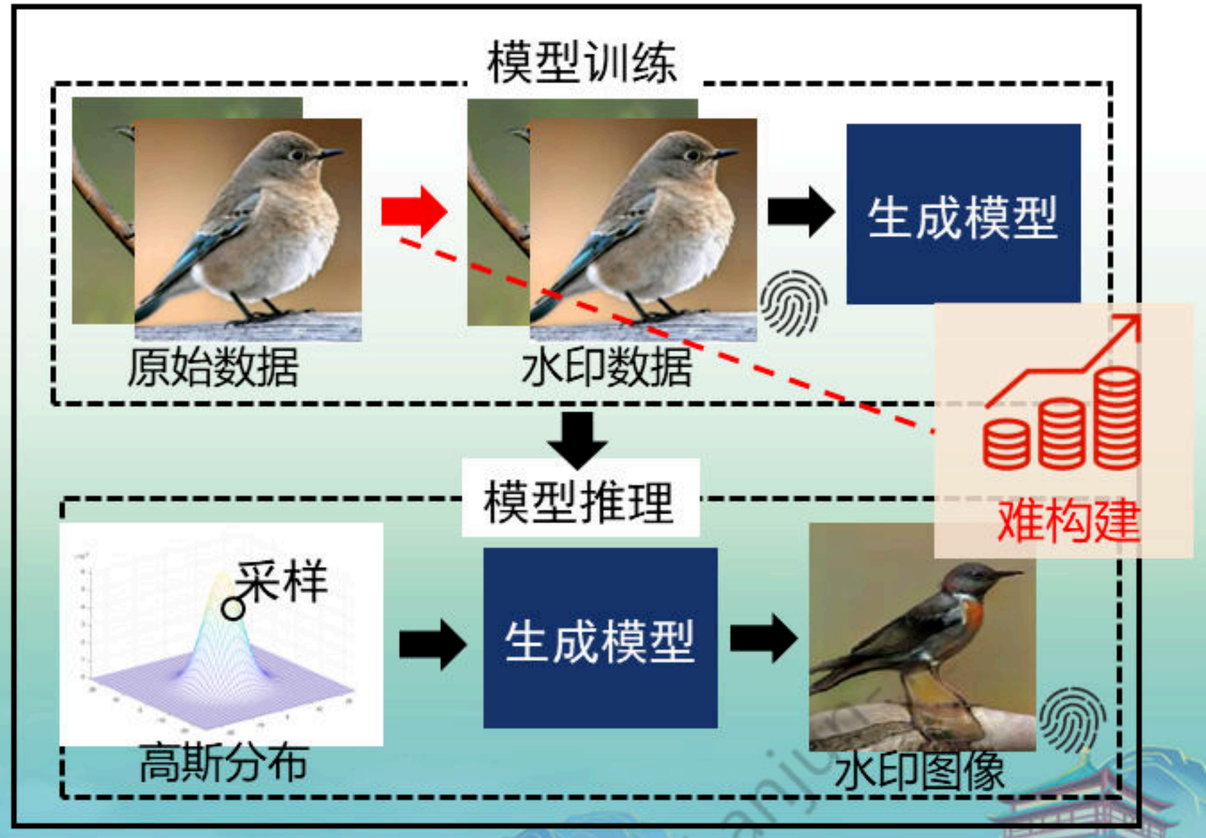
联合生成的生成图像溯源

■研究方法

□ 后置水印嵌入：逃逸攻击风险[1]



□ 前置水印嵌入：水印数据难构建[2]

