Multimodal AI

Lecture 7.1 – Midterm Review

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

mit media lab

MIT EECS

Assignments for This Coming Week

HW3 due today.

Project proposals will be graded and released this week.

Project midterm instructions will be out this week.

Midterm this Thursday.

Midterm Structure

6 problems, 60 minutes.

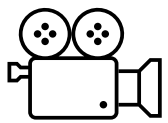
5 problems:

1. MCQ (10, 2 points each = 20 points)
2. Short answers (4, 5 points each = 20 points)
3. Multimodal Fusion (20 points)
4. Multimodal LLMs (20 points)
5. Multimodal Generation (20 points)
6. Bonus open questions (10 points)

What is a Modality?

Modality

Modality refers to the way in which something expressed or perceived.



from a sensor

Raw Modalities
(closest from sensor)

Abstract Modalities
(farthest from sensor)

Examples:

Raw Modalities		Abstract Modalities
Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

multisensory intelligence

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

***Multimodal* is the science of**

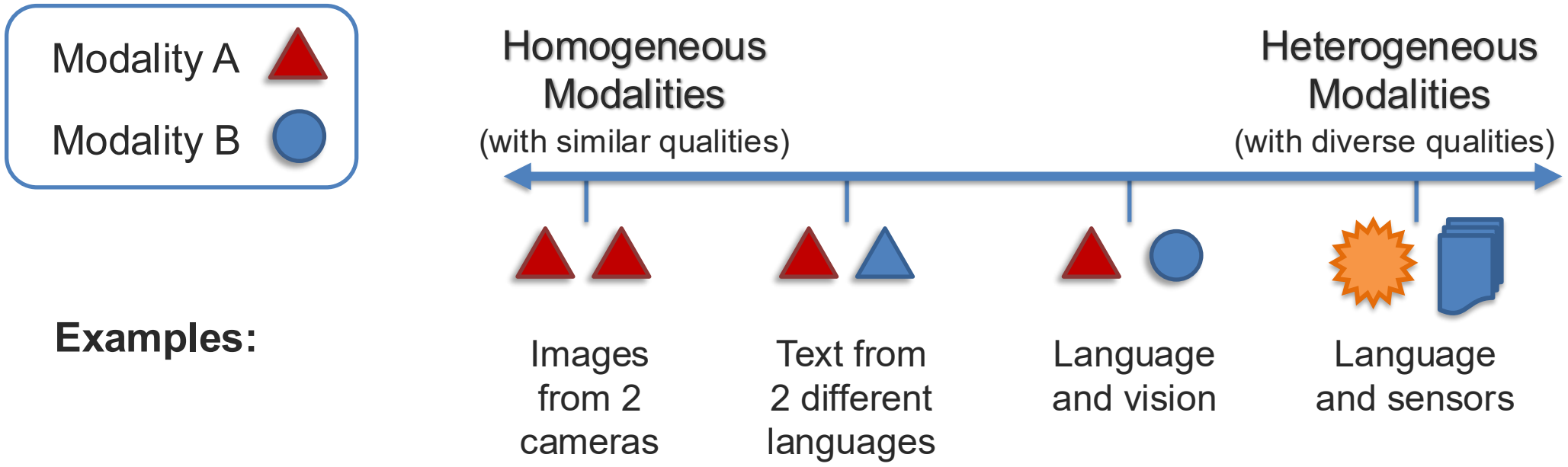
heterogeneous and interconnected data

Connected + Interacting

multisensory intelligence

Heterogeneous Modalities

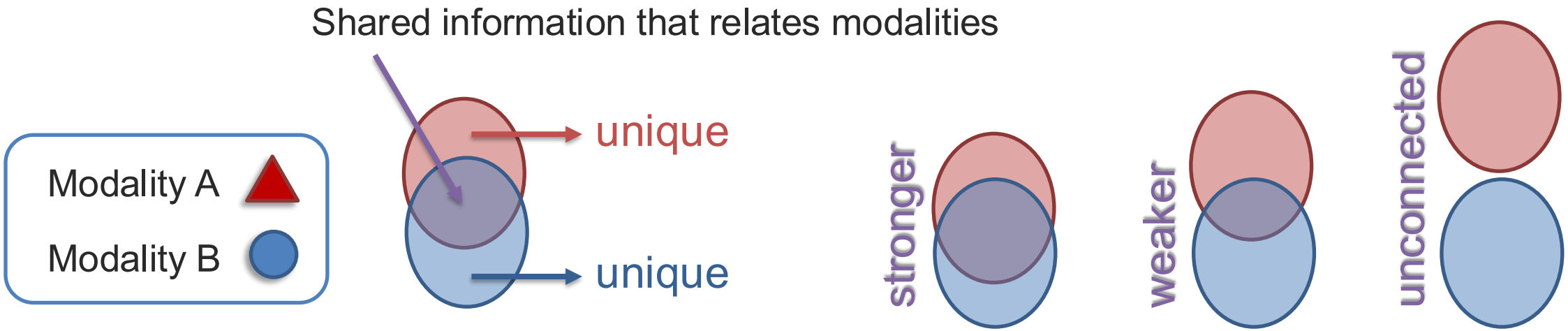
Information in different modalities shows diverse qualities, structures, & representations.



