Vision-Language Modeling:

Representation, Generation, and the Road to Multimodal Reasoning

Yichao Cai

Lecture 10, Statistical Machine Learning, Sem. 2, 2025

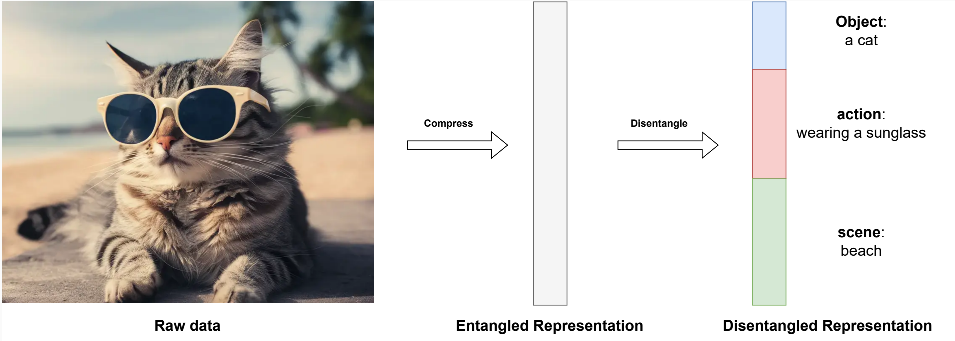
School of Computer and Mathematical Sciences, The University of Adelaide

Contents

Backgrounds on Representation Learning

What is a Representation?

- Raw data is high-dimensional & unstructured.
- Representation = transformed, structured form.
- Captures meaningful semantic factors.
- Enables learning and generalization.



Why Learn Representations?

- Reduce dimensionality.
- Remove noise and redundancy.
- Capture underlying structure.
- Enable transfer across tasks.
- Support interpretability.

Predictive Tasks

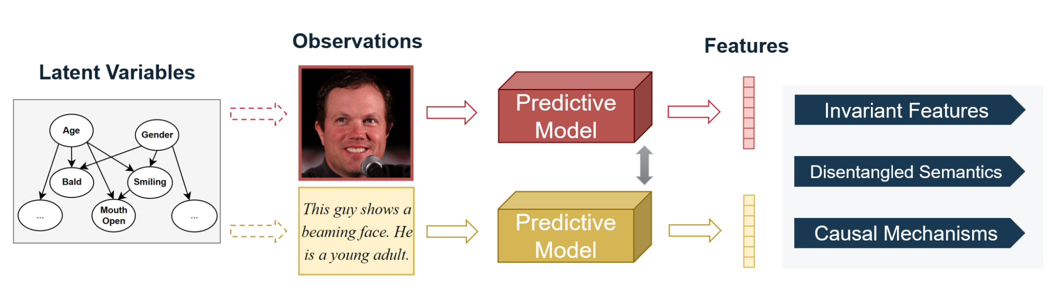
THE UNIVERSITY of ADELAIDE

- Goal: predict target from input.
- Examples:
  - Classification
  - Regression
  - Detection / Segmentation
  - Or even, complex reasoning tasks
- Representation must be **task-relevant** and **environment-shift robust**.
- Evaluated by accuracy / error.

Generative Tasks

- Goal: model or generate data.
- Examples:
    - Image / text generation
    - Inpainting, super-resolution
    - Diffusion / VAE / GANs
- Representation should capture **data structure**, if we desire contraollability.
- Evaluated by fidelity, diversity, likelihood.

Hypothesis – A Generative Perspective: Image–text pairs (or generally speaking, observations) are **generated** from a **shared latent space** (scene semantics, attributes, viewpoint, linguistic intent); our goal is to learn representations that **align with** this latent space.



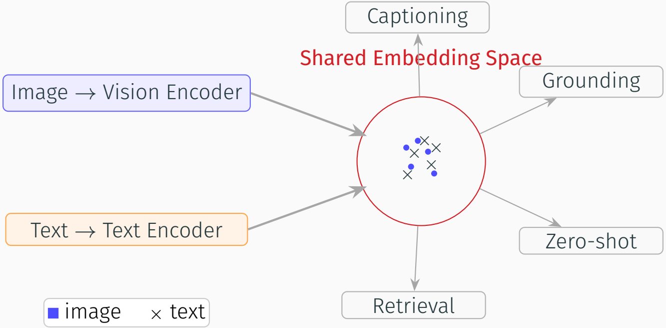
Human face image from the CelebA-HQ dataset.