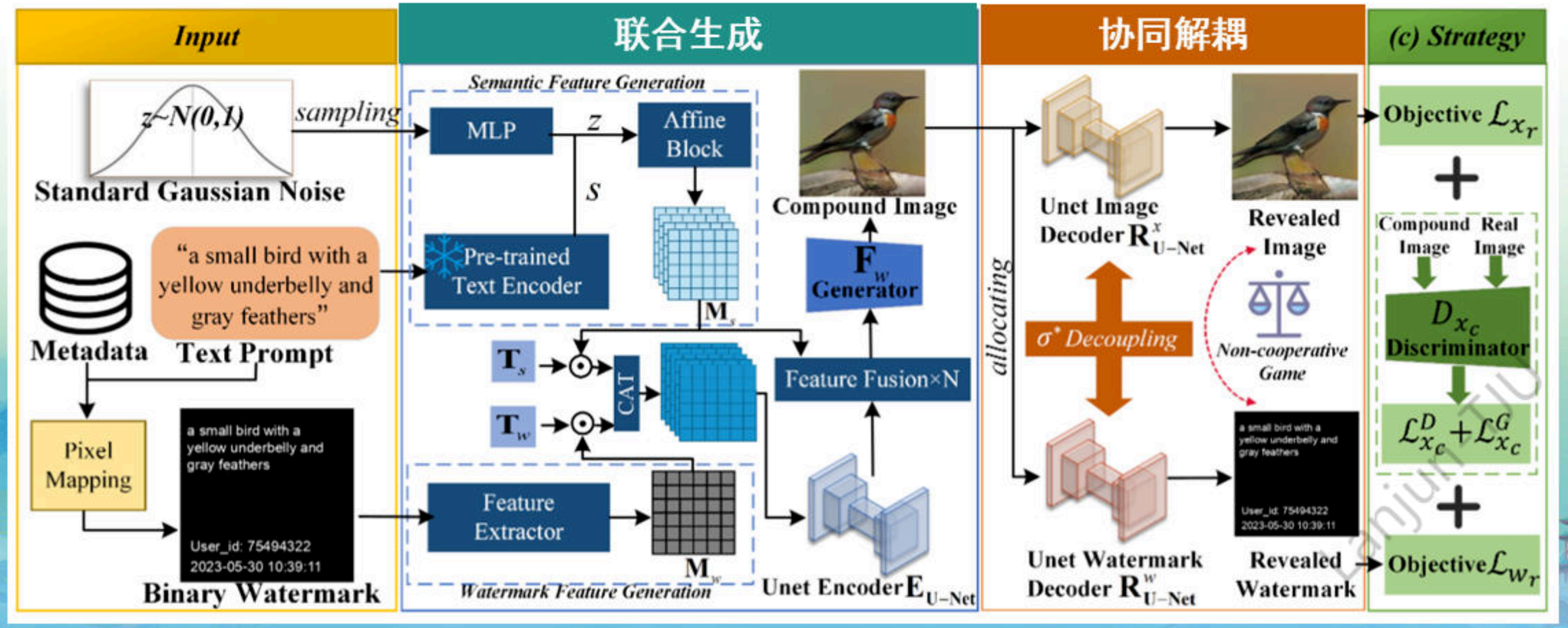


[1] Jianwei Fe, et al. ”Supervised gan watermarking for intellectual property protection.” *WIFS*. 2022.
[2] Ning Yu, et al. ”Artificial Fingerprinting for Generative Models:Rooting Deepfake Attribution in Training Data.” *ICCV*. 2022.

联合生成的生成图像溯源

■研究方法

□依据当前文生图和水印技术结合的迫切需求，[1]提出**文本到嵌入水印的图像跨模态生成**的新任务，做到图像和水印的联合生成与协同解耦，实现**图像高质量**生成和**水印强鲁棒**恢复。



[1] An-An Liu, et al. "T2IW: Joint Text to Image & Watermark Generation." *arXiv*. 2023.