Abstract modalities are more likely to be homogeneous

multisensory intelligence

Connected Modalities

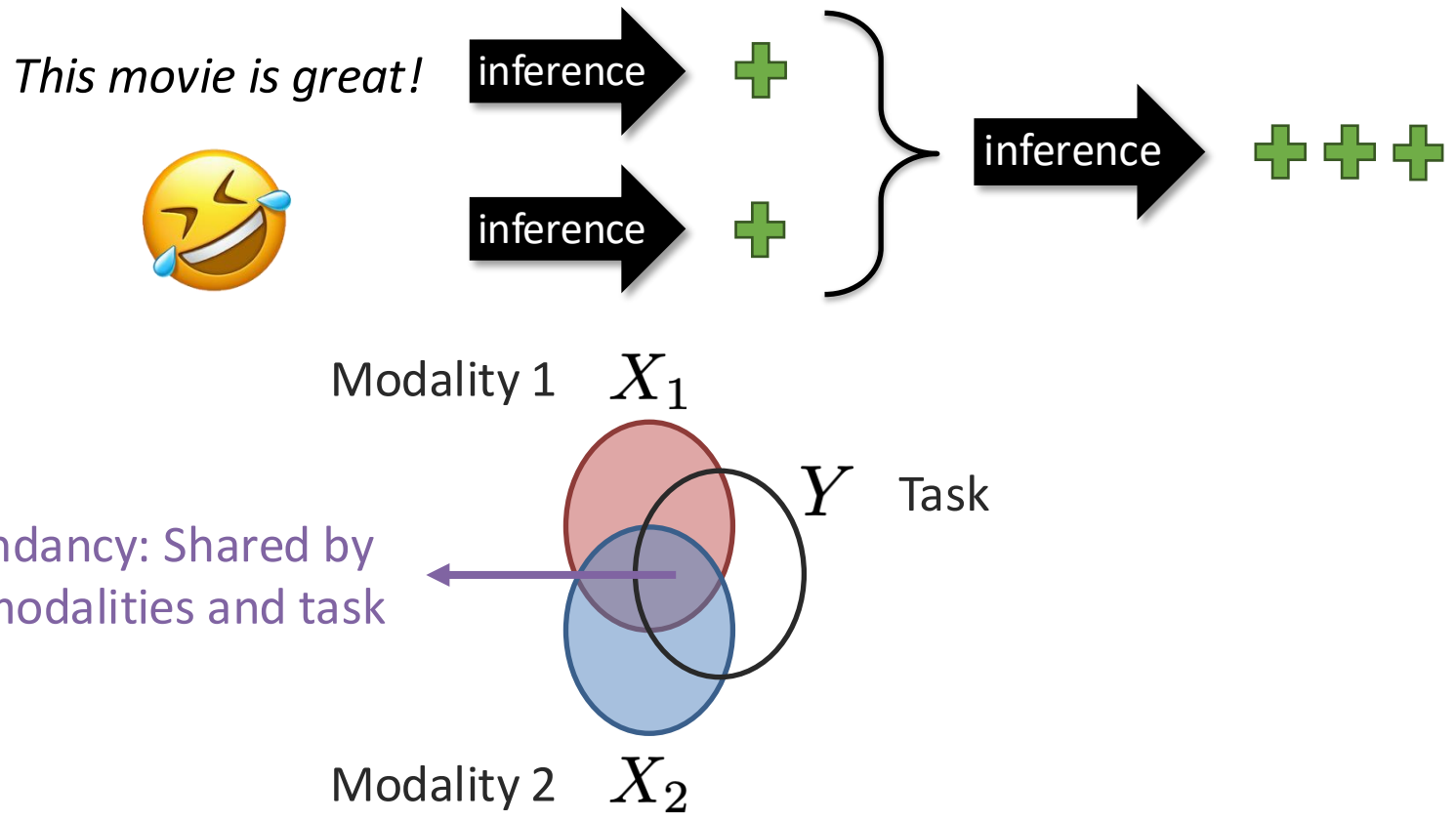


A teacup on the right of a laptop in a clean room.

