GFM4D: Conditional Diffusion Model for Post-event Satellite Image Generation

Presenter: Zhenyuan Chen （bili_sakura@zju.edu.cn）

Date: August 26, 2024

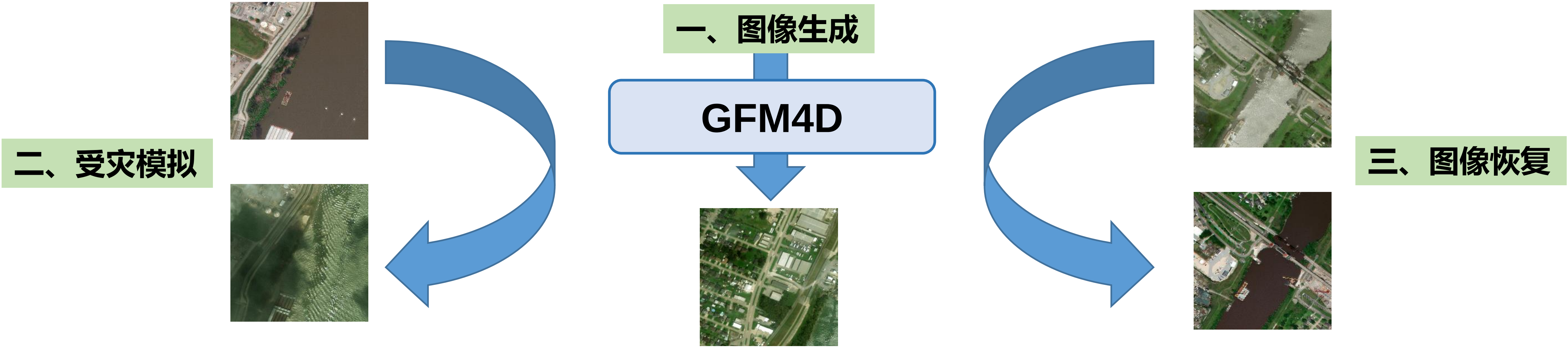
Outline

- Background
- Main Contribution
- Current Progress
- Related Work
  - Generative Geo-Foundation Models
  - RSIs Captioning
- Dataset
- Techniques
  - Backbone
  - Image Editing
  - Temporal Generation
- Evaluation

Background

遥感图像在资源分配、救援路线规划、救援和恢复等人道主义援助与灾害响应任务中具有重要作用。通过卫星影像和计算机视觉算法，可以远程和自动化地进行灾害评估，避免进行危险的现场评估 [1]。灾害具有偶发性和即时性的特点，导致遥感图像获取和制作困难，存在数量稀少、时空分布不均等问题 [2]。扩散模型 [3, 4, 5] 作为生成式 AI 的典型代表，展示出强大的图像合成能力 [6, 7, 8, 9]，为解决灾害遥感图像数据稀少的问题提供了一种解决方案。然而，现有的扩散模型基于通用自然图像数据集 [10, 11, 12, 13, 14] 进行训练，缺乏遥感场景的特化知识学习，在遥感图像合成中表现欠佳。本研究中，我们构建了一个以扩散模型为底座的生成式遥感基模型 Generative Foundation Model for Disaster(GFM4D)，用于灾害遥感图像的合成。基于此模型，我们制作一个全球范围的大体量的灾害场景特化的遥感图像文本对数据集，可用于遥感场景预训练模型的继续训练和微调，以进一步提升模型在下游任务 (如图像标注 [15]，场景分类 [16, 17]，目标检测 [18, 19]，变化检测 [20, 21]，语义分割 [22] 和视觉问答 [23] 等) 的表现。

文本描述：卫星影像显示，洪水泛滥，河流水位暴涨，淹没了房屋、道路和农田。整个地区被洪水覆盖，只能看到屋顶和树冠。救援人员正在紧急展开救援和物资运送。
元数据：【位置】（30°N，120°E），【时间】2020/7/10，【分辨率】0.35m



GFM4D是一个生成式遥感基模型，可以支持多种形式的输入，生成高质量的可靠的灾害卫星影像。

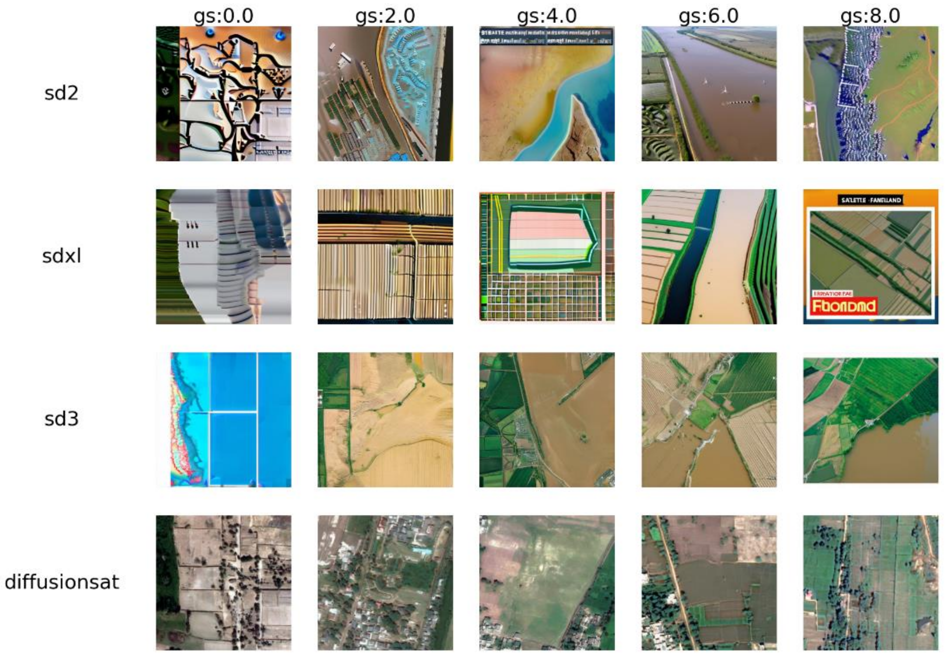
Main Contribution

本研究创新点主要包括以下几个部分：

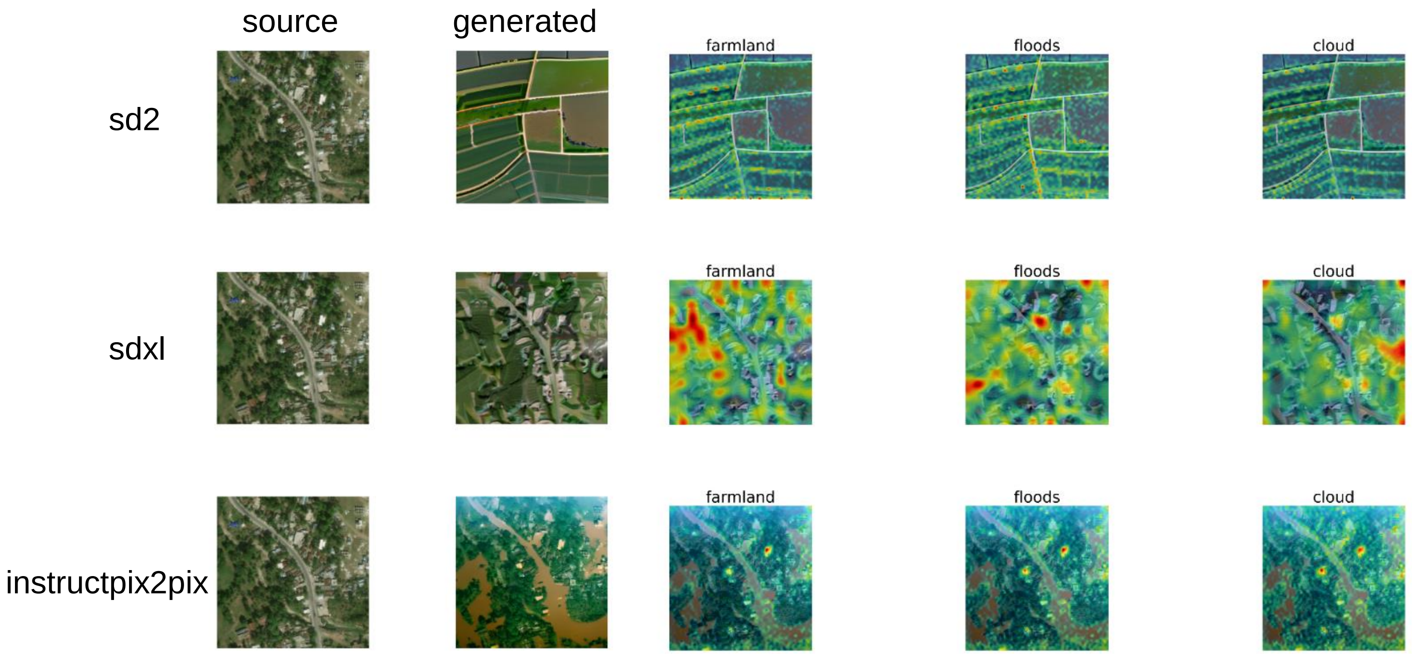
1. 首个将百亿级别开源扩散通用基模型（如 Stable Diffusion 3，Flux.1-schnell, FLUX.1-dev）应用于遥感图像生成，并全面对比和评估当下最先进的条件控制图像生成技术 (如 Uni-Control [68], ControlNet++ [69], ControlNeXt [70])，为后续遥感图像生成研究铺路。
2. 提出一个用于遥感图像时序生成模块RSTemporal Layer，相较于NVIDIA在Video LDM [67] 中提出的应用于DiffusionSat [2]的Temporal Layer能更高效的处理稀疏的遥感时序数据。
3. 基于多个主流多模态大模型，使用 Woodpecker [56] 框架和集成学习的方式为现有的灾害遥感图像数据集生成高质量文本描述，构建高质量预训练数据集。
4. 构建首个以生成式模型制作的灾害遥感合成数据集 RSD5M，预训练模型经过此数据集继续训练在下游任务表现得到提升，证明数据集的有效性。

Current Progress

prompt=“A farmland being flooded in India, no cloud”



how hyper-parameter guiding scale affect results

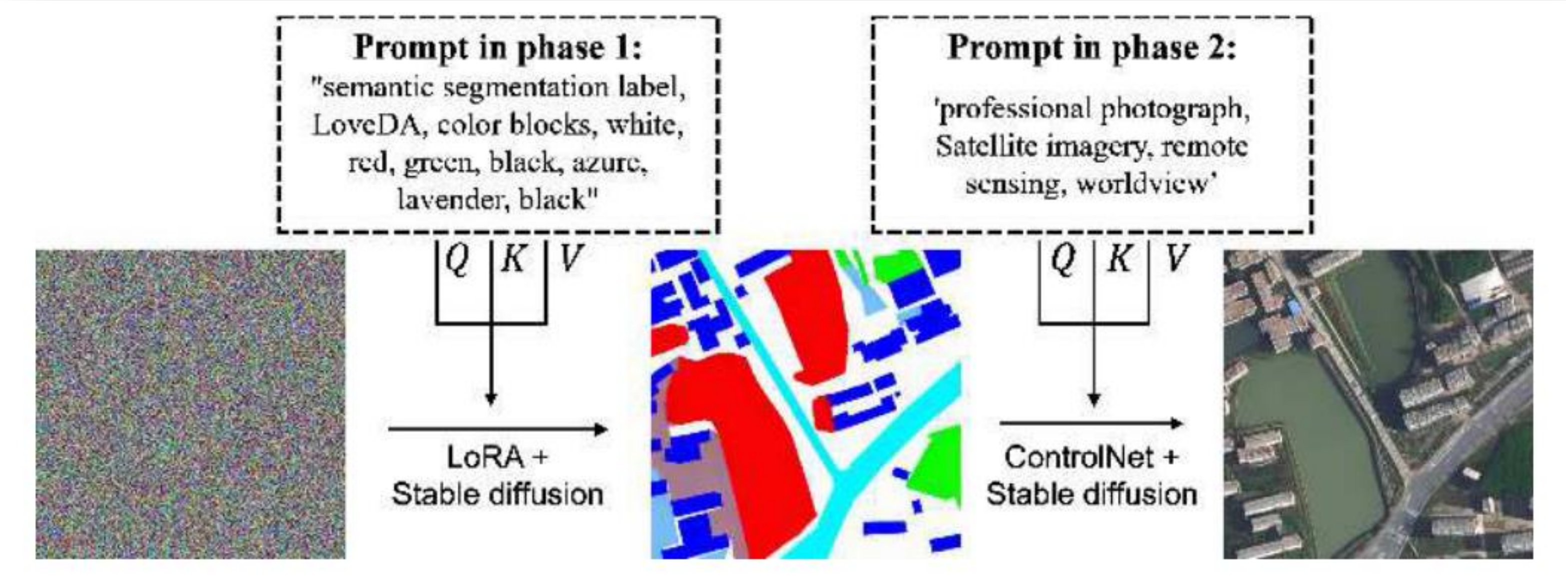


Visualization of attention in prompt using DaaM

R. Tang et al., “What the DAAM: Interpreting Stable Diffusion Using Cross Attention,” presented at the ACL 2023, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5644–5659. doi: 10.18653/v1/2023.acl-long.310.

[44] C. Zhao, Y. Ogawa, S. Chen, Z. Yang, and Y. Sekimoto, “Label Freedom: Stable Diffusion for Remote Sensing Image Semantic Segmentation Data Generation,” in 2023 IEEE International Conference on Big Data (BigData), Dec. 2023, pp. 1022–1030.

Generative Geo-Foundation Models



(a) annotation-image pairs generated from scratch

(b) generated results and comparison with real images.