联合生成的生成图像溯源

■实验结果

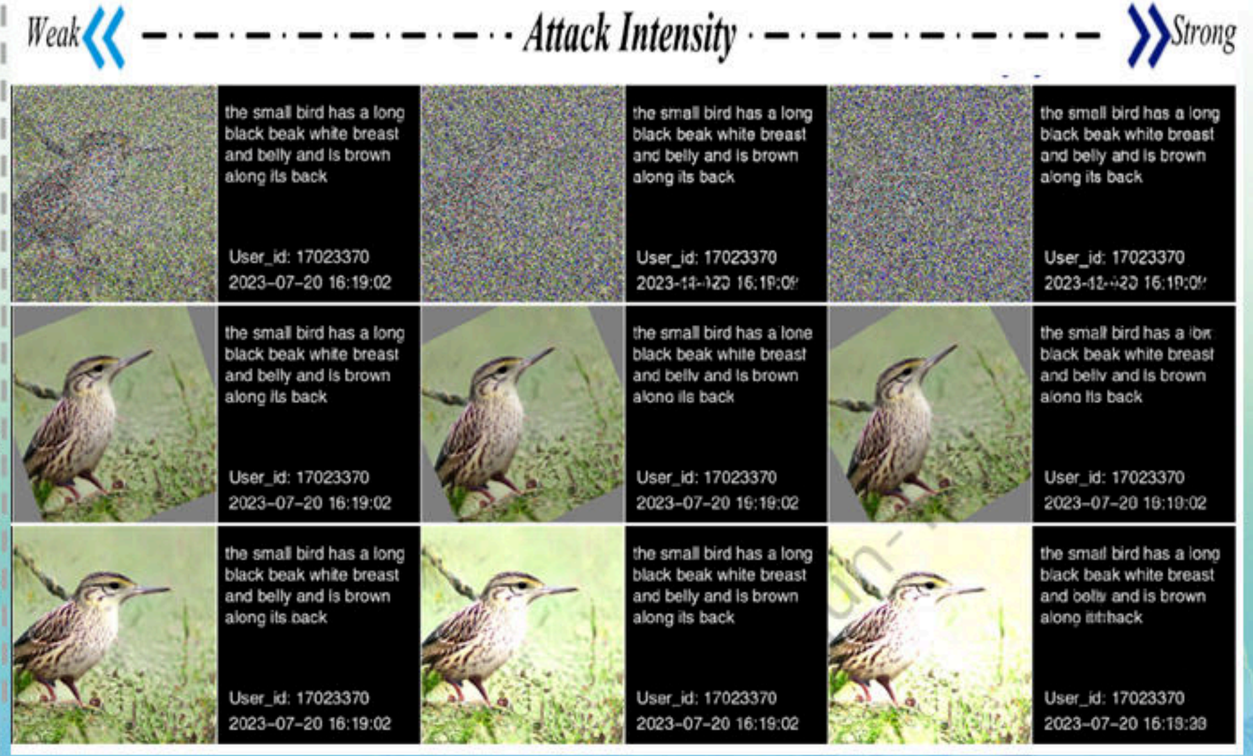
- 嵌入水印对**图像生成质量影响微弱**

Models	Mode	CUB-birds		Oxford-102 flowers	
		IS↑	FID↓	IS↑	FID↓
RAT-GAN [2]	原始图像	5.36 ± 0.20	13.91	4.09 ± 0.06	16.04
	水印图像	4.94 ± 0.06	16.95	3.72 ± 0.07	18.35
	无水印图像	4.98 ± 0.06	17.32	3.81 ± 0.07	19.16
AttnGAN [25]	原始图像	4.36 ± 0.03	23.98	–	–
	水印图像	4.02 ± 0.05	26.49	–	–
	无水印图像	4.09 ± 0.05	26.01	–	–

- 有无水印图像相似性超过98%，**水印隐蔽性强**

Models	CUB-birds			Oxford-102 flowers		
	PSNR (dB)↑	SSIM(%)↑	LPIPS↓	PSNR (dB)↑	SSIM(%)↑	LPIPS↓
RAT-GAN [2]	33.29	98.46	0.0257	33.51	98.60	0.0231
AttnGAN [25]	33.86	98.15	0.0223	–	–	–