multisensory intelligence

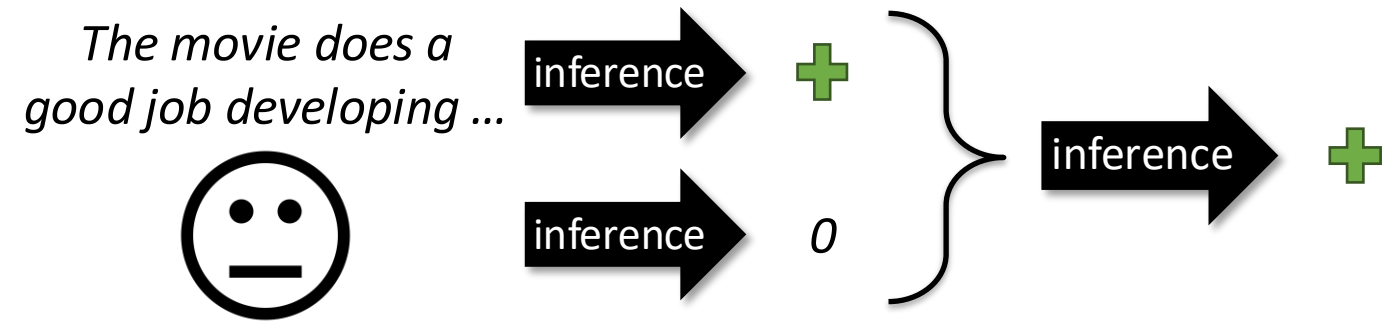
Interacting Modalities

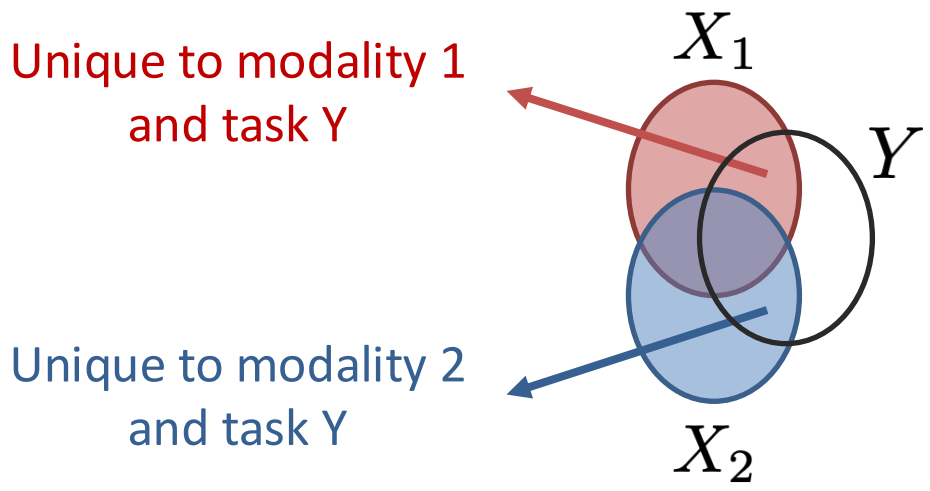
Interactions: How modalities *combine* to provide information for a task.



Interacting Modalities

Interactions: How modalities *combine* to provide information for a task.

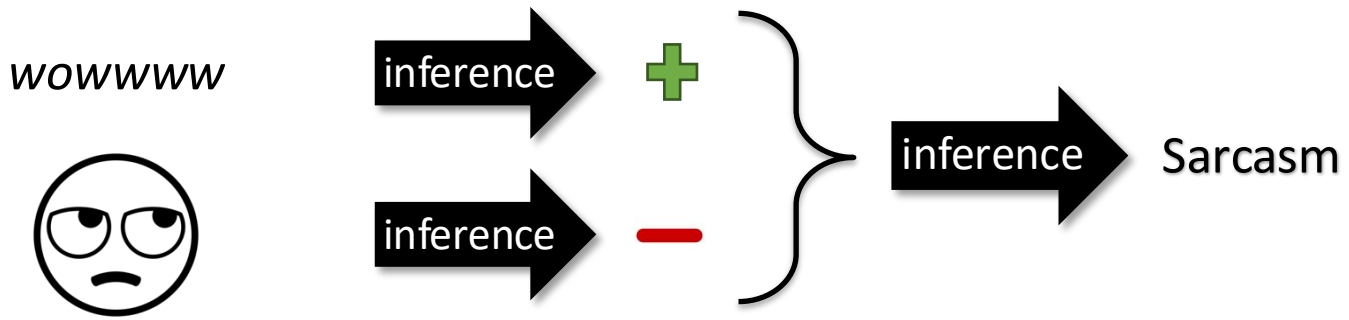


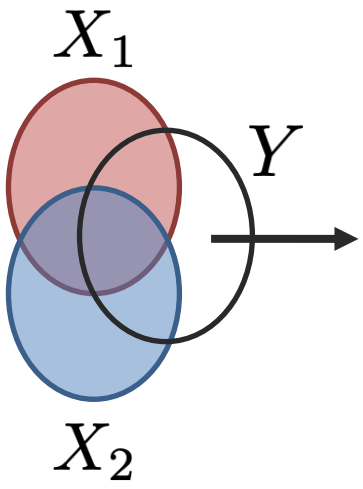


multisensory intelligence

Interacting Modalities

Interactions: How modalities *combine* to provide information for a task.