Motivation of Vision-Language Modeling

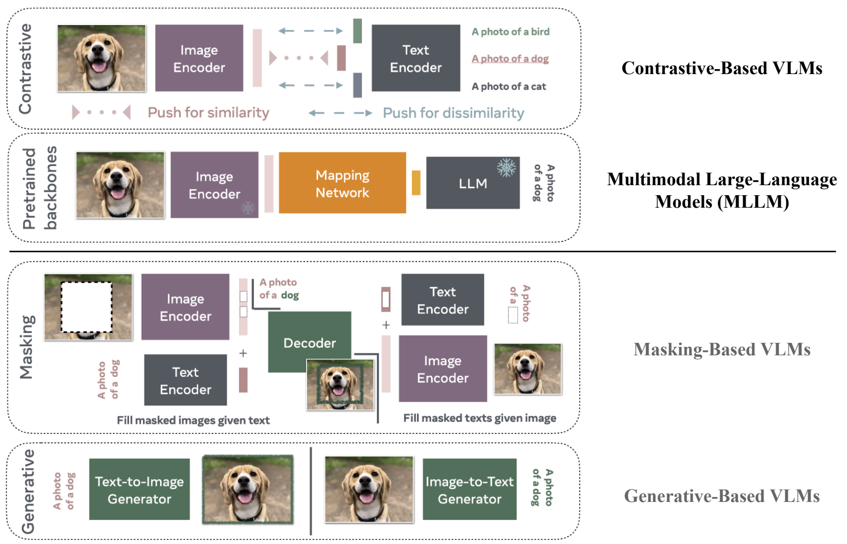
Why Vision-Language Models (VLMs)?

THE UNIVERSITY of ADELAIDE

- Machines see **pixels (i.e., unstructured data)** but often lack semantic understanding; Language provides **semantic structure and abstraction**.
- **Goal:** learn a meaningful, modality-shared representation, that **Enables** interpretable prediction & reasoning, e.g., retrieval, zero-shot, captioning, etc.



Families of VLMs



Florian Bordes, et al., 2024, An Introduction to Vision-Language Modeling.

Fusion vs. Alignment: Two Ways to Combine Modalities

Alignment (dual encoders; CLIP/ALIGN/SigLIP)

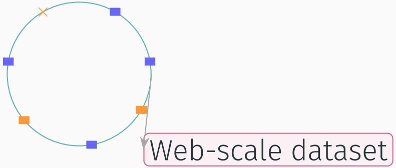
- Learn a *shared* image–text embedding space.
- Contrastive training (InfoNCE; multi-positive), web-scale.
- **Strengths:** highly scalable, fast inference, strong zero-shot & retrieval.
- **Limits:** needs paired data; weaker fine-grained grounding unless augmented.

Fusion (cross-attention; LLaVA/Qwen-VL)

- Joint modeling of image & text tokens (e.g., cross-attn / Q-Former).
- Trained with task supervision (VQA, captioning, grounding).
- **Strengths:** better reasoning, localization, stepwise answers.
- **Limits:** heavier compute/memory; less task-agnostic reuse.

Rule of thumb: Alignment for *breadth & zero-shot*. Fusion for *reasoning & grounding*.

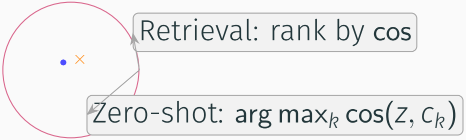
Key Ingredients of VLMs



(1) Large-scale web data

(3) Flexible architectures (ViT/Transformer)

(2) Contrastive or pretext objectives



(4) Meaningful evaluation, e.g., Zero-shot & VQA

Contrastive Vision-Language Foundation Models

CLIP: Learning from Web-scale Supervision

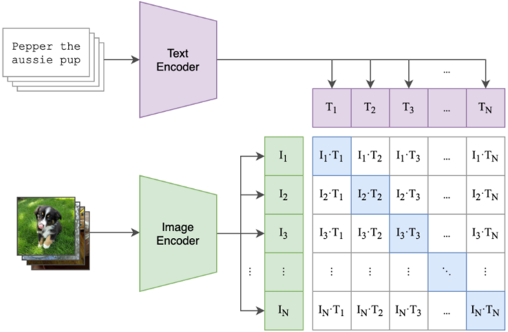
- Dual-encoder: ViT (or, CNNs) + text Transformer.
- Data: 400M image–text pairs.
- Trained with **symmetric contrastive loss**.
- Emergent zero-shot transfer.

high waist sleeveless mini soft jeans dress frilled women ruffles casual summer sundress short denim beach dress cotton

DragonSpeed reveals revised plans for IndyCar, IMSA, ELMS

Man on moving walkway at Dublin Airport

https://huggingface.co/datasets/laion/relaion400m



A. Radford, et al., 2021, CLIP

THE UNIVERSITY of ADELAIDE

Learning objective - Symmetric contrastive loss (InfoNCE):

Normalize: $\tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}$, $\tilde{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|}$, Feature similarity: $s_{ij} = \frac{\tilde{\mathbf{z}}_i^\top \tilde{\mathbf{t}}_j}{\tau}$

$$\mathcal{L}_{\mathrm{i}\to\mathrm{t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N}\exp(s_{ij})}, \qquad \mathcal{L}_{\mathrm{t}\to\mathrm{i}} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s_{jj})}{\sum_{i=1}^{N}\exp(s_{ij})}$$

$\boxed{\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{i}\to\mathrm{t}} + \mathcal{L}_{\mathrm{t}\to\mathrm{i}}\right)}$ with temperature τ (learned).