- 多种不同强度攻击下，**水印表现出强鲁棒性**

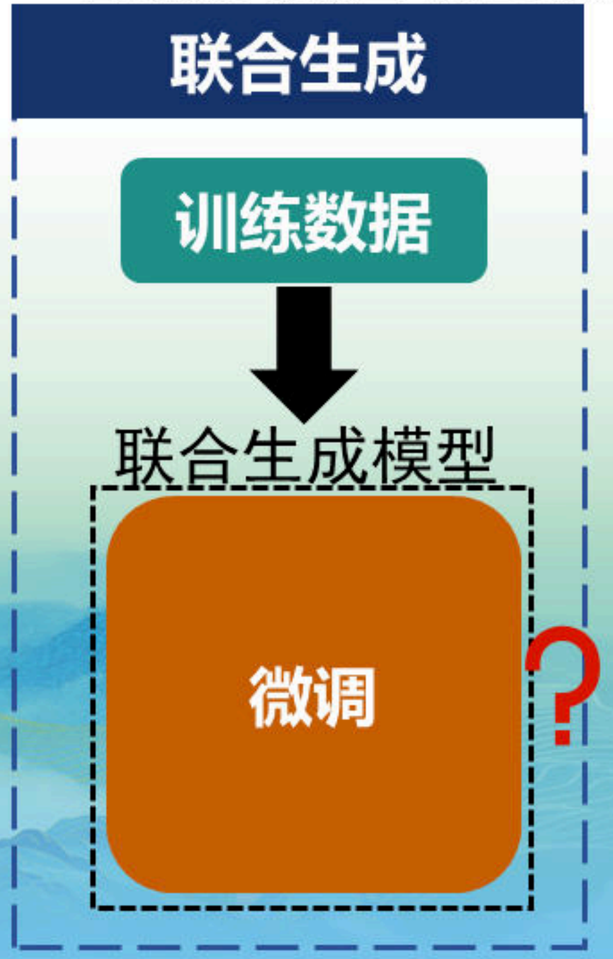


[1] An-An Liu, et al. "T2IW: Joint Text to Image & Watermark Generation." *arXiv*. 2023.

联合生成的生成图像溯源

■任务挑战

□联合生成方法，在图像生成过程中实现水印信息的自适应嵌入，同时显著减少对图像生成过程的影响，最终生成携带水印的图像。



挑战1

水印和语义难耦合

2023年开始转变

挑战2

训练成本高

联合生成的生成图像溯源

■研究动机

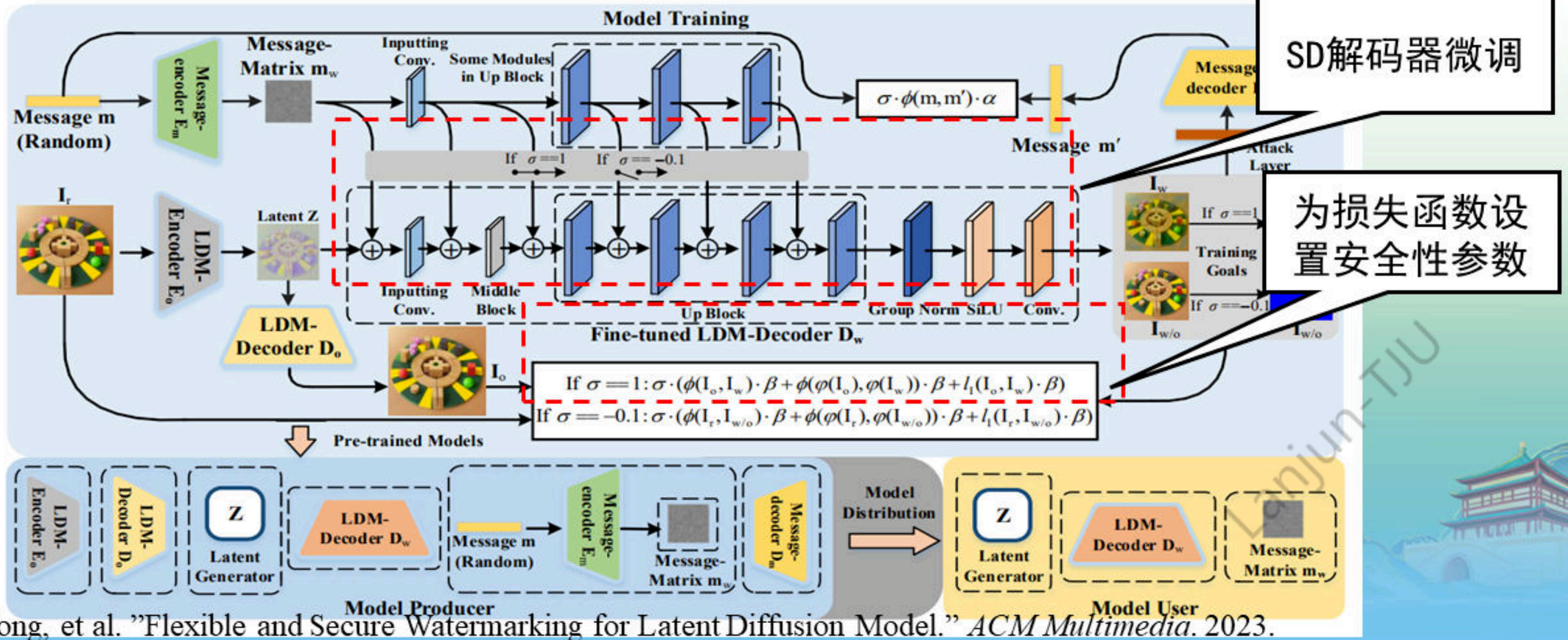
- 不进行端到端训练，仅微调模型
- 恶意用户通过窃取微调部分而获取无水印生成图像，即模型容易受到逃逸攻击。

[1] Cheng Xiong, et al. "Flexible and Secure Watermarking for Latent Diffusion Model." *ACM Multimedia*. 2023.