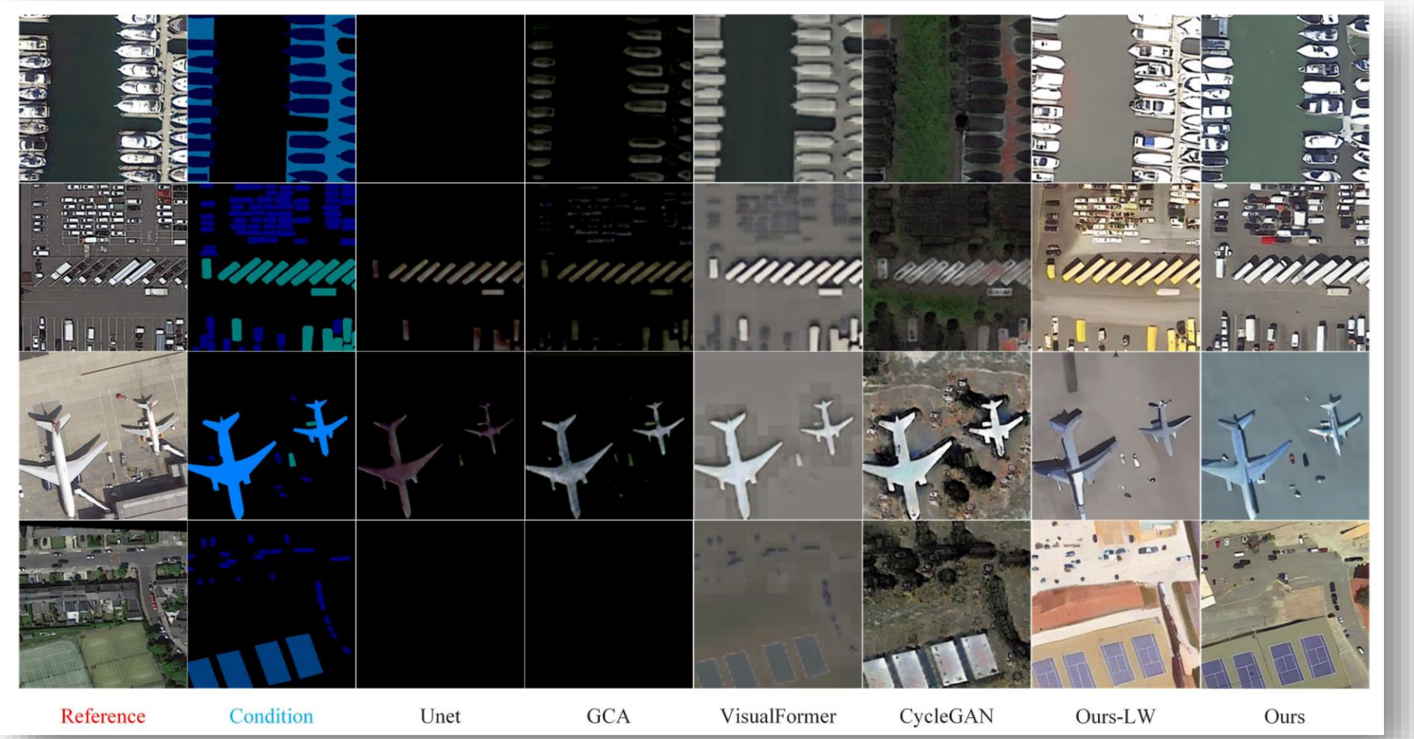
[38] R. Ou, H. Yan, M. Wu, and C. Zhang, "A Method of Efficient Synthesizing Post-disaster Remote Sensing Image with Diffusion Model and LLM," in 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Oct. 2023, pp. 1549–1555.

Generative Geo-Foundation Models

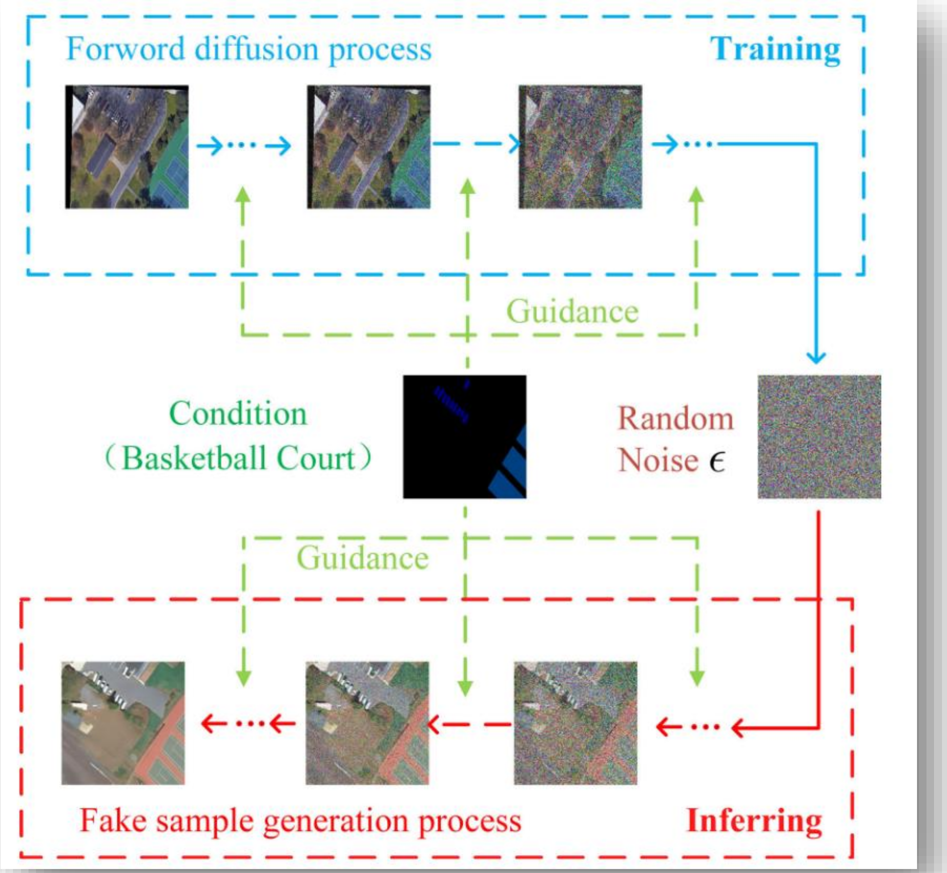
Fig. 1: **Synthesizing Remote Sensing Images Affected by Disaster under Natural Language Guidance with Diffusion Model and LLM.** (1)The images in left block are synthesized in generation manner. The text to the left of the image is condition promtps for synthesizing by the row of images. (2)The images in right block are synthesized in inpainting manner under the same condition of the generation. In addition, synthesizing the images with inpainting manner can obtain the ground truth of disaster as the position label for other interpretation model learning.

[42] Z. Yuan, C. Hao, R. Zhou, J. Chen, M. Yu, W. Zhang, H. Wang, and X. Sun, "Efficient and Controllable Remote Sensing Fake Sample Generation Based on Diffusion Model," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–12, 2023.

Generative Geo-Foundation Models



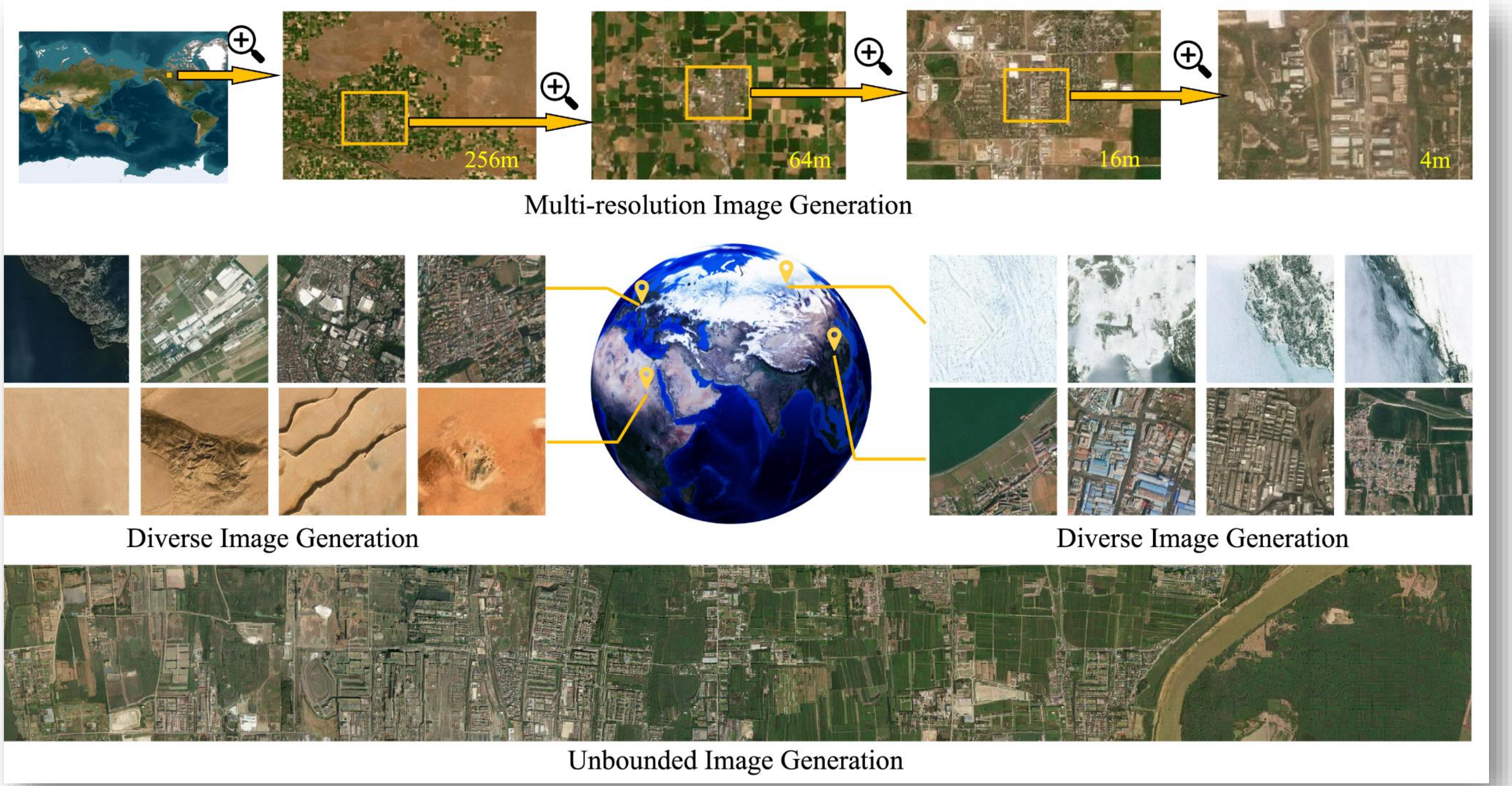
Qualitative comparison of different models



Framework of RSFSG based on the diffusion model.

[37] Z. Yu, C. Liu, L. Liu, Z. Shi, and Z. Zou, “MetaEarth: A Generative Foundation Model for Global-Scale Remote Sensing Image Generation,” May 2024.

Generative Geo-Foundation Models

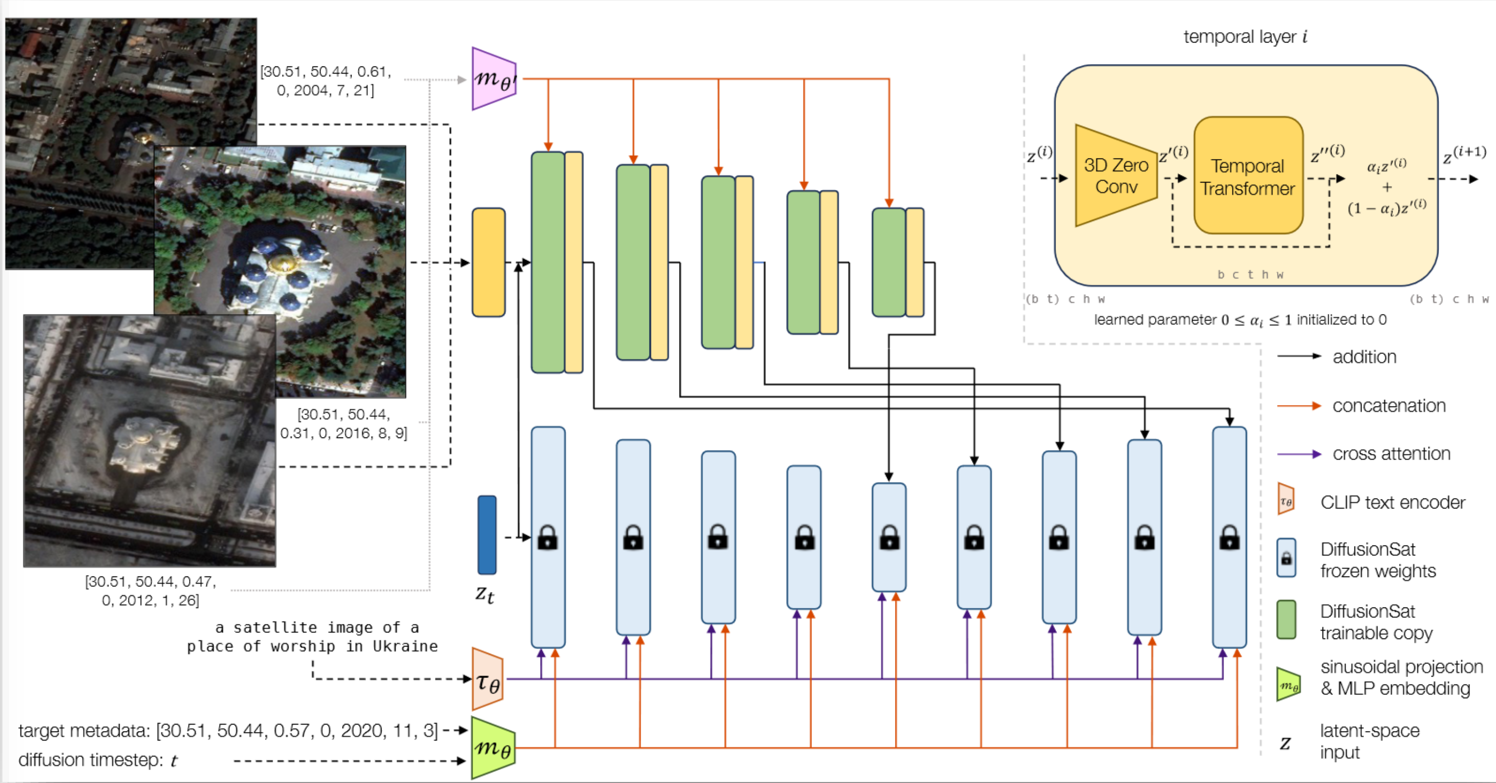


We propose MetaEarth, a generative foundation model that simulates Earth’s visuals from an overhead perspective.