What CLIP Learns

Learning objective - Symmetric contrastive loss (InfoNCE):

$$\text{Normalize:}\quad \tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|},\ \tilde{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|},\qquad \text{Feature similarity:}\quad s_{ij} = \frac{\tilde{\mathbf{z}}_i^\top \tilde{\mathbf{t}}_j}{\tau}$$

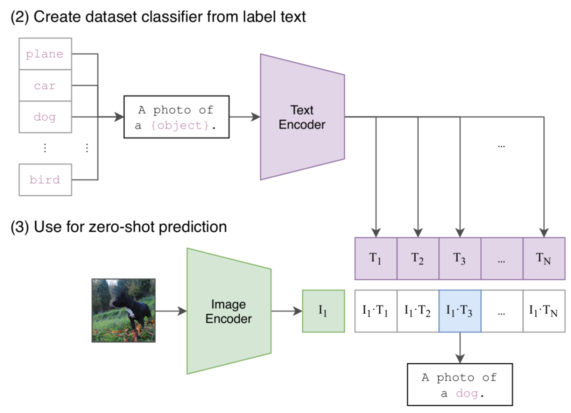
$$\mathcal{L}_{\mathrm{i}\to\mathrm{t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{ii})}{\sum_{j=1}^{N}\exp(s_{ij})},\qquad \mathcal{L}_{\mathrm{t}\to\mathrm{i}} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp(s_{jj})}{\sum_{i=1}^{N}\exp(s_{ij})}$$

$$\boxed{\mathcal{L} = \tfrac{1}{2}\left(\mathcal{L}_{\mathrm{i}\to\mathrm{t}} + \mathcal{L}_{\mathrm{t}\to\mathrm{i}}\right)}\quad \text{with temperature } \tau \text{ (learned).}$$

What CLIP facilitates?

- Semantic alignment across-modality: “red car”, “dog wearing hat”.
- Compositional reasoning in embedding space.
- Transferable representations for zero-shot tasks with text prompts.

After training:

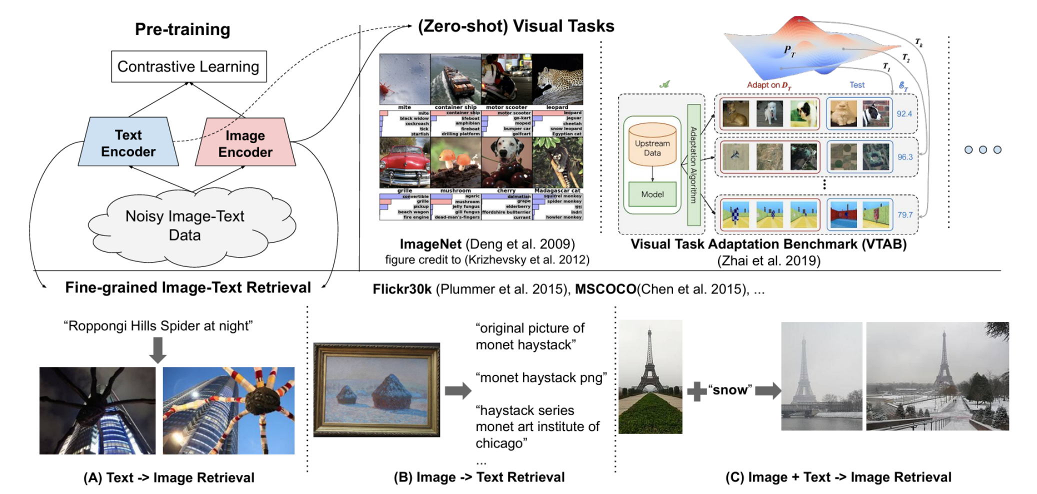


A. Radford, et al., 2021, CLIP

Limitations of CLIP:

- Noisy captions (reporting bias).
- Poor domain transfer (e.g., medical, scientific).
- Lacks causal or spatial grounding.

Larger-Scale Noisy web data (1.8 B) ⇒ More robust representations.

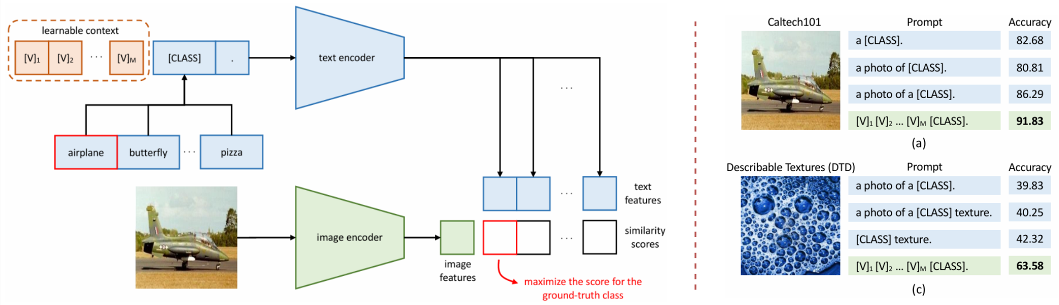


C. Jia, et al., 2021, ALIGN

Improving Contrastive VLMs through Post-Hoc Adaptation

CoOp: Improving CLIP Inference through Context Optimization

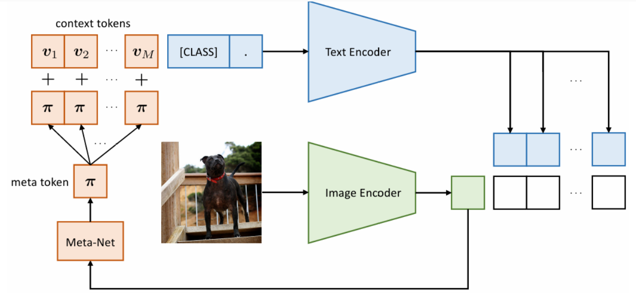
CoOp (Context Optimization) replaces hand-crafted text prompts in CLIP with **learnable context tokens**: it freezes the image/text encoders and **optimizes the prompt vectors** end-to-end on a small labeled set, yielding stronger zero-/few-shot recognition without retraining CLIP.