联合生成的生成图像溯源

■研究方法

- □仅微调SD解码器
- □针对目前水印方法存在逃逸攻击风险，[1]在损失函数中设计安全参数来微调Stable Diffusion图像解码器，能够破坏未添加水印的生成图像。



[1] Cheng Xiong, et al. "Flexible and Secure Watermarking for Latent Diffusion Model." *ACM Multimedia*, 2023.

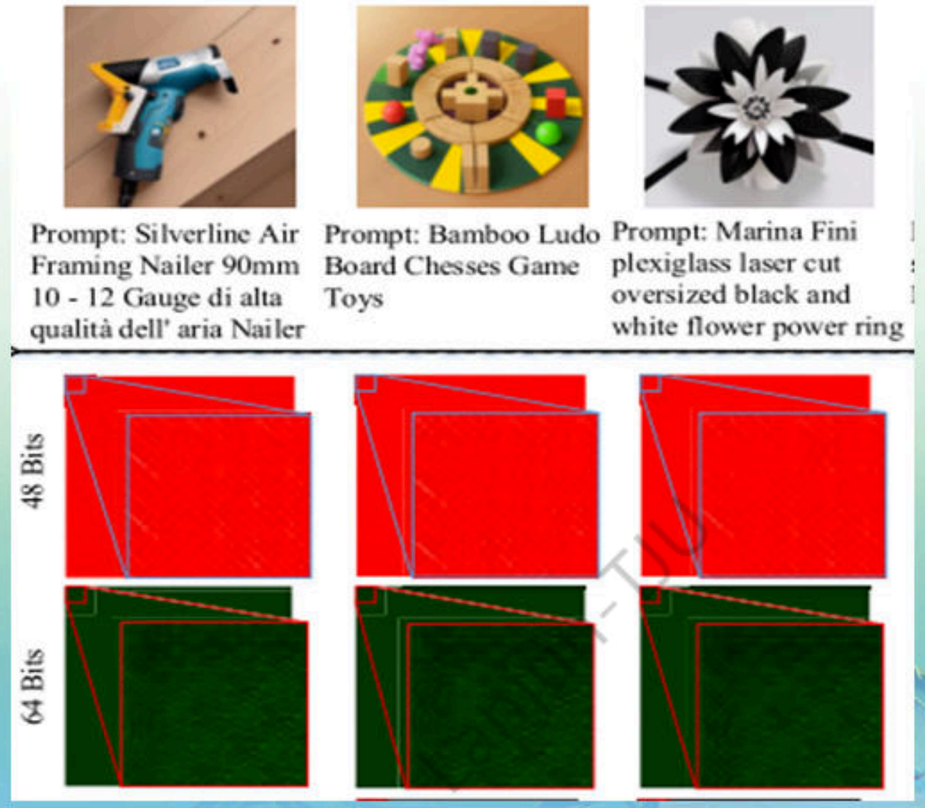
联合生成的生成图像溯源

■实验结果

- 尽管视觉质量有所下降，但水印恢复准确率能够达到99.9%

Methods	PSNR ↑ (dB)	SSIM ↑	FID ↓	Bit Acc↑
DCT [18]	**44.58**	**0.99**	**0.40**	0.999
HiDDeN [28]	26.33	0.95	13.56	0.999
StegaStamp [20]	28.40	0.88	12.26	0.999
MBRS [10]	35.36	0.96	21.17	0.879
Proposed	32.51	0.93	6.35	**0.999**