Synergy: Emerging information from multimodal interaction

multisensory intelligence

What is Multimodal? → *Why is it hard?* → *What is next?*

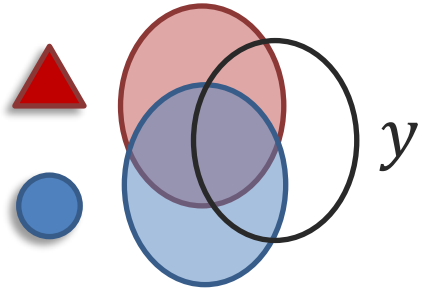
Heterogeneous

+

Connected

+

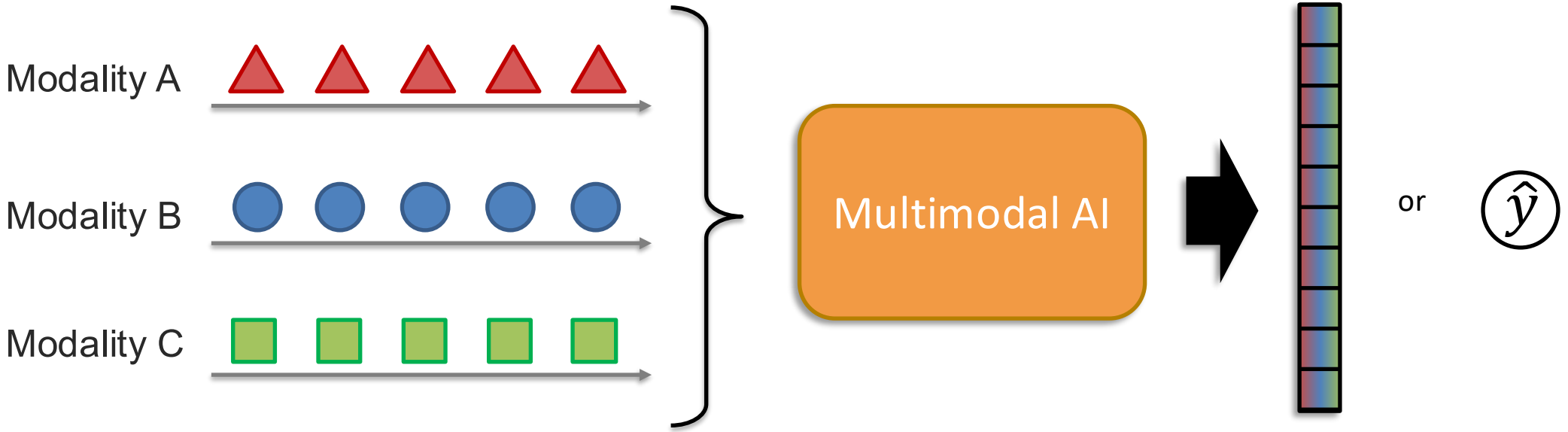
Interacting



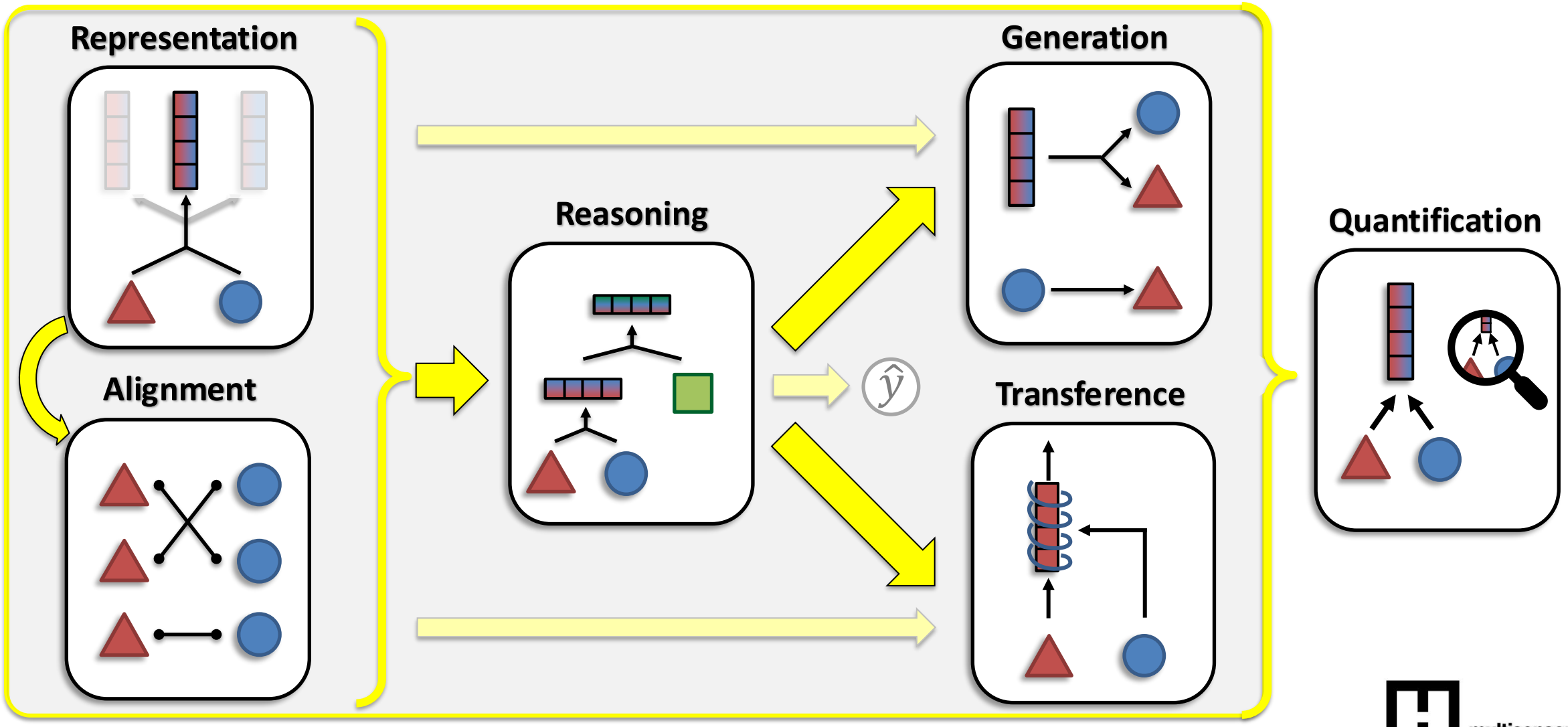
Multimodal is the scientific study of heterogeneous and interconnected data ☺