[2] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, "DiffusionSat: A Generative Foundation Model for Satellite Imagery," in The TwelfthInternational Conference on Learning Representations, Oct. 2023.

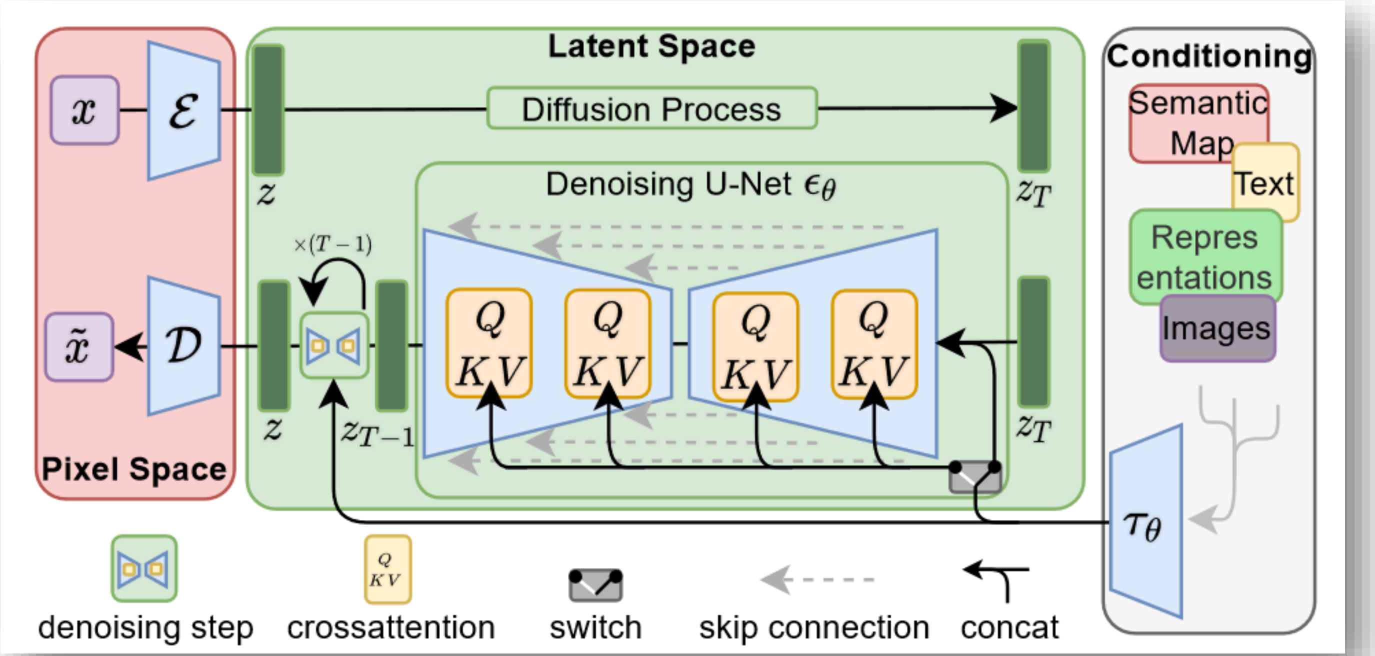
Generative Geo-Foundation Models

Inpainting results. DiffusionSat successfully reconstructs damaged roads and houses from floods, fires, and wind, even when large portions of the conditioning image are masked by clouds or damage.



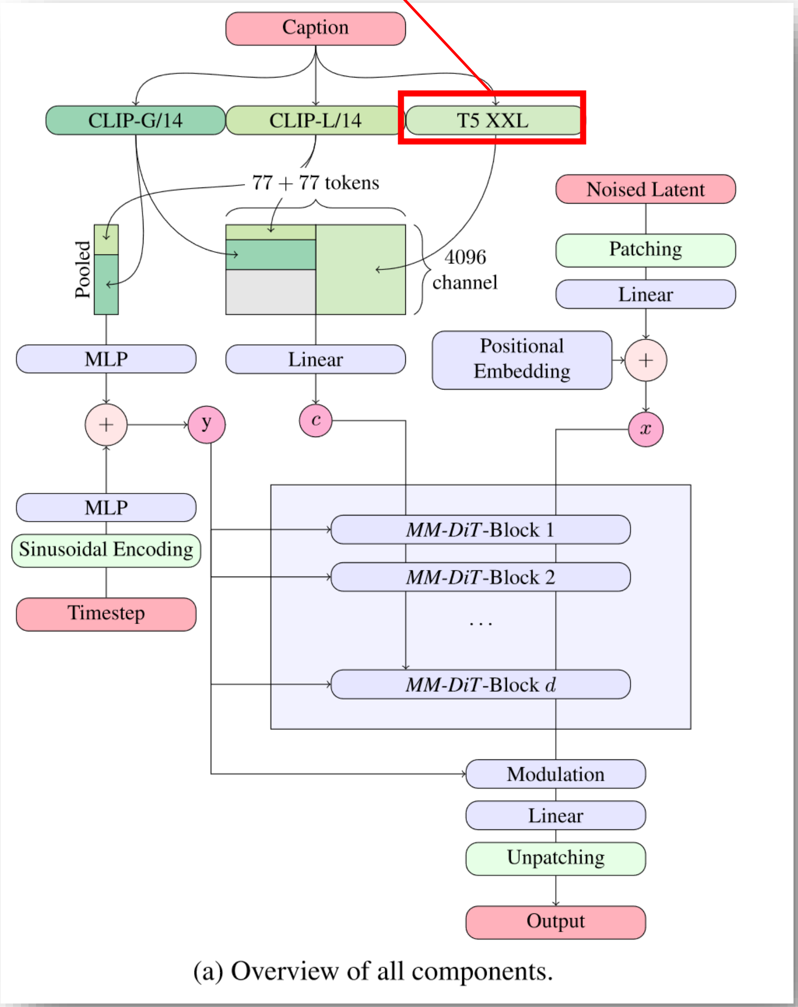
Architecture of DiffusionSat for conditional generation tasks. A novel 3D version of a ControlNet [71] is adopted.

Backbone



Latent Diffusion Model (Stable Diffusion)

Large Language Model T5 Encoder Part (5B)



Stable Diffusion 3

P. Esser et al., "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis," in Forty-First International Conference on Machine Learning, Jun. 2024.
R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.

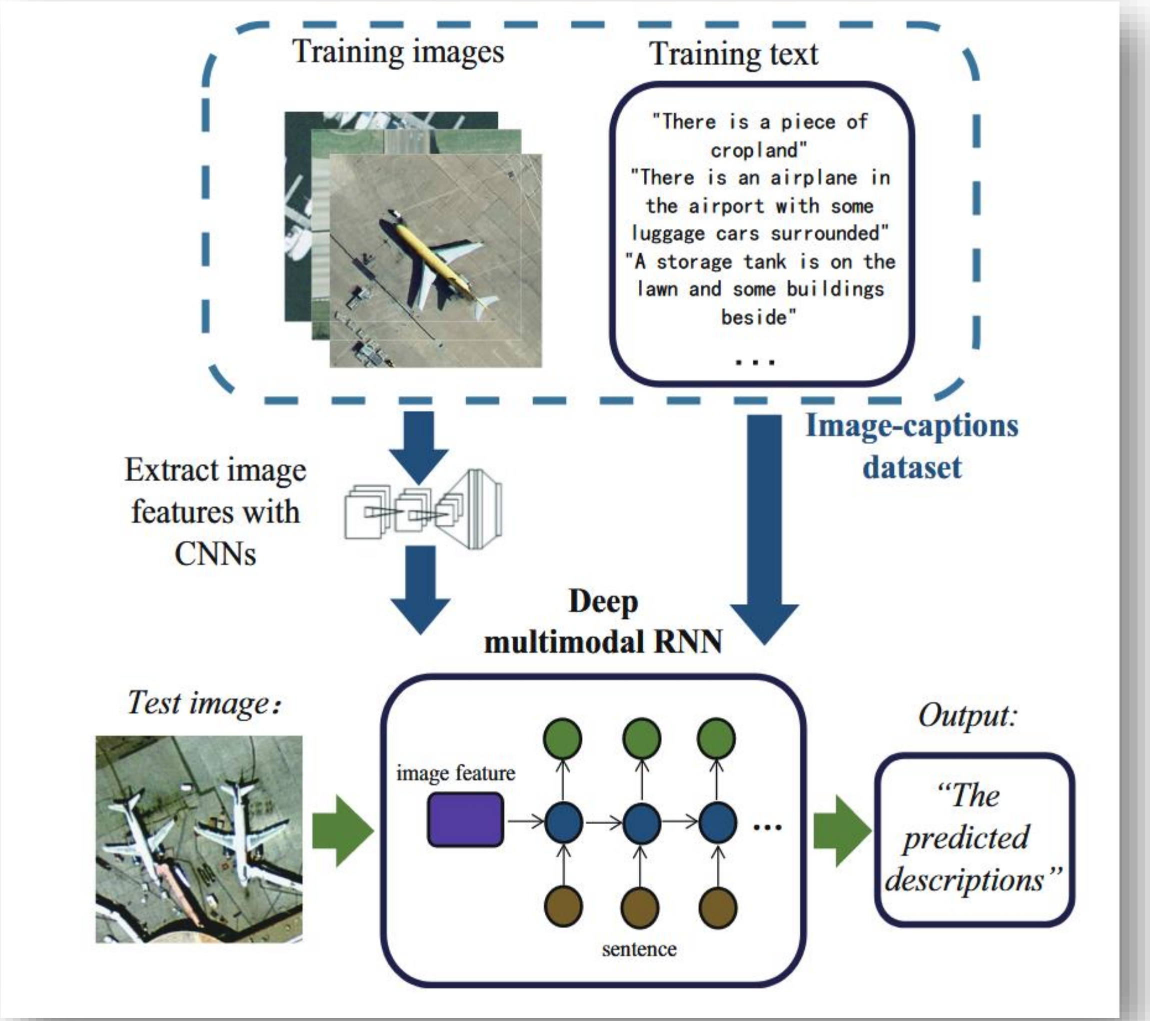
Generate High Quality Descriptions for Disaster RSIs

- Imagen [45] showcases the effectiveness of frozen large pretrained language models as text encoders for the text-to-image generation using diffusion models. Inspired by this idea, DALL-E 3 [8] shows that prompt following abilities of text-to-image models can be substantially improved by training on highly descriptive generated image captions. Stable Diffusion 3 [9] follows previous findings, leverage 3 different text encoder for better image synthesis.
- In terms of remote sensing images (RSIs), high-quality descriptive captions for RSIs are scarce. Even for inspiring work as DiffusionSat [2], the caption of training dataset is quite short and less informative.
- Therefore, there is a strong need for high-quality informative text-image RSIs dataset.

Dataset	Caption
fMoW	"a [fmow] satellite image [of a <object>] [in <country>]"
SpaceNet	"a [spacenet] satellite image [of <object>] [in <city>]"
Satlas	"a [satlas] satellite image [of <object>]"
Texas Housing	"a [satlas] satellite image [of houses] [built in <year_built>] [covering <num_acres> acres]"
xBD	"a [fmow] satellite image [<before/after>] being affected by a <disaster_type> natural disaster"

Captions created for each dataset type based on available label information from DiffusionSat [2].

Classic Image Captioning Method



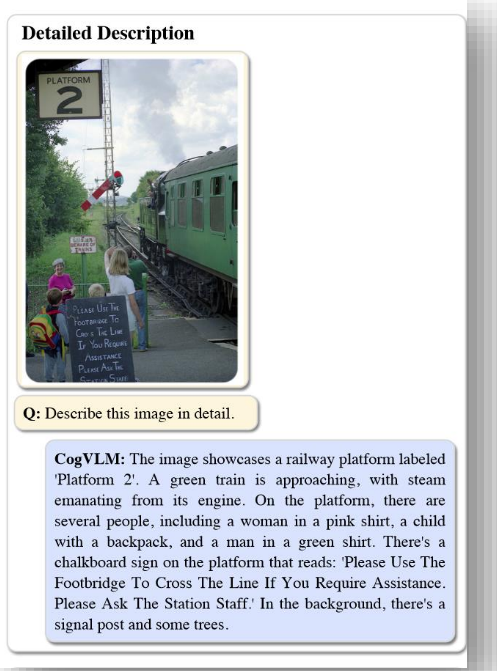
Left: Overview of model proposed in [15]. Right: The result of HSR image caption generation.