- 水印图像正常生成，但经过逃逸攻击的无水印图像被破坏

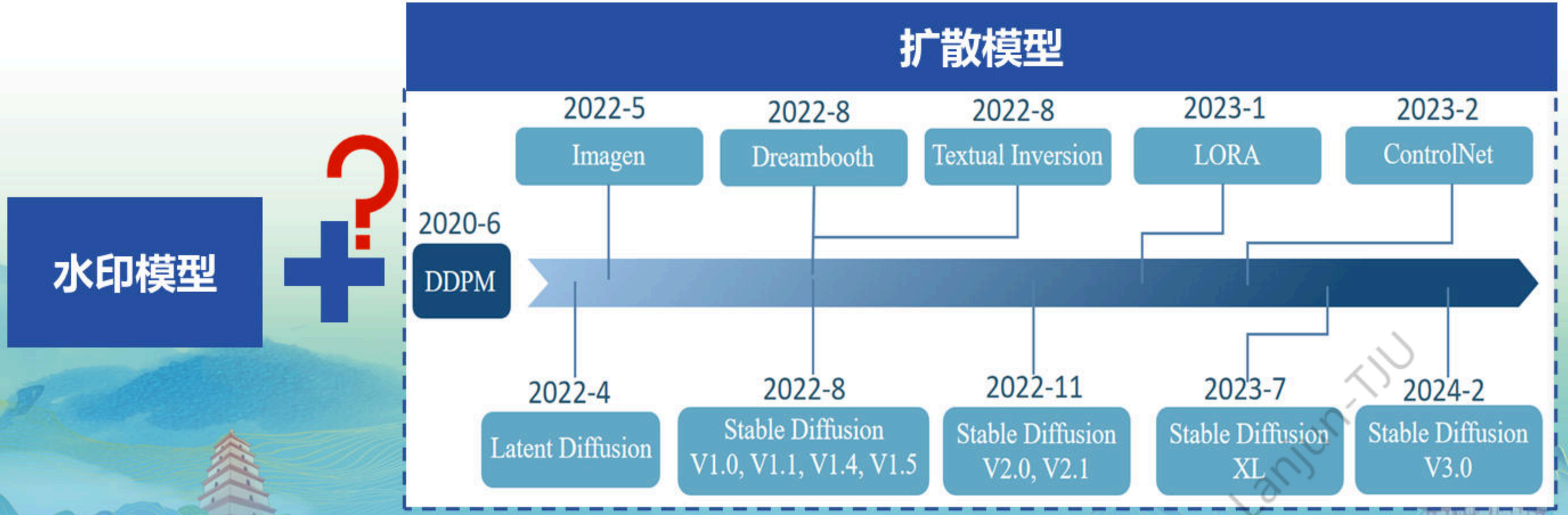


[1] Cheng Xiong, et al. "Flexible and Secure Watermarking for Latent Diffusion Model." *ACM Multimedia*. 2023.