multisensory intelligence

Multimodal AI Challenges



Summary of Core Multimodal Challenges

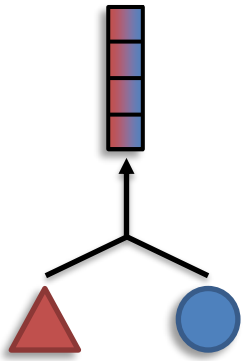


Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities.

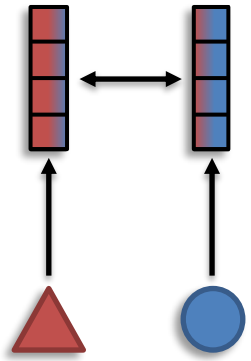
Sub-challenges:

Fusion



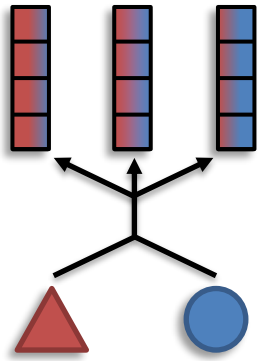
modalities > # representations

Coordination



modalities = # representations

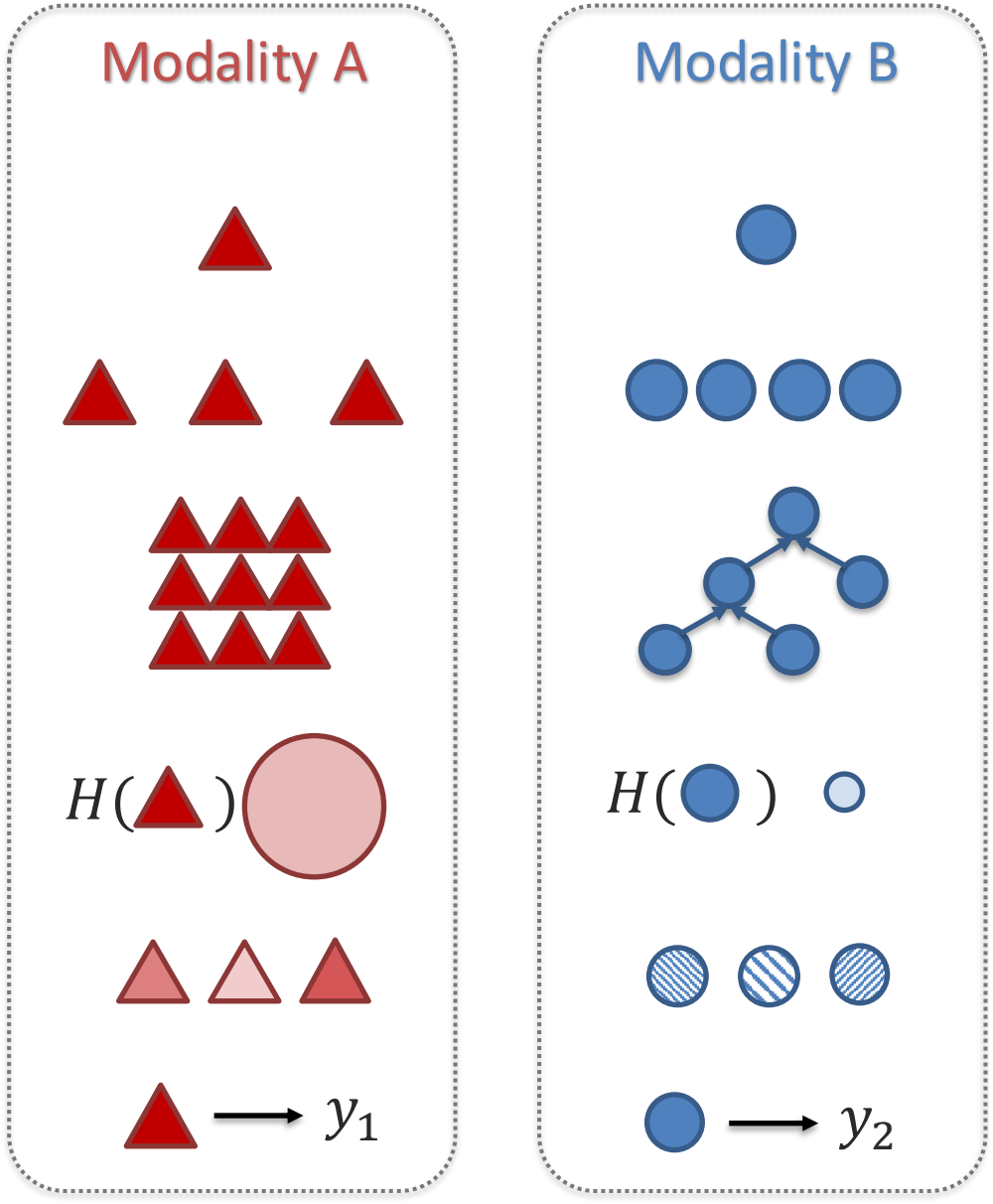
Fission



modalities < # representations

Modality Profile

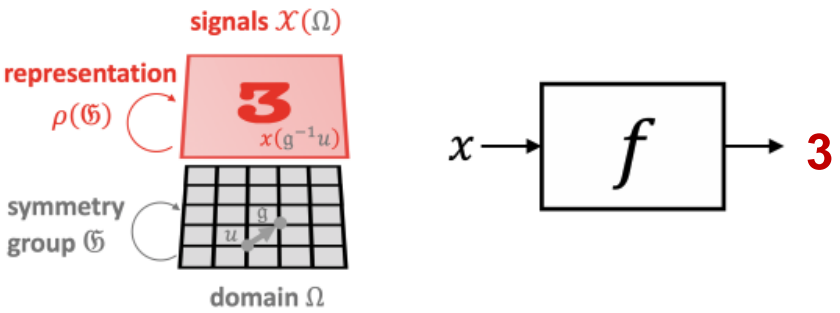
1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence



Unimodal Structure

Data invariances – example of image classification