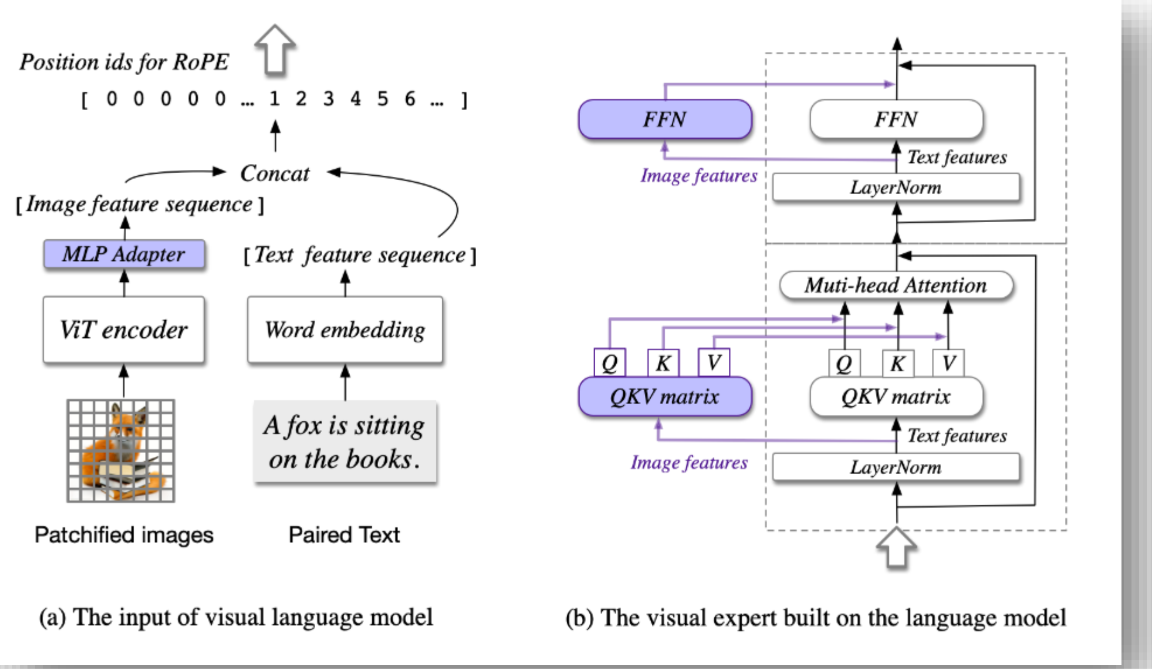
List of Captioning Models

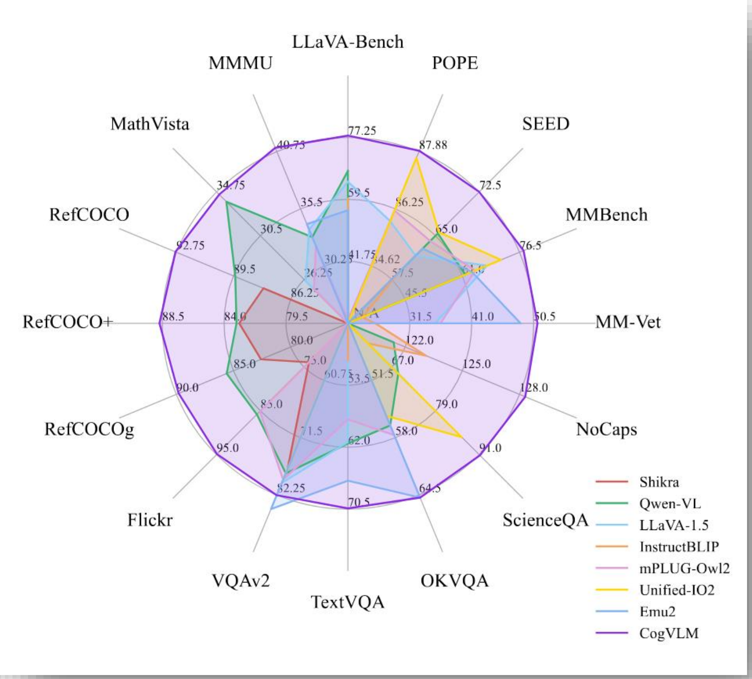
Model	Type	Text Model	Image Model	Total Size	Release Date
CogVLM 2	General	LLaMA-3 (8B)	EVA2-CLIP-E (5B)	**19B**	2024.05.20
MiniCPM-V 2.6	General	Qwen2 (7B)	SigLIP (400M)	**8B**	2024.08.17
mPLUG-Owl3	General	Qwen2 (7B)	SigLIP (400M)	**8B**	2024.08.12
xGen-MM (BLIP-3)	General	Phi3-mini (3.8B)	SigLIP (400M)	**4.3B**	2024.08.17
InternVL 2	General	InternLM 2.5 (7B)	InternViT (300B)	**8B**, 40B,76B,108B	2024.07.04-now
Cambrian	General	LLaMA-3 (8B)	Aggretion with 4 encoders CLIP (300M)/ SigLIP (400M)/ DINOv2 (300M)/ OpenCLIP (800M)	**8B**,13B,34B	2024.06
PKG-Transformer	Remote Sensing	Transformer	ResNet+Faster R-CNN	**32M?** (no checkpoint)	TGRS2023
MG-Transformer	Remote Sensing	Transformer	CLIP + ResNet-152	**38M?** (no checkpoint)	TGRS2024
BITA	Remote Sensing	OPT (2.7B)	CLIP (300M)	**~3B**	TGRS2024

CogVLM 2

- Developed by Tsinghua University and Zhipu AI
- Checkpoint accessed here





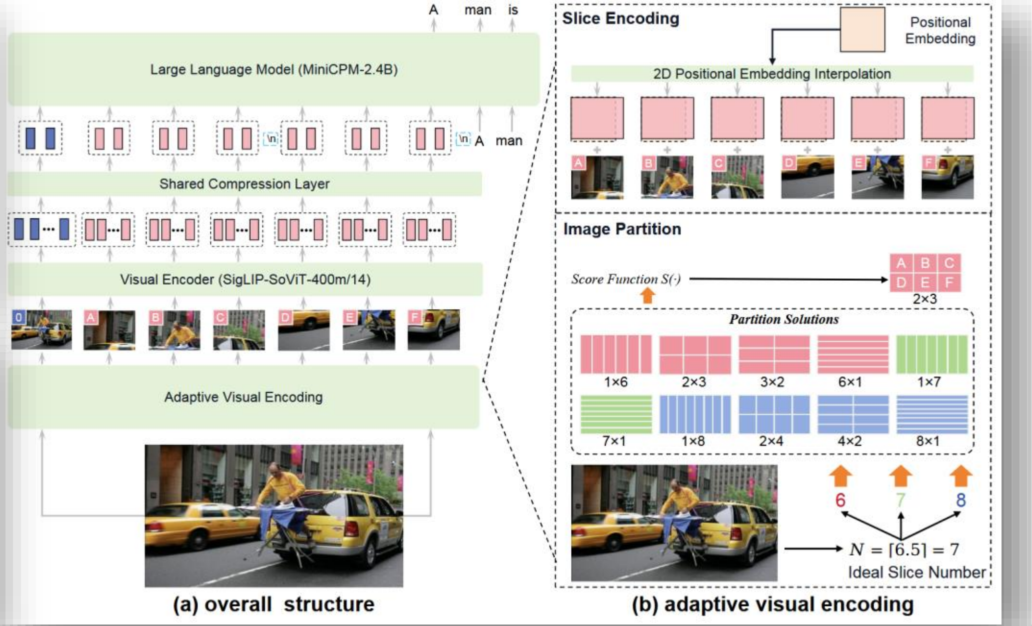


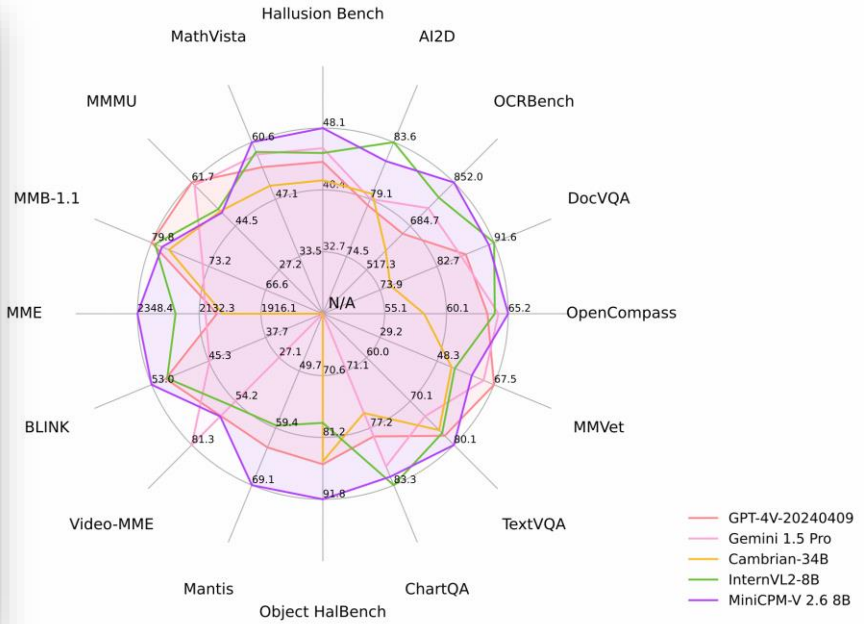
Left: CogVLM image captioning example. Mid: Architecture of CogVLM. Right: Evaluation of CogVLM.

W. Wang et al., 'CogVLM: Visual Expert for Pretrained Language Models', Feb. 04, 2024, arXiv: arXiv:2311.03079. doi: 10.48550/arXiv.2311.03079.

MiniCPM-V 2.6

- Developed by OpenBMB
- Checkpoint accessed [here]



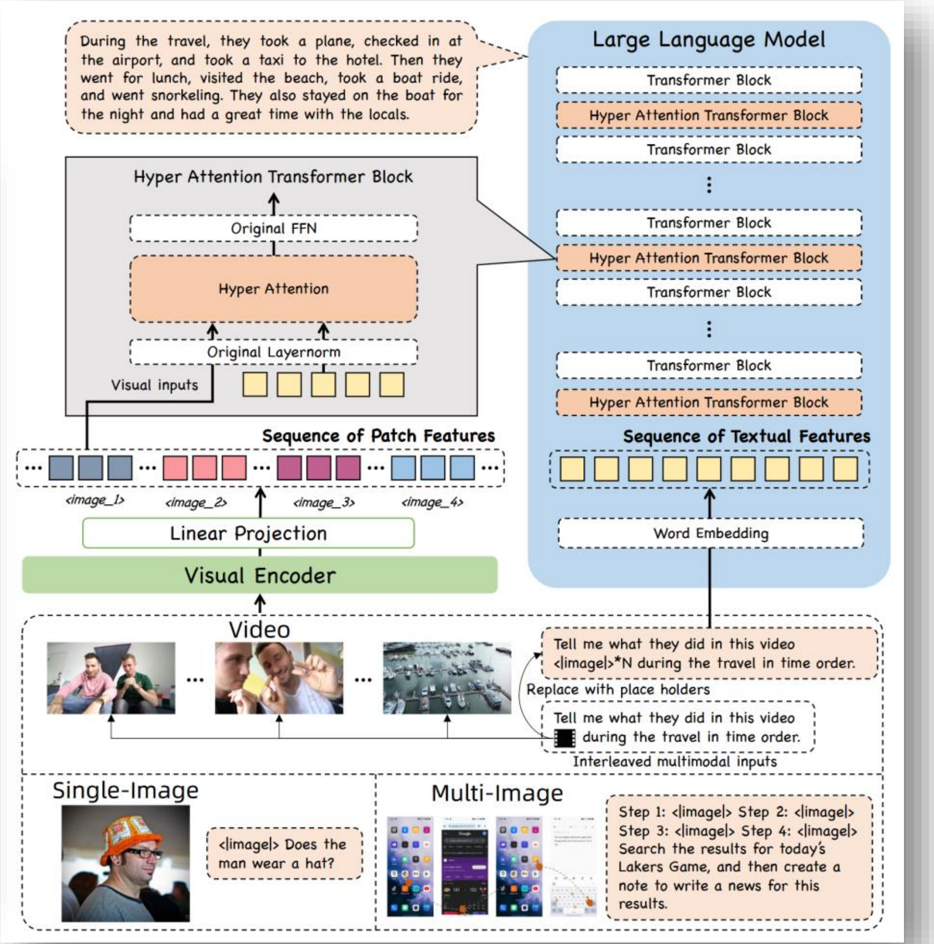


Left: MiniCPM-V 2.6 image captioning example. Mid: Architecture of MiniCPM-V. Right: Evaluation of MiniCPM-V 2.6.