联合生成的生成图像溯源

■研究动机

□针对生成式模型的水印方法需要**训练部分或者全部的扩散模型参数**，难以适应当前扩散模型技术的高速迭代。

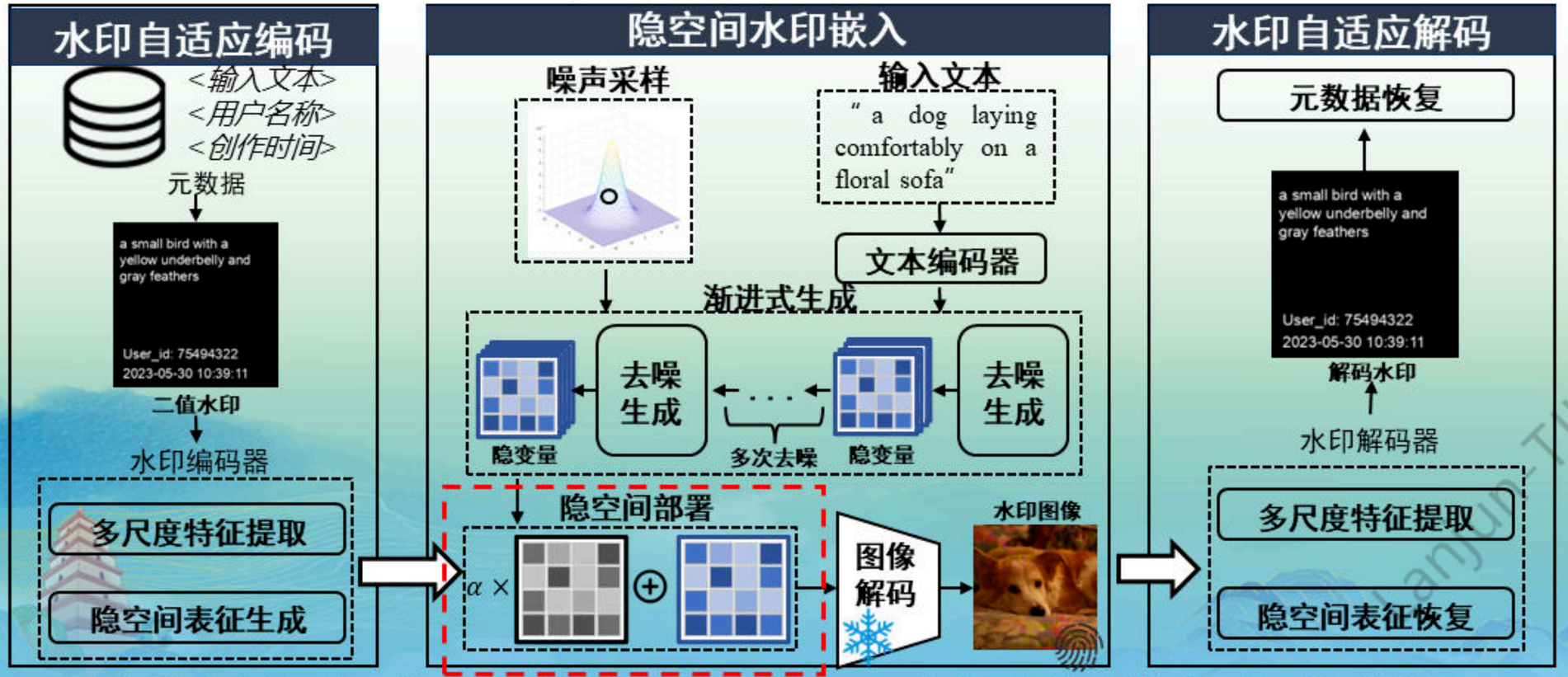


[1] Guokai Zhang, et al. ”A Training-Free Plug-and-Play Watermark Framework for Stable Diffusion.” *arXiv*. 2024.

联合生成的生成图像溯源

■研究方法

□针对现有水印模型对不同的生成式大模型缺乏可迁移性，[1]提出**即插即用、无需训练**的Stable Diffusion水印方法，通过在隐空间将水印信息部署在隐变量单通道中，实现**低成本、可扩展**的生成内容溯源。



[1] Guokai Zhang, et al. "A Training-Free Plug-and-Play Watermark Framework for Stable Diffusion." *arXiv*. 2024.

联合生成的生成图像溯源

■实验结果

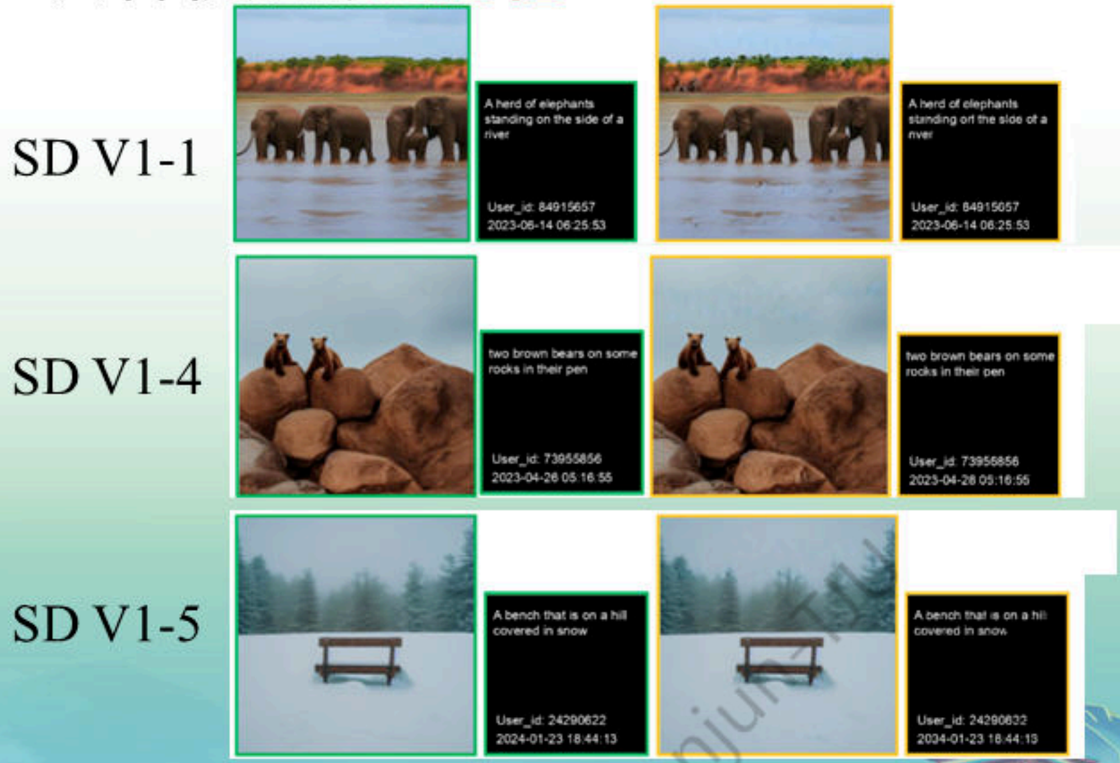
- 迁移到其他版本SD模型中，性能未出现下降，证明实现**即插即用且无需训练**