K. Zhou, et al., 2022, Learning to Prompt for Vision-Language Models

CoCoOp: Conditional Prompt Learning for Vision-Language Models

CoCoOp improves on CoOp by making the prompt **input-conditional**—a lightweight meta-network generates prompt context from each image's features, yielding better generalization to unseen classes/domains while keeping CLIP frozen.

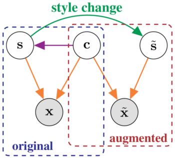


K. Zhou, et al., 2022, Conditional Prompt Learning for Vision-Language Models

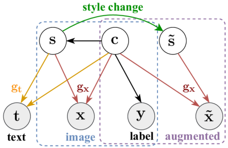
CLAP: Contrastive Learning with Augmented Prompts

Motivation: Image–text share latent semantics ⇒ transfer across modalities; enforcing invariances via *text* is more precise and efficient than editing *images*.

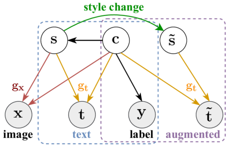
Theory (von Kügelgen, et al., 2021)	Our findings
• Contrastive learning with data augmentation; • All style changes through DA for pure content factors.	• The shared latent space of two-modalities; • High semanticity and logical structure in text data.



von Kügelgen, et al., 2021



(a) Image augmentation



(b) Text Augmentation

c - Latent content variable
s - Latent style variable

x - Image data sample
t - Text data sample

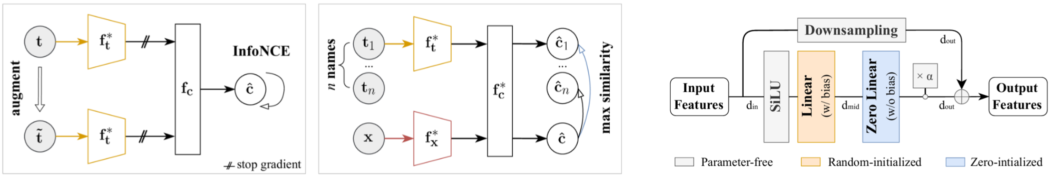
$\mathbf{g_x}$ - Image generating process
$\mathbf{g_t}$ - Text generating process

y - Label of a class
$\tilde{*}$ - Changed view

Y. Cai, et al., 2024, CLAP

CLAP: Contrastive Learning with Augmented Prompts

Method: Adapt a pretrained CLIP with contrastive learning and augmented text prompts (CLAP) to disentangle **object-class** semantics from nuisances, improving downstream accuracy and robustness.



CLAP

Zero-shot inference

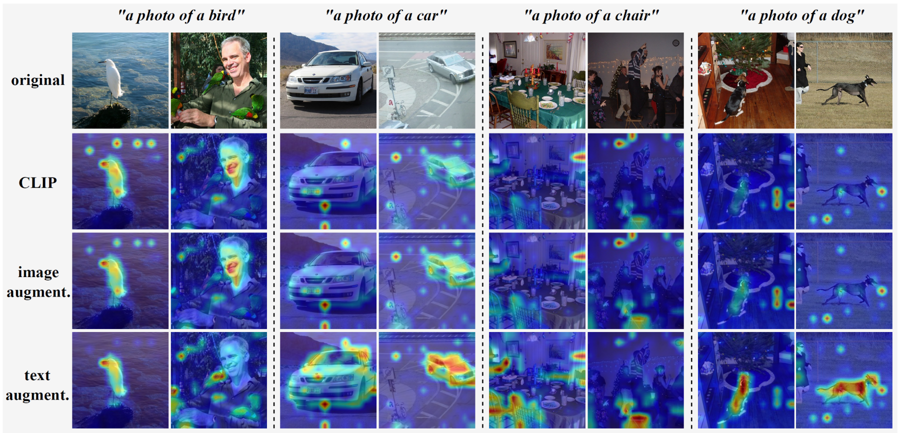
$$\mathcal{L}(\mathbf{f}; \{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}_{i=1}^b, \tau) = -\frac{1}{b}\sum_{i=1}^b \log \frac{\exp[\langle \mathbf{f}(\mathbf{x}_i), \mathbf{f}(\tilde{\mathbf{x}}_i)\rangle/\tau]}{\sum_{j=1}^b \exp[\langle \mathbf{f}(\mathbf{x}_i), \mathbf{f}(\tilde{\mathbf{x}}_j)\rangle/\tau]}$$

$$\mathbf{f}_c^* = \operatorname*{argmin}_{\mathbf{f}_c} \mathop{\mathbb{E}}_{\{\mathbf{t}_i\}_{i=1}^b \in \mathcal{D}_t} \mathcal{L}(\mathbf{f}_c \circ \mathbf{f}_t^*; \{\mathbf{t}_i, \tilde{\mathbf{t}}_i\}_{i=1}^b, \tau) + \lambda \mathcal{L}(\mathbf{f}_c \circ \mathbf{f}_t^*; \{\mathbf{t}_i^c, \tilde{\mathbf{t}}_i\}_{i=1}^b, 1)$$

Y. Cai, et al., 2024, CLAP

CLAP: Contrastive Learning with Augmented Prompts

Qualitative results: Embeddings from baseline CLIP vs image-augmented adaptation vs text/prompt-augmented adaptation.