Y. Yao et al., 'MiniCPM-V: A GPT-4V Level MLLM on Your Phone', Aug. 03, 2024, arXiv: arXiv:2408.01800. doi: 10.48550/arXiv.2408.01800.

mPLUG-Owl3

- Developed by Alibaba Group
- Checkpoint accessed here

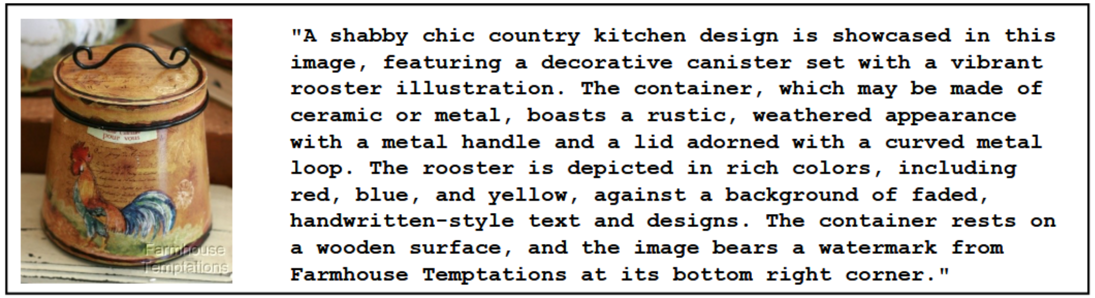


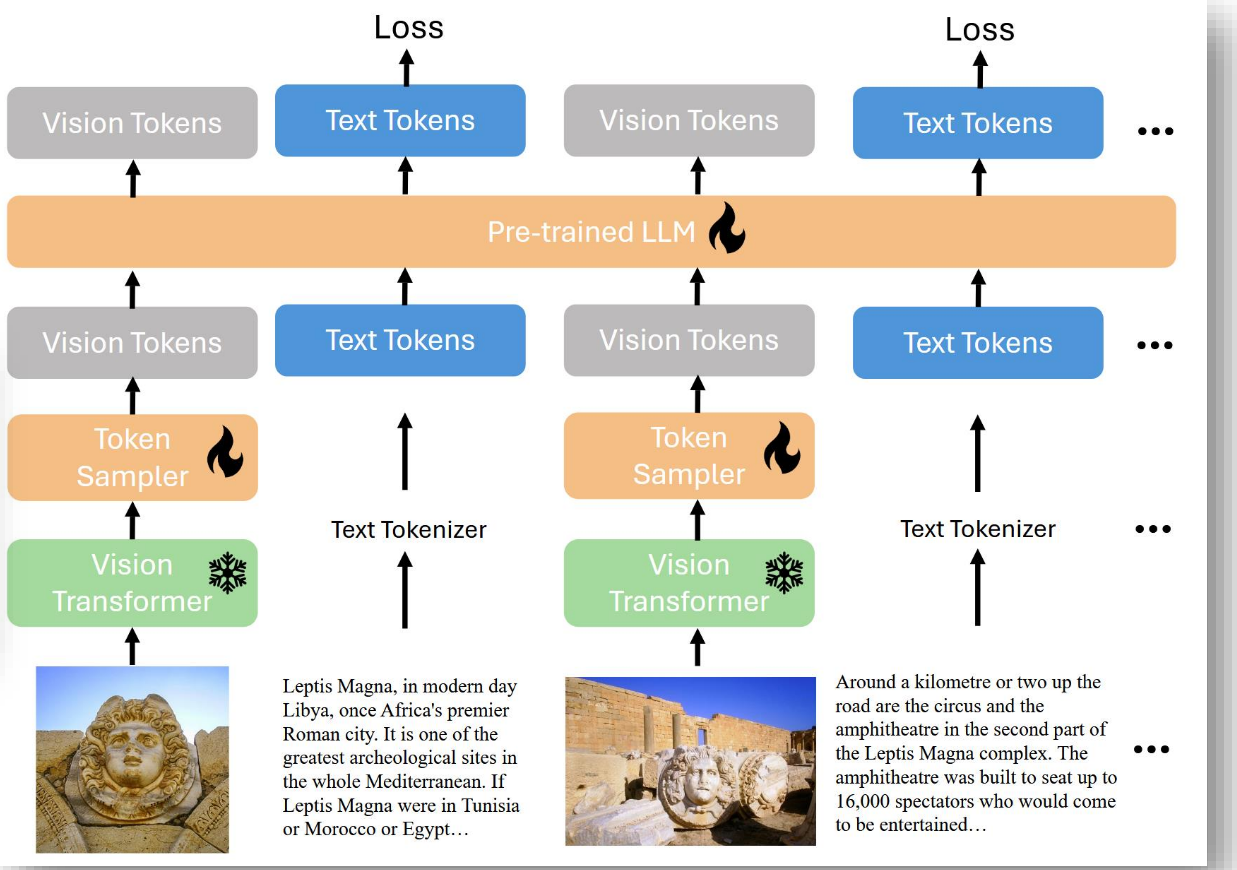
Left: mPLUG-Owl3 multi-images understanding example. Right: Architecture of mPLUG-Owl3.

J. Ye et al., 'mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models', Aug. 13, 2024, arXiv: arXiv:2408.04840.

xGen-MM (BLIP-3)

- Developed by Salesforce Research
- Checkpoint accessed here



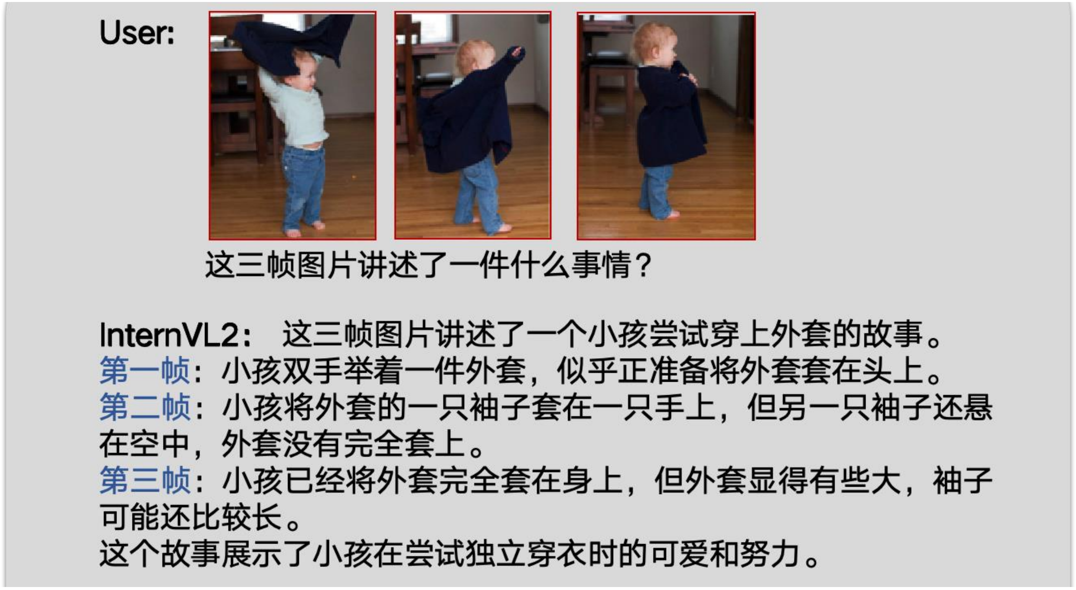


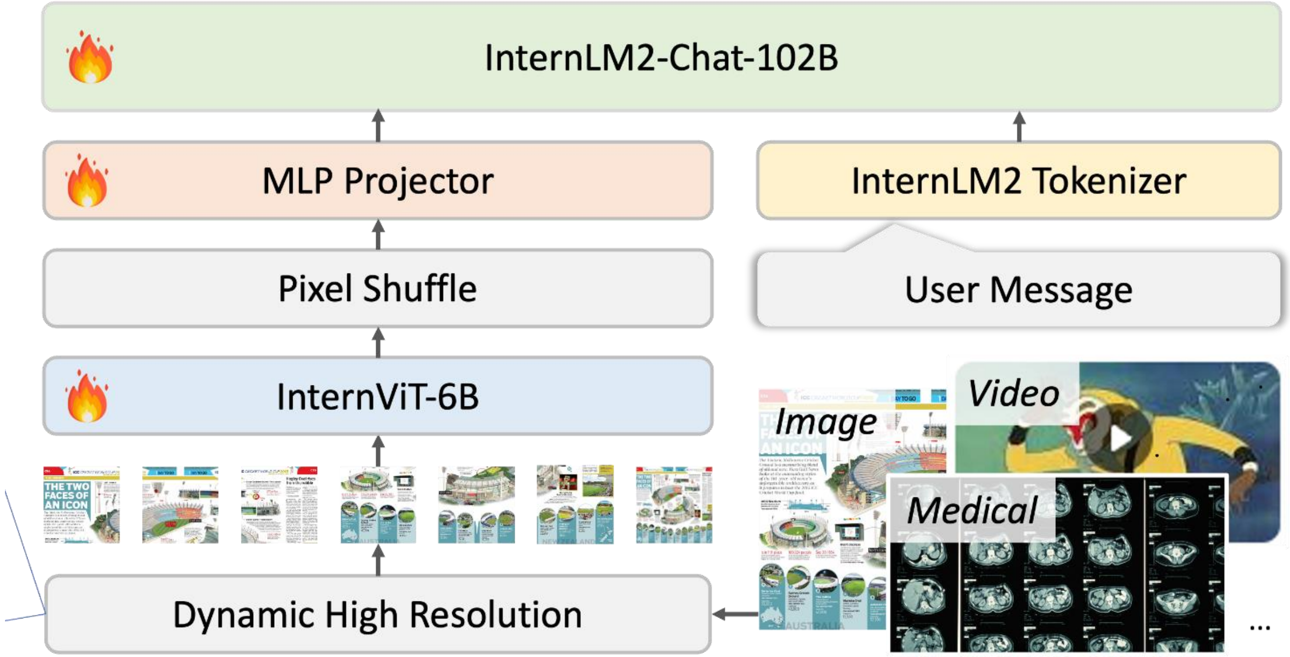
Left: BLIP-3 image captioning example. Right: Architecture of BLIP-3.

L. Xue et al., "xGen-MM (BLIP-3): A Family of Open Large Multimodal Models," Aug. 16, 2024, arXiv:2408.08872. doi: 10.48550/arXiv.2408.08872.

InternVL 2

- Developed by OpenGVLab, Shanghai AI Laboratory
- Checkpoint accessed here



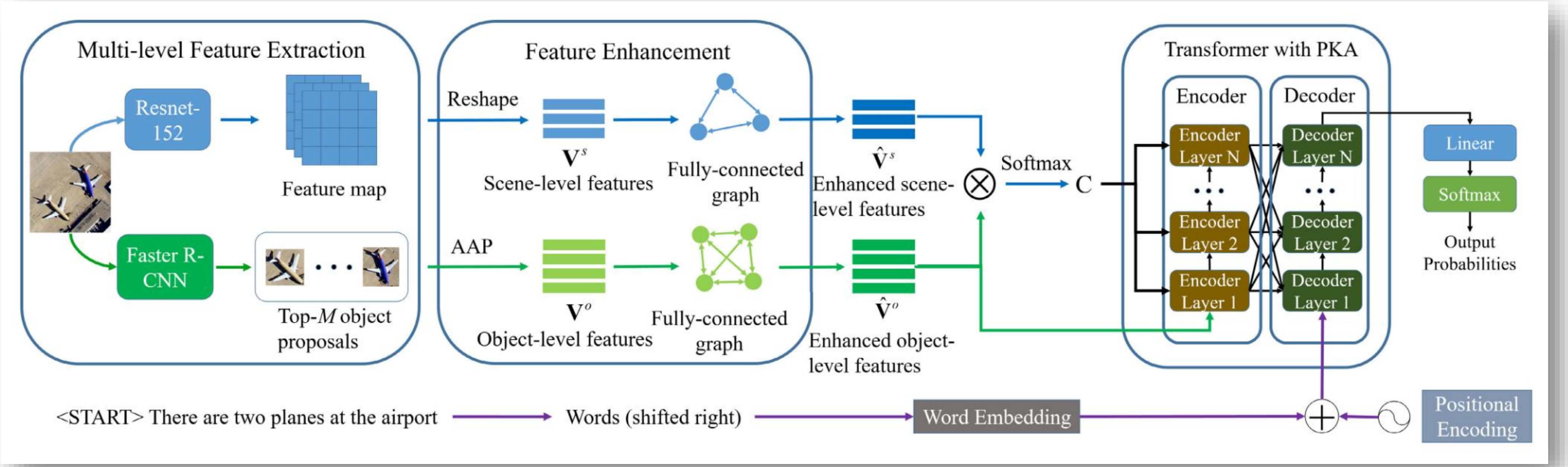


Left: InternVL 2 video understanding example. Right: Architecture of InternVL 2.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu,B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, “InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24 185–24 198.
Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma et al., “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” arXiv preprint arXiv:2404.16821, 2024.

PKG-Transformer & MG-Transformer

- PKG-Transformer (TGRS 2023) checkpoint accessed here
- MG-Transformer (TGRS 2024) checkpoint accessed here

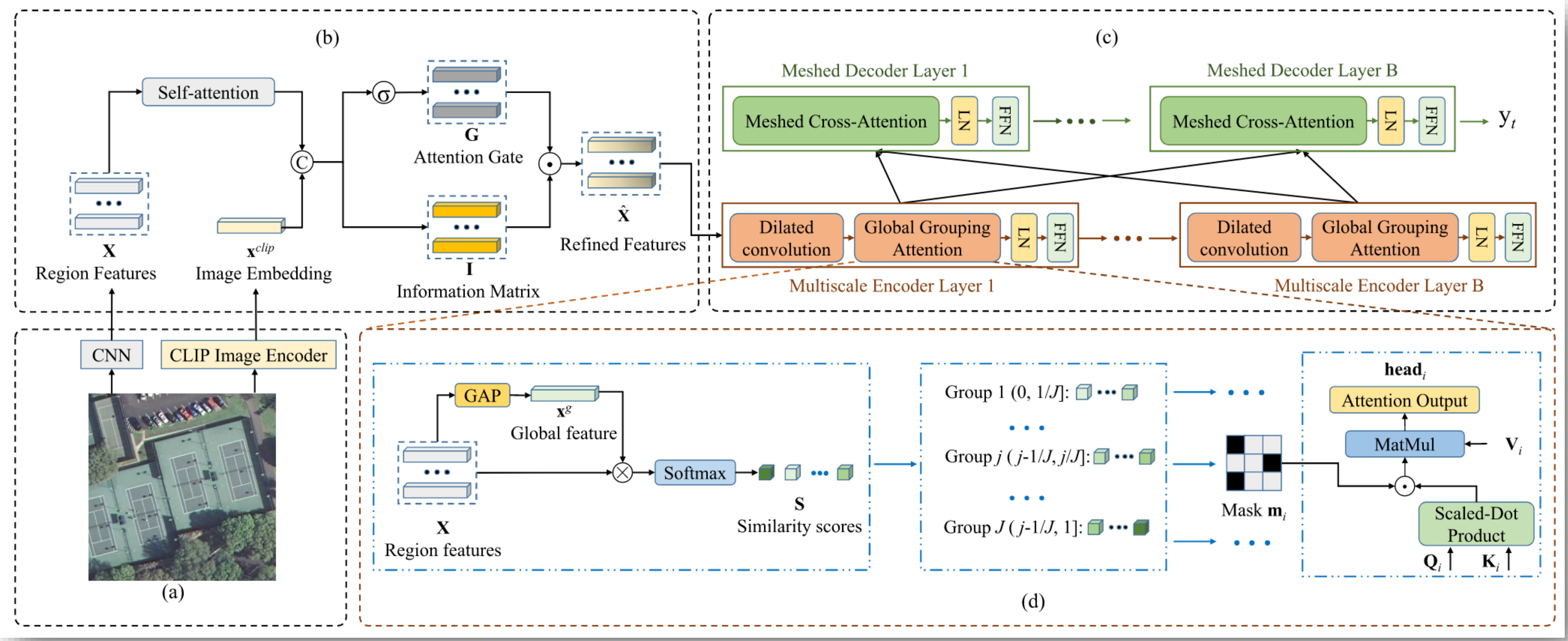


Framework of PKG-Transformer.

