

AI Security

Large Multimodal Models (LMM)

Sangdon Park

POSTECH

April 9, 2026

Outline

- 1 Introduction
- 2 Diffusion Models
- 3 Vision-Language Models
- 4 Vision-Language-Action Models
- 5 Conclusion

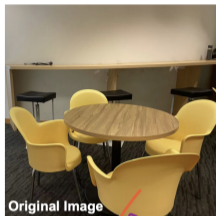
Motivation: LMMs

Why LMMs?

An LMM is a clearly richer model than an LLM.

- LLM: text \rightarrow text
- LMM:
 - ▶ (text, image) \rightarrow text
 - ▶ text \rightarrow image
 - ▶ ...

Motivation: Attacks on LMMs



Question: How many black chairs are there in the image?

Response from ChatGPT4o

Answer to Original Image:
There are **three** black chairs.

Answer to Digital SceneTAP:
There are **two** black chairs.

Answer to Physical SceneTAP:
There are **two** black chairs.



Generation

Print & Paste

- typographic attacks [Cao et al., 2025]

Outline

- 1 Introduction
- 2 Diffusion Models**
- 3 Vision-Language Models
- 4 Vision-Language-Action Models
- 5 Conclusion

Overview: Diffusion Models

Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads
"Latent Diffusion"'

'A zombie in the
style of Picasso'

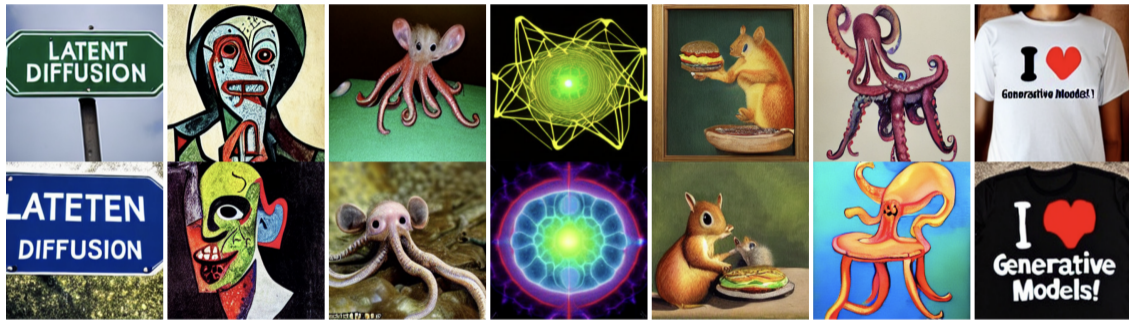
'An image of an animal
half mouse half octopus'

'An illustration of a slightly
conscious neural network'

'A painting of a
squirrel eating a burger'

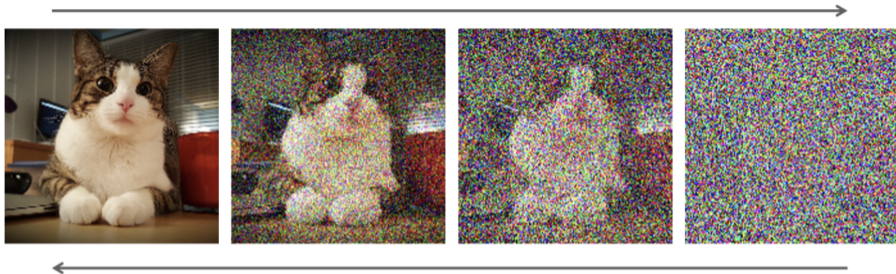
'A watercolor painting of a
chair that looks like an octopus'

'A shirt with the inscription:
"I love generative models!"'



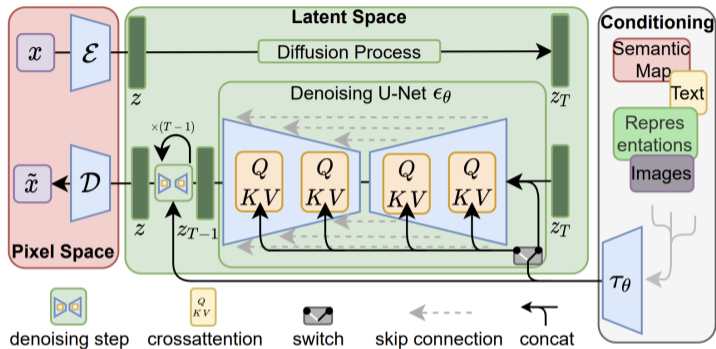
- A kind of density estimators along with VAE, GAN, and normalizing flow.
- Diffusion models achieve amazing performance.
- Useful for text-to-image generation