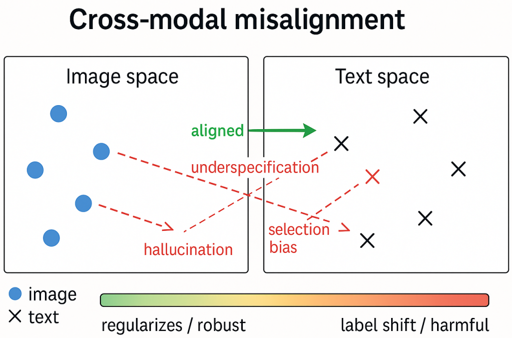
Y. Cai, et al., 2024, CLAP

Theoretical Perspective on Multimodal Contrastive Learning

Phenomenon: Images and Captions are Not Pefectly aligned

Motivation: Image–caption datasets exhibit **cross-modal misalignment** between visual content and text (e.g., underspecification, or perturbation).

Can we model these noise patterns formally? And, When do they hurt vs. help (e.g., robustness, generalization)?



What Does Multimodal Contrastive Learning Optimize?

Standard MMCL loss:

$$\mathcal{L}_{\text{MMCL}}(f_x, f_t) = -\frac{1}{2K}\sum_{i=1}^{K}\left[\log\frac{e^{\kappa(f_x(x_i), f_t(t_i))/\tau}}{\sum_j e^{\kappa(f_x(x_i), f_t(t_j))/\tau}} + \log\frac{e^{\kappa(f_x(x_i), f_t(t_i))/\tau}}{\sum_j e^{\kappa(f_x(x_j), f_t(t_i))/\tau}}\right]$$

Asymptotically (large K):

$$\mathcal{L}_{\text{MMCL}} \implies \mathbb{E}\,\|f_x(x) - f_t(t)\|^2 - \tfrac{1}{2}\Big[H(f_x(x)) + H(f_t(t))\Big]$$

What Does Multimodal Contrastive Learning Optimize?

Standard MMCL loss:

$$\mathcal{L}_{\mathrm{MMCL}}(f_x, f_t) = -\frac{1}{2K}\sum_{i=1}^{K}\left[\log \frac{e^{\kappa(f_x(x_i), f_t(t_i))/\tau}}{\sum_j e^{\kappa(f_x(x_i), f_t(t_j))/\tau}} + \log \frac{e^{\kappa(f_x(x_i), f_t(t_i))/\tau}}{\sum_j e^{\kappa(f_x(x_j), f_t(t_i))/\tau}}\right]$$

Asymptotically (large K):

$$\mathcal{L}_{\mathrm{MMCL}} \implies \mathbb{E}\,\|f_x(x) - f_t(t)\|^2 \; - \; \tfrac{1}{2}\Big[H(f_x(x)) + H(f_t(t))\Big]$$

Interpretation: Contrastive Learning =

- **Alignment:** pull paired features together

$$\downarrow \|f_x(x) - f_t(t)\|^2$$

- **Entropy maximization:** preserve information within each modality

$$\uparrow H(f_x(x)), \; \uparrow H(f_t(t))$$

A Theory On the Value of Cross-Modal Misalignment

Theory framework: Y. Cai et al., 2025 introduce a latent variable model where each modality is generated from shared semantics with modality-specific biases.

Selection Bias (θ): Shared semantics excluded from the text modality.
Perturbation Bias (ρ): Spurious or altered semantics added to text.

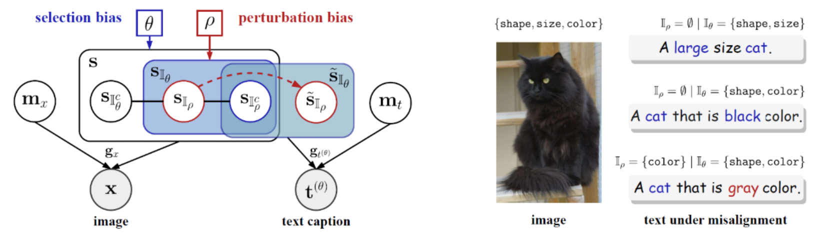
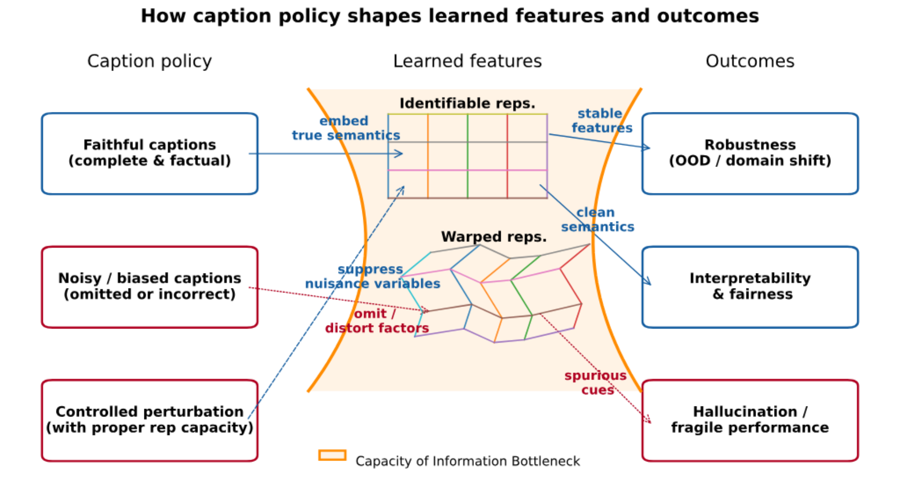