L. Meng, J. Wang, Y. Yang, and L. Xiao, 'Prior Knowledge-Guided Transformer for Remote Sensing Image Captioning', IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–13, Aug. 2023, doi: 10.1109/TGRS.2023.3328181.

PKG-Transformer & MG-Transformer

- PKG-Transformer (TGRS 2023) checkpoint accessed here
- MG-Transformer (TGRS 2024) checkpoint accessed here

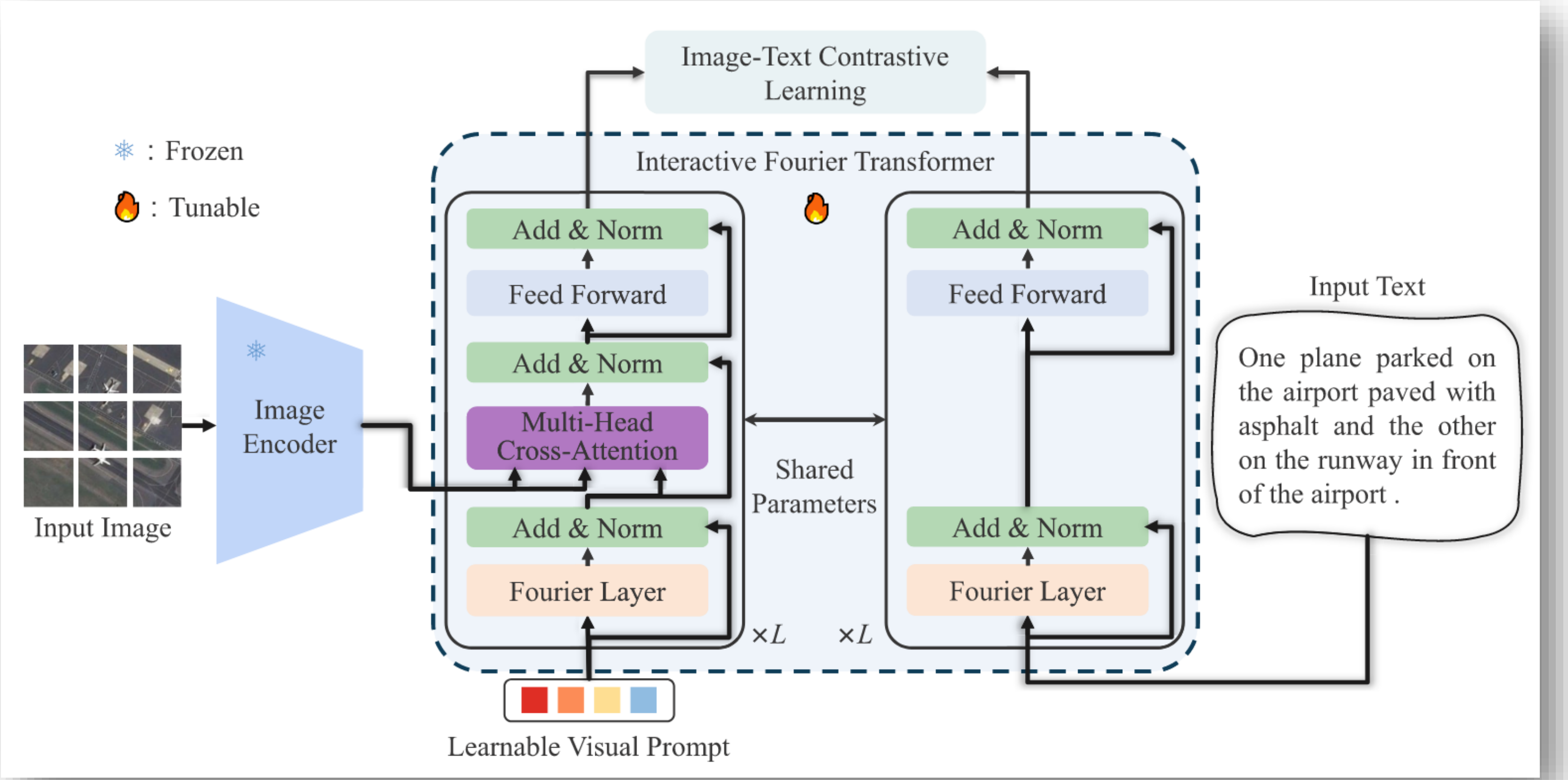


Framework of MG-Transformer.

L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, 'A Multiscale Grouping Transformer With CLIP Latents for Remote Sensing Image Captioning', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–15, Apr. 2024, doi: 10.1109/TGRS.2024.3385500.

Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning

- BITA checkpoint accessed here

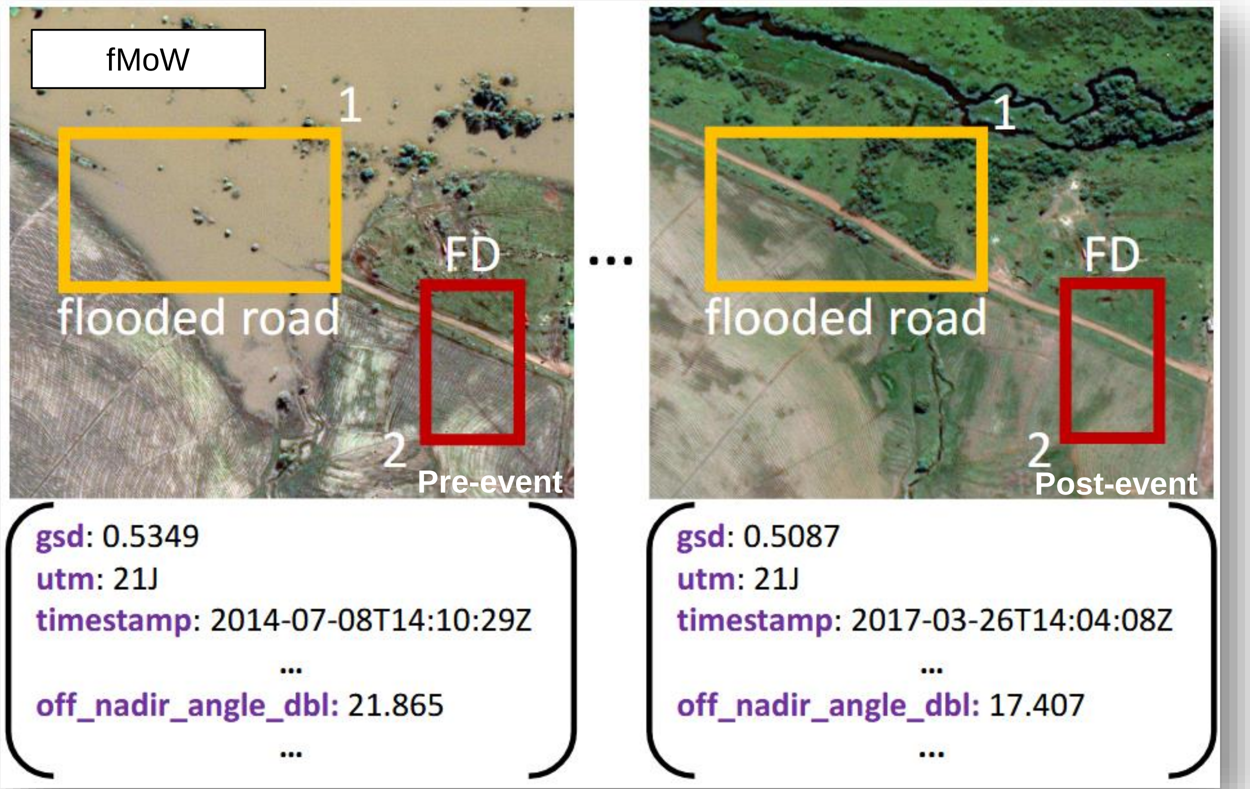


Framework of BITA.

C. Yang, Z. Li, and L. Zhang, 'Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning', TGRS 2024, vol. 62, pp. 1–12, Jan. 2024, doi: 10.1109/TGRS.2024.3359316.

Dataset

Dataset	Types of Images	Resolution of images	GSD	Pixels	Object Bounding Box	Semantic Segmentation	Caption/VQA
fMoW (2018)	Temporal images	Varies in size	~0.5m	437B	√	×	×
xBD (2019)	Pre and Post Disaster Images	1024×1024	~0.5m	~2B	√	×	×
SpaceNet 8 (2022)	Pre and Post Disaster Images	Varies in size	~0.5m	~0.3B	√	×	×



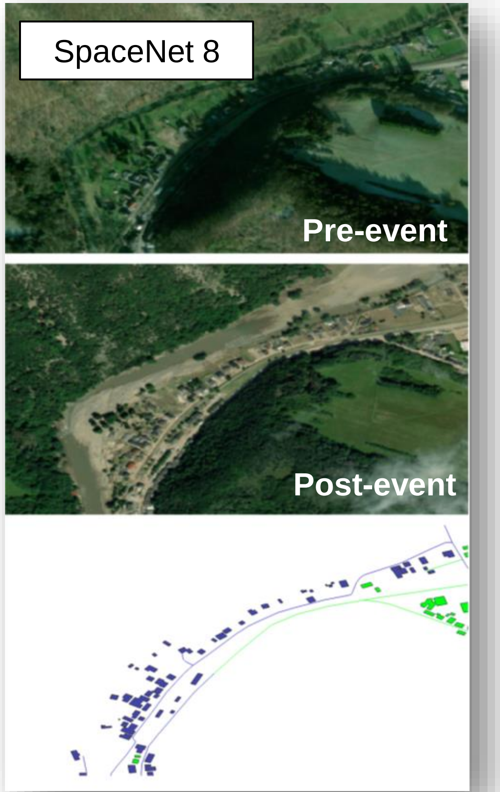


Image Editing

Text-to-Image

Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

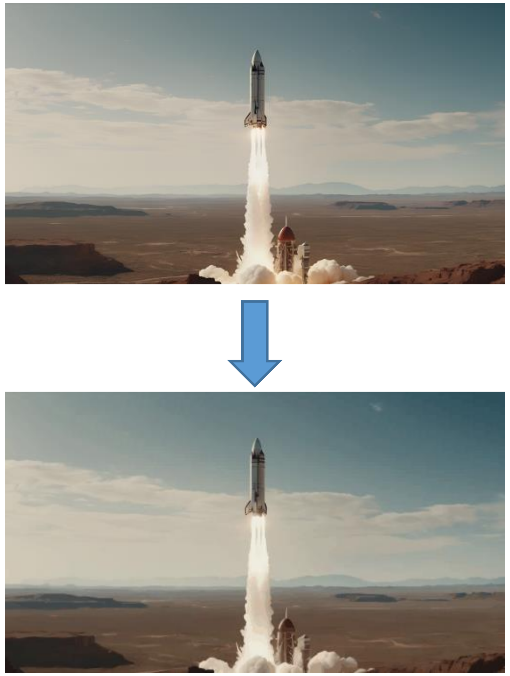
Image-to-Image

Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

Inpainting

concept art digital painting of an elven castle, inspired by lord of the rings, highly detailed, 8k

Text/Image-to-Video

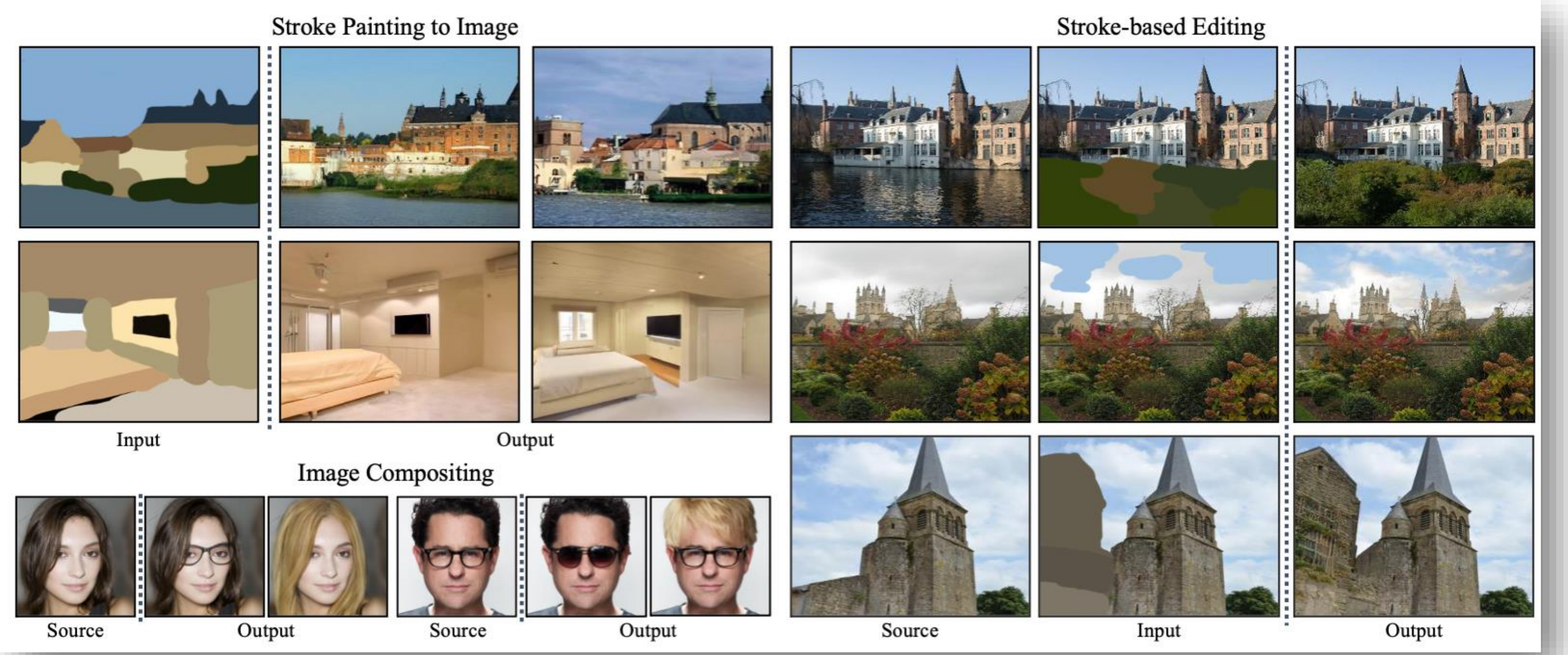


fps=7

All examples derive from [huggingface]*. From left to right, images are generated by SD 1.5, SD 1.5, SD 1.5 Inpainting and SVD XT respectively.