A function $f : \mathcal{X}(\Omega) \to \mathcal{Y}$ is $\mathfrak{G}$-*invariant* if $f(\rho(\mathfrak{g})x) = f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$ and $x \in \mathcal{X}(\Omega)$, i.e., its output is unaffected by the group action on the input.



sunset

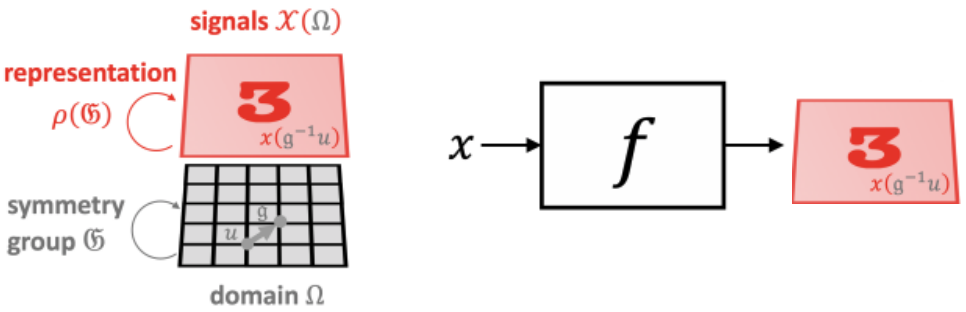
[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

multisensory intelligence

Unimodal Structure

Data equivariances – example of image segmentation

A function $f : \mathcal{X}(\Omega) \to \mathcal{X}(\Omega)$ is $\mathfrak{G}$-*equivariant* if $f(\rho(\mathfrak{g})x) = \rho(\mathfrak{g})f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$, i.e., group action on the input affects the output in the same way.