Illustration of our latent variable model with semantic misalignment.

Main Result: Contrastive learning consistently identifies unbiased shared semantics, regardless of latent causal structure, explaining why CLIP-like models succeed despite noise.

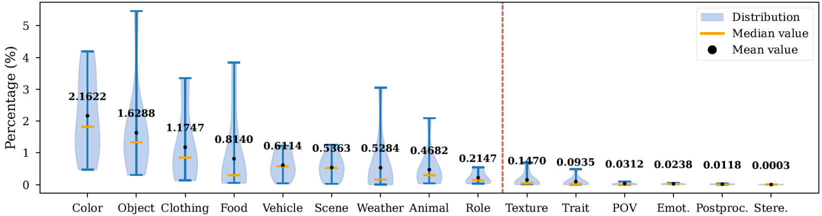
THE UNIVERSITY of ADELAIDE

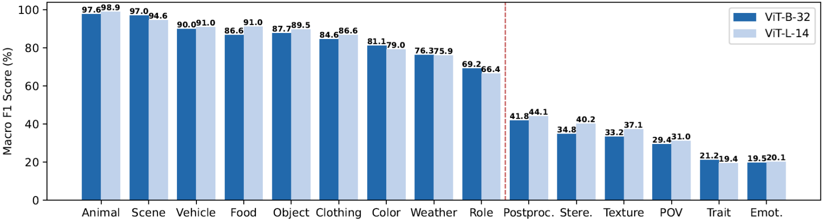
Practical implications:



A Theory On the Value of Cross-Modal Misalignment

Zero-shot evaluation on OpenCLIP trained with LAION-400M dataset:





Key Takeaways of Contrastive VLMs

- **Simple & scalable:** dual encoders + ℓ_2 normalization + cosine similarity with learned temperature (τ).
- **Data scale & diversity** beat cleanliness; web-scale noise is often tolerable (sometimes helpful).
- **Strengths:** excellent zero-shot transfer & retrieval; strong open-vocabulary classification.
- **Caption quality matters (pretraining):** identifies semantics *shared across modalities*; biased captions limit what can be learned.
- **Prompts matter (inference):** templates/ensembles and learned prompts (CoOp/CoCoOp) improve transfer.
- **Foundation layer:** strong vision encoder to pair with fusion/instruction-tuned models for multimodal reasoning.

Application: Generative Diffusion Models Conditioned on VLM Encoders

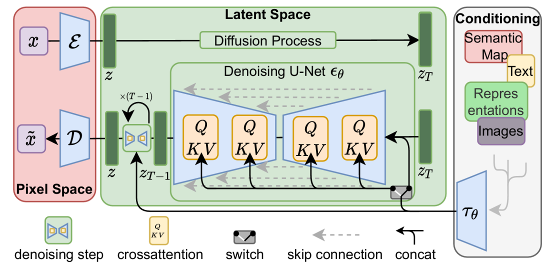
Latent Diffusion Models: Text Conditioning & Classifier-Free Guidance

How text conditions LDM

- Frozen **CLIP text encoder** → prompt embeddings.
- U-Net conditioned via **cross-attn**.
- **CFG** at sampling:

$$\hat{\varepsilon} = (1 - w)\,\varepsilon_\theta(x_t, t, \varnothing) \; + \; w\,\varepsilon_\theta(x_t, t, c)$$

$w \uparrow$ = stronger prompt adherence.



LDM: latent U-Net with a CLIP text encoder.

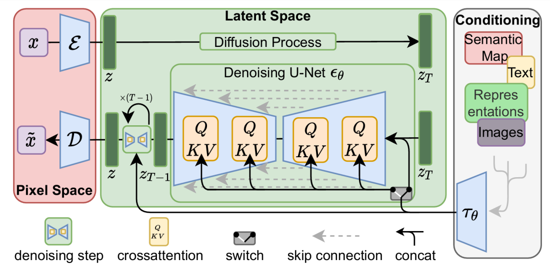
THE UNIVERSITY of ADELAIDE

How text conditions LDM

- Frozen **CLIP text encoder** → prompt embeddings.
- U-Net conditioned via **cross-attn**.
- **CFG** at sampling:

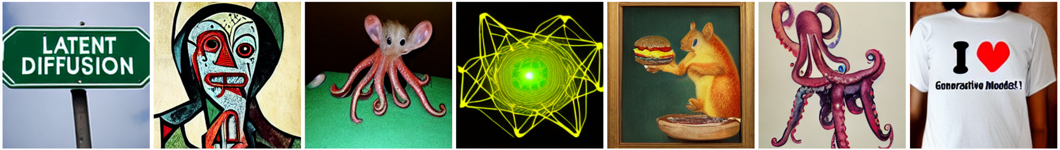
$$\hat{\varepsilon} = (1 - w)\,\varepsilon_\theta(x_t, t, \varnothing) \;+\; w\,\varepsilon_\theta(x_t, t, c)$$

$w \uparrow$ = stronger prompt adherence.