Stochastic Differential Editing (SDEdit)

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)



Stochastic Differential Editing (SDEdit) is a **unified** image synthesis and editing framework based on stochastic differential equations. SDEdit allows stroke painting to image, image compositing, and stroke-based editing **without** task-specific model training and loss functions.

C. Meng et al., 'SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations', presented at the International Conference on Learning Representations, Oct. 2021.

Stochastic Differential Editing (SDEdit)

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)

initial image

generated image

Stable Diffusion v1.5 with SDEdit technique for (better?) image editing.

(Example is derived from huggingface.co/docs/diffusers.*)

C. Meng et al., 'SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations', presented at the International Conference on Learning Representations, Oct. 2021.

CycleDiffusion

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)

Examples of CycleDiffusion for zero-shot image editing. Within each pair of source and target texts, overlapping text spans are marked in purple in the source text and abbreviated as [. . .] in the target text. Visual comparison to the baselines, DDIB and SDEdit.

C. H. Wu and F. De La Torre, 'A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance', in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 7344–7353. doi: 10.1109/ICCV51070.2023.00678.

InstructPix2Pix

- Fine-tuned on SD 1.5
- InstructPix2Pix checkpoint is accessible [here]

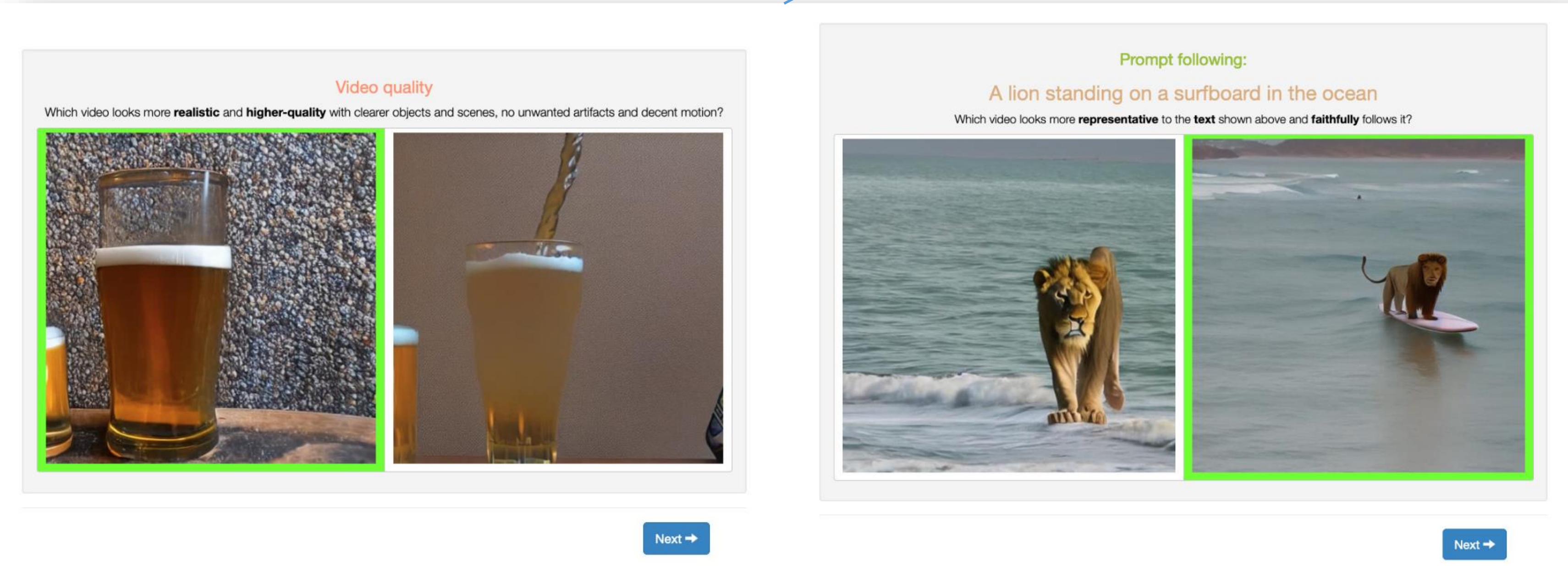
Given an image and an instruction for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

T. Brooks, A. Holynski, and A. A. Efros, 'InstructPix2Pix: Learning To Follow Image Editing Instructions', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18392–18402.

Stable Video Diffusion (SVD)

- Developed by Stability AI
- Start from SD 2.1 checkpoint

Post-Training (e.g. reinforce leaning from human preference)



(a) Sample instructions for evaluating visual quality of videos.

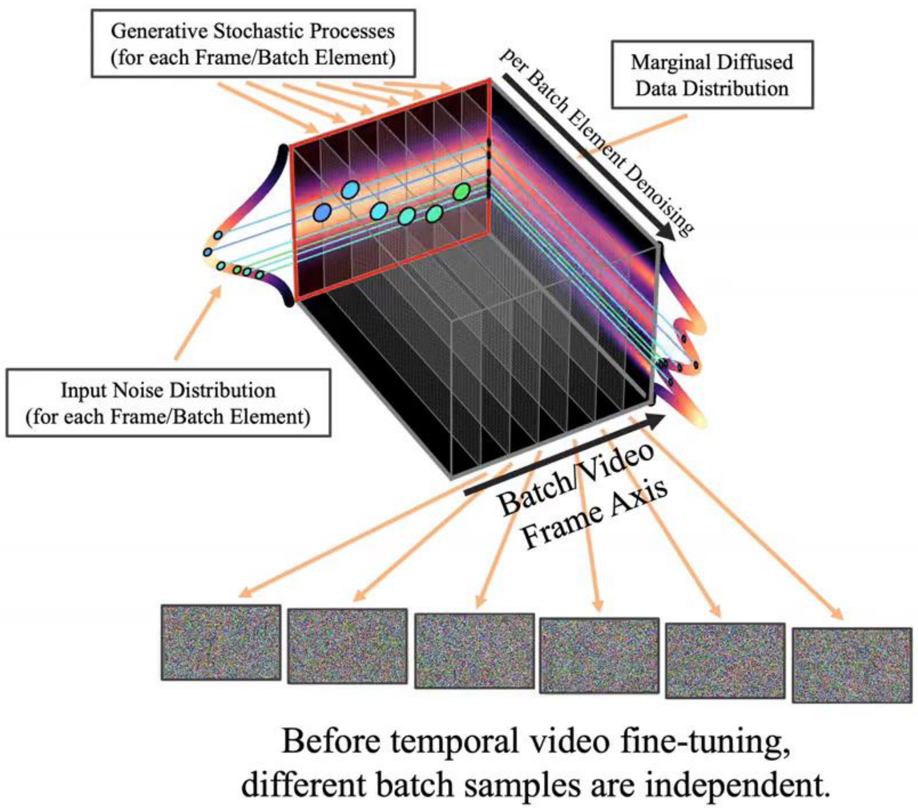
(b) Sample instructions for evaluating the prompt following of videos.

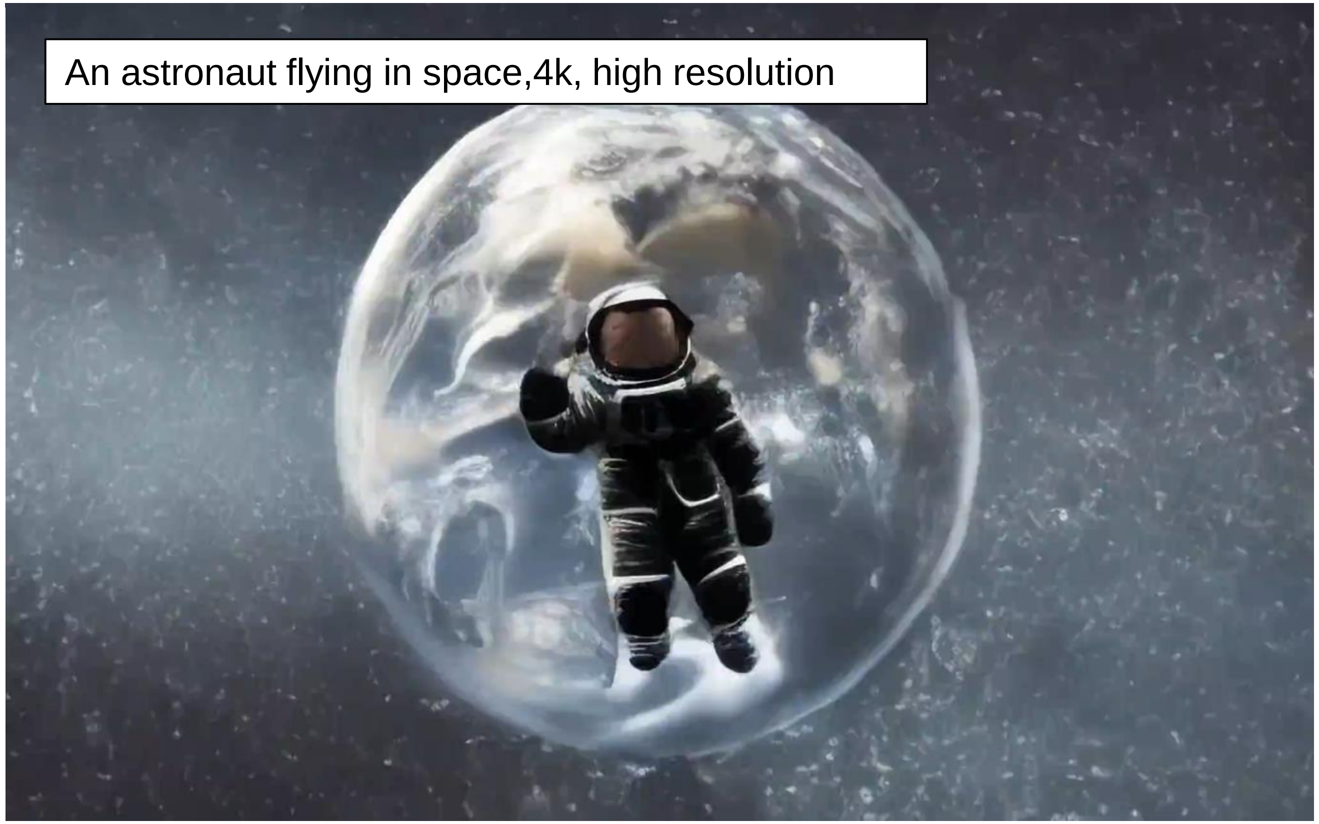
Left: Our human evaluation framework, as seen by the annotators. The prompt & task order and model choices are fully randomized. Right: Image-to-video examples.

[6] A. Blattmann et al., 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', Nov. 25, 2023, arXiv: arXiv:2311.15127. doi: 10.48550/arXiv.2311.15127.