Diffusion and Denoising Processes with Examples



- → direction: diffusion (or noising) process.
 - ▶ Add small noise at each step
 - ▶ Being used in a learning phase
- ← direction: denoising process.
 - ▶ Remove small noise at each step
 - ▶ Being used in a generation phase

Latent Diffusion Model Architecture [Rombach et al., 2022]



- Consists of two main parts:
 - 1 Learning: a pipeline for diffusion processes
 - 2 Generation: a model for denoising processes
- latent space vs image space?
- Conditioning via cross-attention

Generation Step of Diffusion Models

Generation Step

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_{\theta}(z_t, t, c) \right) + \sigma_t \varepsilon$$

- generation = denoising process
- ε_{θ} : a denoising model
- $\varepsilon \sim \mathcal{N}(0, I)$: noise
- c : a condition (e.g., text embedding vectors)
- α_t : a noise schedule parameter
- $\bar{\alpha}_t := \prod_{s=t}^T \alpha_s$
- But, all we need is...this generation step is differentiable w.r.t. c .
 - ▶ assume that an attacker injects an adversarial prompt to generate a harmful image.

Outline

- 1 Introduction
- 2 Diffusion Models
- 3 Vision-Language Models**
- 4 Vision-Language-Action Models
- 5 Conclusion

Overview: Vision-Language Models (VLMs)



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

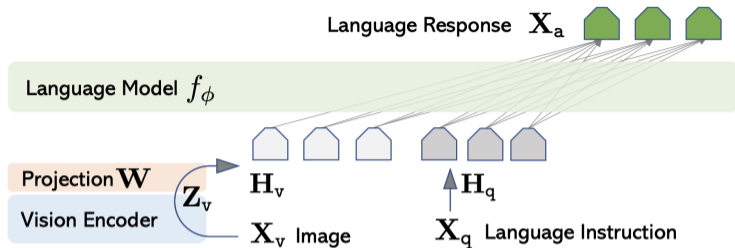
User
LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

- Often we use VLMs for text generation from an image and an instruction.
- VLMs consist of (roughly) three modules:
 - 1 a vision encoder
 - 2 a language model
 - 3 a fusion mechanism

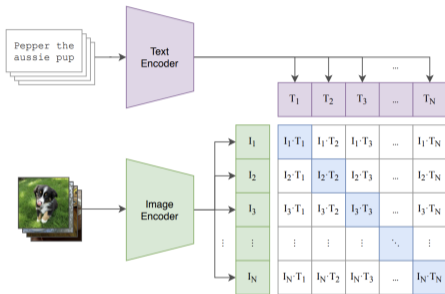
LLaVA [Liu et al., 2023]



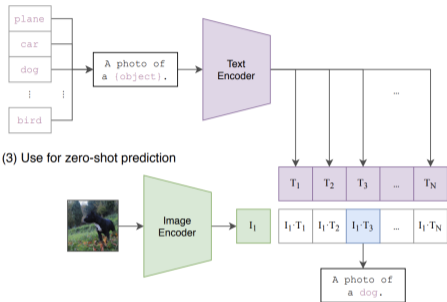
- Input:
 - ▶ image: convert an image into “visual” tokens, and then converted to a list of token vectors
 - ▶ text: converted the same way as LLMs.
- Output:
 - ▶ text: generated in the same way as LLMs.
- Models:
 - ▶ Visual encoder: CLIP image encoder (ViT-L/14)
 - ▶ Projection: a linear layer
 - ▶ Language model: Vicuna

CLIP [Radford et al., 2021]

(1) Contrastive pre-training



(2) Create dataset classifier from label text



- CLIP jointly trains an image and a text encoder.
- High cosine similarity between a visual token and text token represents their correlation.
- CLIP can be used as a classifier.

Reference I

- Y. Cao, Y. Xing, J. Zhang, D. Lin, T. Zhang, I. Tsang, Y. Liu, and Q. Guo. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25050–25059, 2025.
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- 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*, pages 10684–10695, 2022.