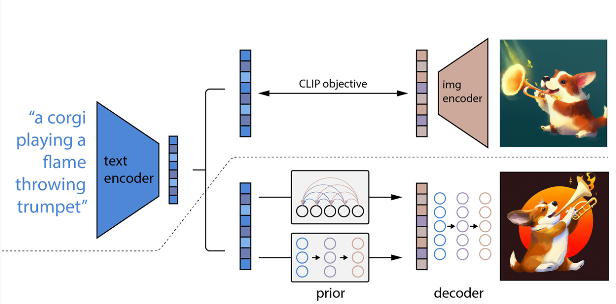
LDM: latent U-Net with a CLIP text encoder.

'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!" '



How it works

- **CLIP-V (frozen)** defines image embedding z_{clip}.
- **Text→prior** samples z_{clip}.
- **Diffusion decoder** generates $x \sim p(x \mid z_{\text{clip}})$.
- **CFG:** $\hat{\varepsilon} = (1-w)\varepsilon_\theta(x_t, t, \varnothing) + w\,\varepsilon_\theta(x_t, t, z_{\text{clip}})$
- *Modes:* text→image, image variations, edits.



UnCLIP pipeline: prior over CLIP-z → diffusion decoder.

A. Ramesh et al., 2022 (DALL·E 2 / unCLIP): prior over CLIP image embeddings + diffusion decoder.

UnCLIP: High-Fidelity “Imagined” Generation via CLIP–Vision Conditioning

Samples from DALL·E 2 / unCLIP (text → image): a prior samples *CLIP vision* embeddings, then a diffusion decoder renders high-fidelity scenes.


an espresso machine that makes coffee from human souls, artstation


panda mad scientist mixing sparkling chemicals, artstation


a corgi’s head depicted as an explosion of a nebula

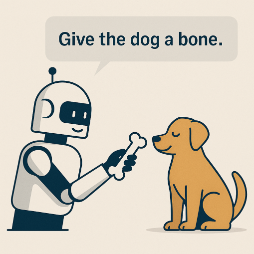
Ramesh et al., 2022 — unCLIP: prior over CLIP image embeddings ⟶ diffusion decoder.

How CLIP Representations Interface with Generators (At a Glance)

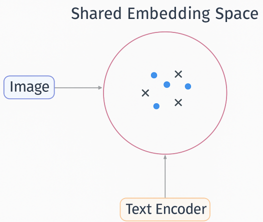
- **Text-side conditioning:** Use CLIP *text* embeddings as prompts/keys for conditioning modules (e.g., cross-attn keys/values).
- **Vision-side priors:** Map text to a prior over *CLIP vision* embeddings (z_{clip}); decode images conditioned on z_{clip} (UnCLIP-style).
- **Retrieval cues:** Nearest neighbors in CLIP space provide exemplar style/structure to condition or prompt the generator.
- **Personalization:** Learn new tokens/embeddings in CLIP *text* space (e.g., textual inversion) for concept-specific generation.
- **Diagnostics:** Monitor modality gap/prompt sensitivity in CLIP space when adapting generators.

Multimodal LLMs: From Representation to Reasoning

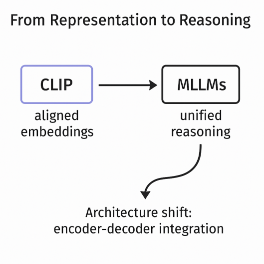
Representation learning (alignment) ⇒ Reasoning with LLM backbones