Models	Watermark Invisibility			Watermark Quality		
	PSNR(dB)↑	SSIM(%)↑	LPIPS↓	NC(%)↑	CA↓	CER(%)↓
COCO-sub						
Stable Signature [7]	31.06	90.67	0.04993	N/A	N/A	N/A
SD-wm v1-1*	**37.04**	94.35	**0.04207**	96.15	13.31	14.20
SD-wm v1-4†	36.93	94.10	0.04323	**97.14**	11.97	**13.32**
SD-wm v1-5†	36.97	**94.46**	0.04237	96.78	**11.85**	13.33
Flikr-8K						
Stable Signature [7]	30.84	90.28	0.05062	N/A	N/A	N/A
SD-wm v1-1†	**36.77**	93.74	0.04463	96.15	**13.31**	**14.20**
SD-wm v1-4†	36.64	93.34	0.04514	**96.67**	13.67	14.83
SD-wm v1-5†	36.72	**93.82**	**0.04366**	96.44	14.40	15.13

- 与现有方法相比，运算次数和训练参数显著下降，证明该方法的**高效性**

Models	FLOPs(G)	Params(M)
Stable Signature [7]	2767.15	50.84
ENDE [41]	3254.42	67.13
Ours	**9.72**	**0.97**

- 各个版本SD模型生成图像表明，该方法具备较强的**泛化性**



[1] Guokai Zhang, et al. ”A Training-Free Plug-and-Play Watermark Framework for Stable Diffusion.” *arXiv*. 2024.

联合生成的生成图像溯源

■总结

- 无需事先构建携带水印的训练数据，并且不存在多阶段间逃逸攻击风险，所以该类方法能够**低成本、低风险**地应用于生成模型中。

- 该类方法预计短时间内不会随着生成技术的革新或者数据集的变更而失效，且能够适配各类模型，是未来为生成式技术保驾护航的主要研究路线。

生成图像溯源领域进展

■总结

- 水印后置嵌入溯源方法**避免生成和嵌入相互干扰**，但存在**信息篡改**和**逃逸攻击**等安全性问题

- 水印前置嵌入溯源方法对**生成图像质量损害小**，但**水印数据构建代价高**。

- 联合生成方法**适应各类生成模型的生成特性**，是未来生成图像溯源技术的主要研究路线。

总结与展望

总结与展望

图像生成领域

总结

可控生成式技术发展比较成熟，**已经具备实际应用能力**。

展望

1. 生成模型能力边界的进一步探索
2. 更贴近实际应用需求的可控生成技术

生成图像溯源领域

总结

随着生成式大模型的快速发展，研究重心由传统的水印后置嵌入方法向**水印前置嵌入方法**和**联合生成方法**转移。

展望

1. 面向生成式大模型的联合生成式水印
2. 可验证性能无损的水印
3. 针对生成图像水印的鲁棒性检测基准

敬请批评指正

王岚君
天津大学
wanglanjun@tju.edu.cn