Video LDM

- Developed by NVIDIA



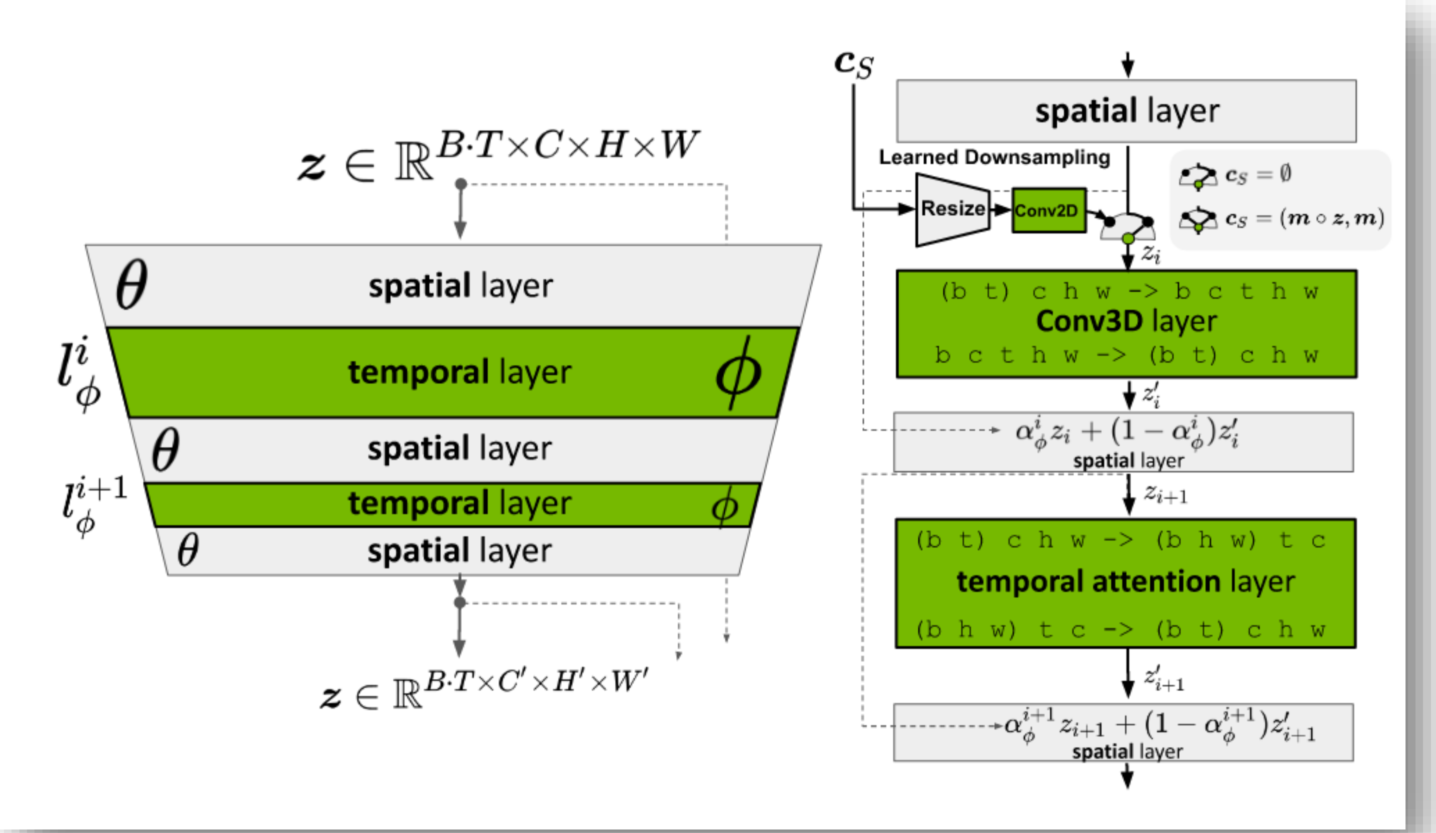


Left: Animation of temporal video fine-tuning in our Video Latent Diffusion Models (Video LDMs). Right: A example from Video LDMs.

A. Blattmann et al., 'Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02161.

Video LDM

- Developed by NVIDIA



Temporal block design in Video LDM.

Left: We turn a pre-trained LDM into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l^i_ϕ are trained. Right: During training, the base model θ interprets the input sequence of length T as a batch of images. For the temporal layers l^i_ϕ, these batches are reshaped into video format. Their output **z'** is combined with the spatial output z, using a learned merge parameter α. During inference, skipping the temporal layers (a^i_ϕ =1) yields the original image model. For illustration purposes, only a single U-Net Block is shown. B denotes batch size, T sequence length, C input channels and H and W the spatial dimensions of the input. c_s is optional context frame conditioning, when training prediction models.

A. Blattmann et al., 'Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02161.

Evaluation

图像生成质量评估

视觉质量指标：Fréchet Inception Distance (FID) 、Inception Score (IS) 和 CLIP-score

像素质量指标：SSIM、PSNR、LPIPS 和 MSE

人类偏好评估

专家打分，GPT打分

下游场景继续训练

场景分类

References and Related Works

[1] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas,H. Choset, and M. Gaston, “Creating xBD: A Dataset for Assessing Building Damage fromSatellite Imagery,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 10–17.
[2] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, “DiffusionSat: A Generative Foundation Model for Satellite Imagery,” in The TwelfthInternational Conference on Learning Representations, Oct. 2023.
[3] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” in Proceedings of the 32nd Interna-tional Conference on Machine Learning. PMLR, Jun. 2015, pp. 2256–2265.
[4] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in Advancesin Neural Information Processing Systems, vol. 33. Curran Associates, Inc., 2020, pp.6840–6851.
[5] A. Q. Nichol and P. Dhariwal, “Improved Denoising Diffusion Probabilistic Models,” in Proceedings of the 38th International Conference on Machine Learning. PMLR, Jul.2021, pp. 8162–8171.
[6] Imagen-Team-Google, “Imagen 3,” Aug. 2024.
[7] BlackForestLab, “Announcing Black Forest Labs,” Aug. 2024.
[8] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo,W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh, “Improving Image Generation with Better Captions,” 2023.
[9] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz,A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis,” in Forty-First InternationalConference on Machine Learning, Jun. 2024.
[10] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning,” in Proceedings of the56th Annual Meeting of the Association for Computational Linguistics (Volume 1: LongPapers), I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Compu-tational Linguistics, Jul. 2018, pp. 2556–2565.
[11] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts,” in Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3558–3568.
[12] V. Ordonez, G. Kulkarni, and T. Berg, “Im2Text: Describing Images Using 1 Million Captioned Photographs,” in Advances in Neural Information Processing Systems, vol. 24. Cur-ran Associates, Inc., 2011.
[13] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes,A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt,R. Kaczmarczyk, and J. Jitsev, “LAION-5B: An open large-scale dataset for training next generation image-text models,” in Advances in Neural Information Processing Systems,vol. 35, Dec. 2022, pp. 25 278–25 294.
[14] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, “Coyo-700m: Image-text paird ataset,” https://github.com/kakaobrain/coyo-dataset, 2022.
[15] B. Qu, X. Li, D. Tao, and X. Lu, “Deep semantic understanding of high resolution remote sensing image,” in 2016 International conference on computer, information and telecommunication systems (Cits). IEEE, 2016, pp. 1–5.
[16] J. Roberts, K. Han, and S. Albanie, “Satin: A multi-task metadataset for classifying satellite imagery using vision-language models,” arXiv preprint arXiv:2304.11619, 2023.
[17] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, “Graph convolutional net-works for hyperspectral image classification,” IEEE Transactions on Geoscience and Re-mote Sensing, vol. 59, no. 7, pp. 5966–5978, 2020.
[18] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images,” IEEE transactions on geoscience and remote sensing, vol. 54, no. 12, pp. 7405–7415, 2016.
[19] X. Yang, G. Zhang, X. Yang, Y. Zhou, W. Wang, J. Tang, T. He, and J. Yan, “Detecting rotated objects as gaussian distributions and its 3-d generalization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4335–4354, 2022.
[20] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, “Change detection based on deep siamese convolutional network for optical aerial images,” IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1845–1849, 2017.

References and Related Works

[21] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, “Changemamba: Remote sensing change detection with spatiotemporal state space model,” IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–20, 2024.
[22] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753, 2018.
[23] S. Lobry, D. Marcos, J. Murray, and D. Tuia, “Rsvqa: Visual question answering for remote sensing data,” IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 12, pp. 8555–8566, 2020.
[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., “Llama 2: Open foundation and fine-tuned chat models,” arXiv preprint arXiv:2307.09288, 2023.
[25] Llama Team,AI@Meta, “The llama 3 herd of models,” 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
[26] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in Neural Information Processing Systems, May 2020.
[27] “Introducing ChatGPT,” https://openai.com/index/chatgpt/.
[28] OpenAI, “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
[29] OpenAI, “Hello GPT-4o,” https://openai.com/index/hello-gpt-4o/.
[30] K. He, X. Chen, S. Xie, Y. Li, P. Doll ár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 000–16 009.
[31] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, “SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery,” in Advances in Neural Information Processing Systems, vol. 35, Dec. 2022, pp. 197–211.
[32] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyt- tendaele, and T. Darrell, “Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4088–4099.
[33] M. Tang, A. L. Cozma, K. Georgiou, and H. Qi, “Cross-Scale MAE: A Tale of Multiscale Exploitation in Remote Sensing,” in Thirty-Seventh Conference on Neural Information Processing Systems, Nov. 2023.
[34] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, “RingMo: A Remote Sensing Foundation Model With Masked Image Modeling,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–22, 2023.
[35] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, “SpectralGPT: Spectral Remote Sensing Foundation Model,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–18, 2024.
[36] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, H. He, J. Wang, J. Chen, M. Yang, Y. Zhang, and Y. Li, “SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery,” in Computer Vision and Pattern Recognition (CVPR). arXiv, Mar. 2024.
[37] Z. Yu, C. Liu, L. Liu, Z. Shi, and Z. Zou, “MetaEarth: A Generative Foundation Model for Global-Scale Remote Sensing Image Generation,” May 2024.
[38] R. Ou, H. Yan, M. Wu, and C. Zhang, “A Method of Efficient Synthesizing Post-disaster Remote Sensing Image with Diffusion Model and LLM,” in 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Oct. 2023, pp. 1549–1555.
[39] D. Tang, X. Cao, X. Hou, Z. Jiang, and D. Meng, “CRS-Diff: Controllable Generative Remote Sensing Foundation Model,” Jun. 2024.

References and Related Works

[40] A. Sebaq and M. ElHelw, “RSDiff: Remote Sensing Image Generation from Text Using Diffusion Model,” Sep. 2023.
[41] M. Espinosa and E. J. Crowley, “Generate Your Own Scotland: Satellite Image Generation Conditioned on Maps,” Aug. 2023.
[42] Z. Yuan, C. Hao, R. Zhou, J. Chen, M. Yu, W. Zhang, H. Wang, and X. Sun, “Efficient and Controllable Remote Sensing Fake Sample Generation Based on Diffusion Model,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–12, 2023.
[43] O. Baghirli, H. Askarov, I. Ibrahimli, I. Bakhishov, and N. Nabiyev, “SatDM: Synthesizing Realistic Satellite Image with Semantic Layout Conditioning using Diffusion Models,” Sep. 2023.
[44] C. Zhao, Y. Ogawa, S. Chen, Z. Yang, and Y. Sekimoto, “Label Freedom: Stable Diffusion for Remote Sensing Image Semantic Segmentation Data Generation,” in 2023 IEEE International Conference on Big Data (BigData), Dec. 2023, pp. 1022–1030.
[45] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” in Advances in Neural Information Processing Systems, vol. 35, Dec. 2022, pp. 36 479–36 494.
[46] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh, “Improving Image Generation with Better Captions,” 2023.
[47] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, “CogVLM: Visual Expert for Pretrained Language Models,” Feb. 2024.
[48] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He et al., “Minicpm-v: A gpt-4v level mllm on your phone,” arXiv preprint arXiv:2408.01800, 2024.
[49] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, “mplug-owl3: Towards long image-sequence understanding in multi-modal large language models,” arXiv preprint arXiv:2408.04840, 2024.
[50] L. Meng, J. Wang, Y. Yang, and L. Xiao, “Prior knowledge-guided transformer for remote sensing image captioning,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–13, Aug 2023.
[51] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, “A multiscale grouping transformer with clip latents for remote sensing image captioning,” IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–15, Apr 2024.
[52] C. Yang, Z. Li, and L. Zhang, “Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning,” IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–12, Jan. 2024.
[53] A. Vahdat, K. Kreis, and J. Kautz, “Score-based generative modeling in latent space,” Advances in neural information processing systems, vol. 34, pp. 11 287–11 302, 2021.
[54] A. Sinha, J. Song, C. Meng, and S. Ermon, “D2c: Diffusion-decoding models for few-shot conditional generation,” Advances in Neural Information Processing Systems, vol. 34, pp. 12 533–12 548, 2021.
[55] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.
[56] S. Yin, C. Fu, S. Zhao, T. Xu, H. Wang, D. Sui, Y. Shen, K. Li, X. Sun, and E. Chen, “Woodpecker: Hallucination Correction for Multimodal Large Language Models,” Oct. 2023.
[57] R. Hänsch, J. Arndt, D. Lunga, M. Gibb, T. Pedelose, A. Boedihardjo, D. Petrie, and T. M. Bacastow, “SpaceNet 8 - The Detection of Flooded Roads and Buildings,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1472–1480.
[58] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, “Functional Map of the World,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6172–6180.
[59] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023, pp. 3813–3824.
[60] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, “Adversarial Diffusion Distillation,” Nov. 2023.
[61] L. Xue et al., “xGen-MM (BLIP-3): A Family of Open Large Multimodal Models,” Aug. 16, 2024, arXiv:2408.08872. doi: 10.48550/arXiv.2408.08872.

References and Related Works

[61] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” in The Twelfth International Conference on Learning Representations, Oct. 2023.
[62] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., 2017.
[63] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, “Improved Techniques for Training GANs,” in Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc., 2016.
[64] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, “CLIPScore: A Reference-free Evaluation Metric for Image Captioning,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7514–7528.
[65] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[67] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “AlignYour Latents: High-Resolution Video Synthesis with Latent Diffusion Models,” in 2023IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver,BC, Canada: IEEE, Jun. 2023, pp. 22 563–22 575.
[68] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, “Uni-ControlNet:All-in-One Control to Text-to-Image Diffusion Models,” in Neural Information ProcessingSystems, vol. 36. Curran Associates, Inc., 2023, pp. 11 127–11 150.
[69] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, “ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback,” Apr. 2024.
[70] B. Peng, J. Wang, Y. Zhang, W. Li, M.-C. Yang, and J. Jia, “ControlNeXt: Powerful and Efficient Control for Image and Video Generation,” Aug. 2024.
[71] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Dif-fusion Models,” in 2023 IEEE/CVF International Conference on Computer Vision (ICCV),Oct. 2023, pp. 3813–3824.