Motivation: Sometimes real-world tasks can be complex

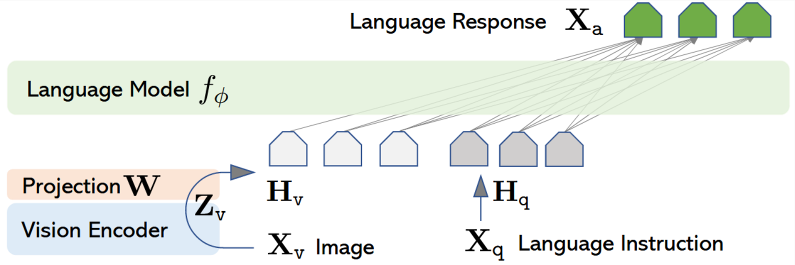


CLIP: shared image–text space zero-shot via prompts



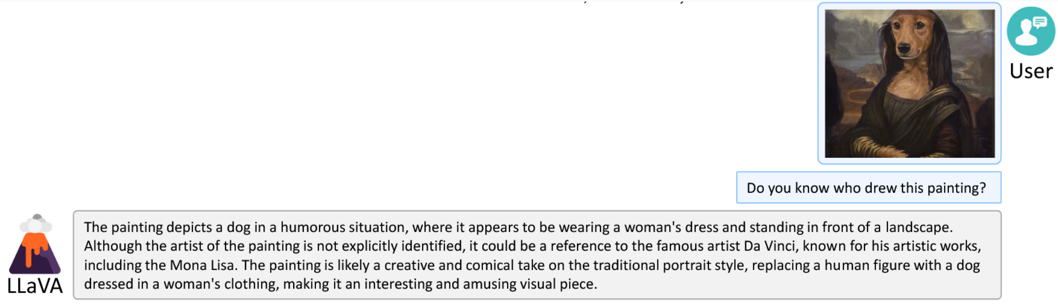
MLLMs: instruction following multi-step reasoning

- **Synthesize instructions** from captions, build (image, prompt, response) pairs (QA, reasoning, visual grounding).
- **Supervised tune** a vision–LLM via the adaptor to map visual tokens into the LLM's reasoning space.
- **Outcome:** better instruction following, reasoning, and task generalization.



H. Liu et al., 2023 (LLaVA): Visual Instruction Tuning.

LLaVA recognizes Leonardo da Vinci's *Mona Lisa* and also explains humorous web parodies that mimic it.



H. Liu et al., 2023 (LLaVA): Visual Instruction Tuning.

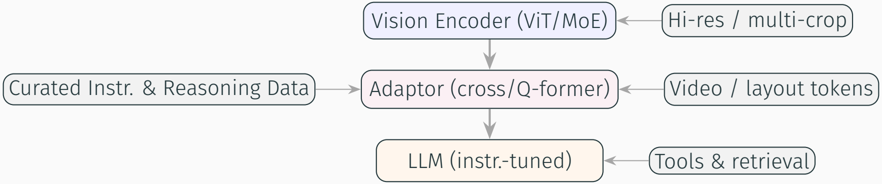
THE UNIVERSITY of ADELAIDE

Common recipe

- Vision encoder → adaptor → Large Language Model
- Instruction tuning; multi-turn chat
- Captioning / Visual Question Answering / Optical Character Recognition / grounding

Differences

- Perception: high resolution, multi-crop, tables/charts
- Context: long context, multiple images or video
- Coverage: multilingual support, domain-specific tools



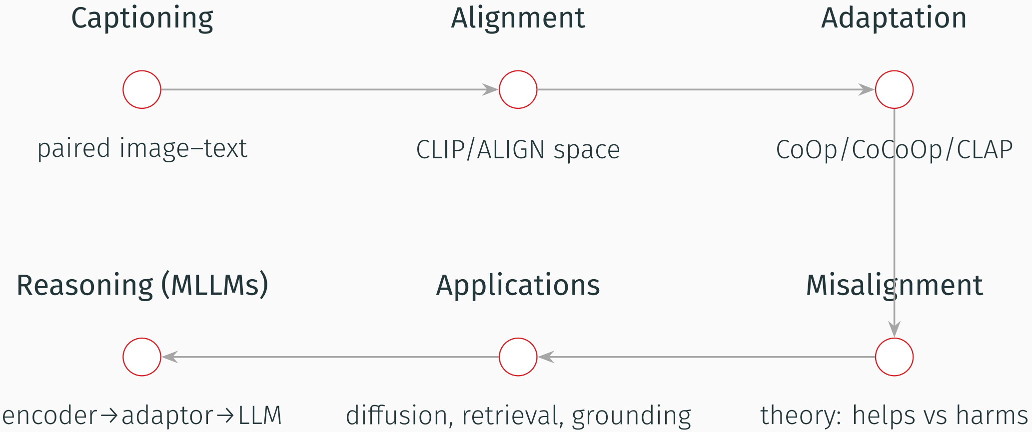
Qwen-VL GPT-4V Gemini 1.5

Challenges and Research Directions

Challenges	Research directions
Hallucination	Factuality and calibration
Lack of explicit grounding	Grounded perception and spatial reasoning
Data imbalance and causal confusion	Compositional and causal learning
Entangled visual encoder representations	Disentangled and causal encoders
Mismatch with human values	Human value alignment and safety

Wrap-Up and Discussions

The Journey in One Slide



Open Questions

- What defines model **performance** in the general-AI era?
- How do we quantify data **quality** for vision–language?
- How to design **better supervision** (data process + objectives)?
- How to evaluate **reasoning & grounding** reliably?

Recommended Reading

THE UNIVERSITY of ADELAIDE

- Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021
- Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, ICML 2021
- Zhou et al., Conditional Prompt Learning for Vision-Language Models, CVPR 2022
- Cai et al., CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts, ECCV 2024
- Cai et al., On the Value of Cross-Modal Misalignment in Multimodal Representation Learning, NeurIPS 2025
- Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022
- Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022
- Liu et al., Visual Instruction Tuning, NeurIPS 2023
- Bai et al., Qwen-VL: AVersatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023

Thank You !

(Q & A)

Interested in VLMs? Email yichao.cai@adelaide.edu.au to discuss more.