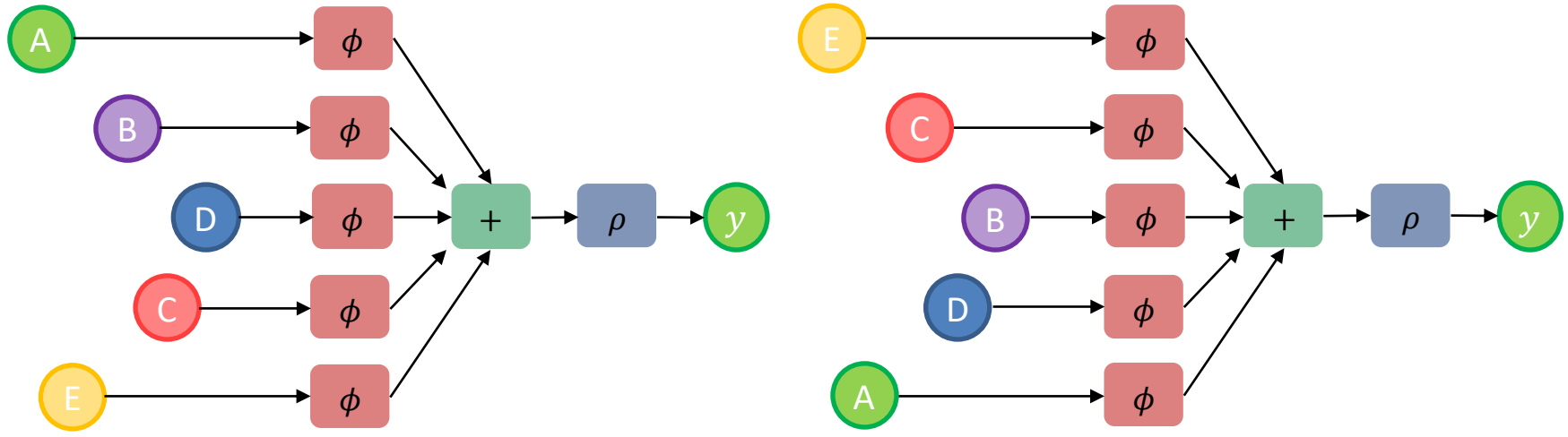
[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

multisensory intelligence

Unimodal Models

Models for set-based data must be invariant to element order.

1. Parameter sharing for each set element
2. Permutation invariant aggregation function

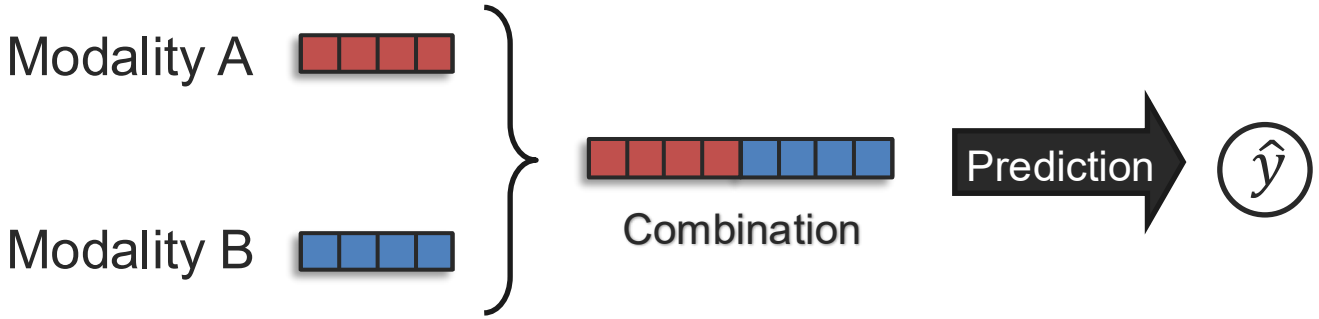


Giving ABDCE also gives ECBDA, BCAED etc…

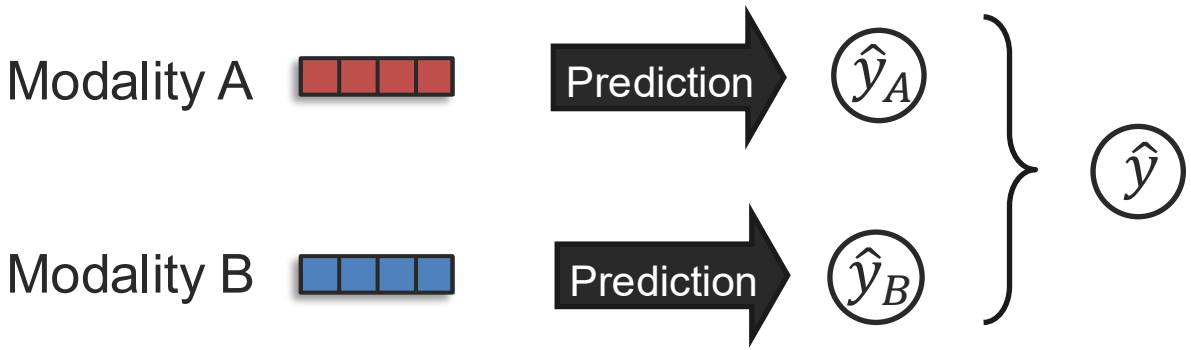
[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

Early and Late Fusion

Early fusion:



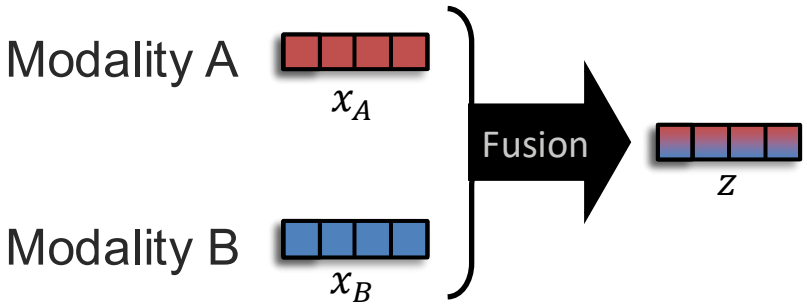
Late fusion:



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Basic Concepts for Representation Fusion



Goal: Model *cross-modal interactions* between the multimodal elements

➡ Let's study the univariate case first

↳ (only 1-dimensional features)

Linear regression:

$$z = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

w_0: constant; $w_1 x_A + w_2 x_B$: Additive terms; $w_3 (x_A \times x_b)$: Multiplicative term; ϵ: error

① Additive interaction:

$$z = w_1 x_A + w_2 x_B + \epsilon$$

② Multiplicative interaction:

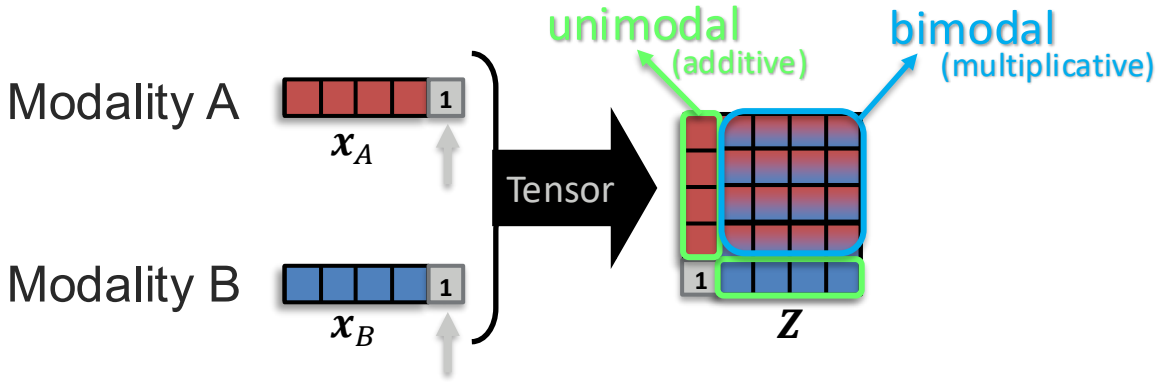
$$z = w_3 (x_A \times x_b) + \epsilon$$

③ Additive and multiplicative interactions:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

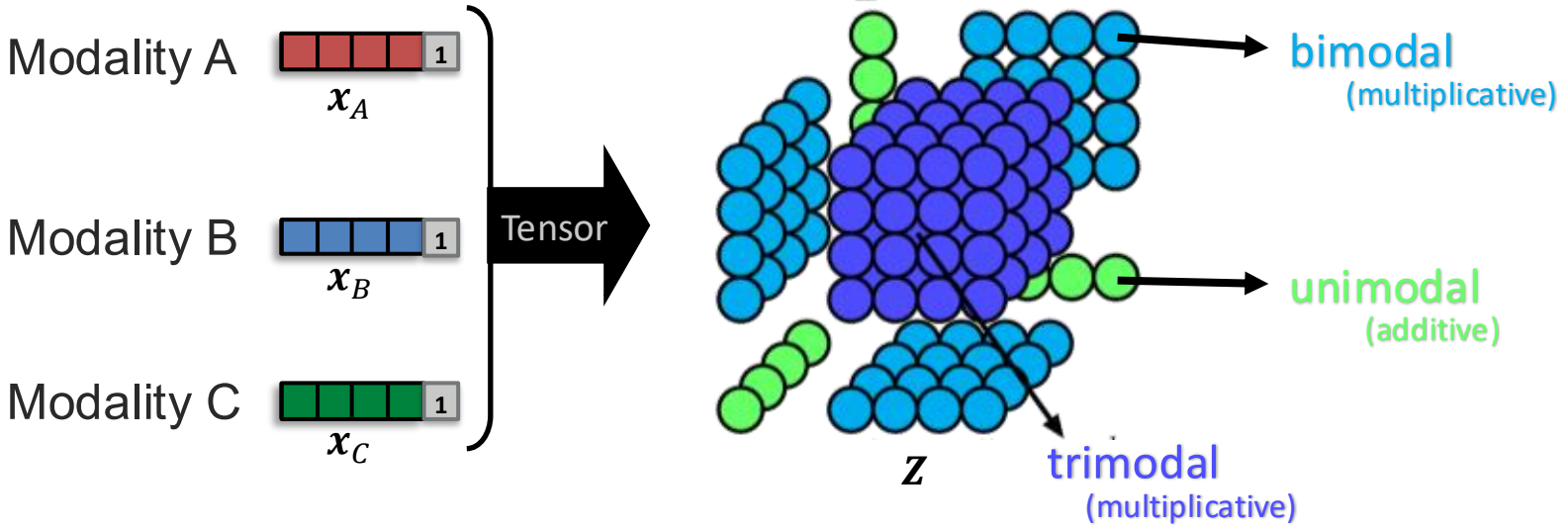
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Tensor Fusion



Tensor Fusion (bimodal):

$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$

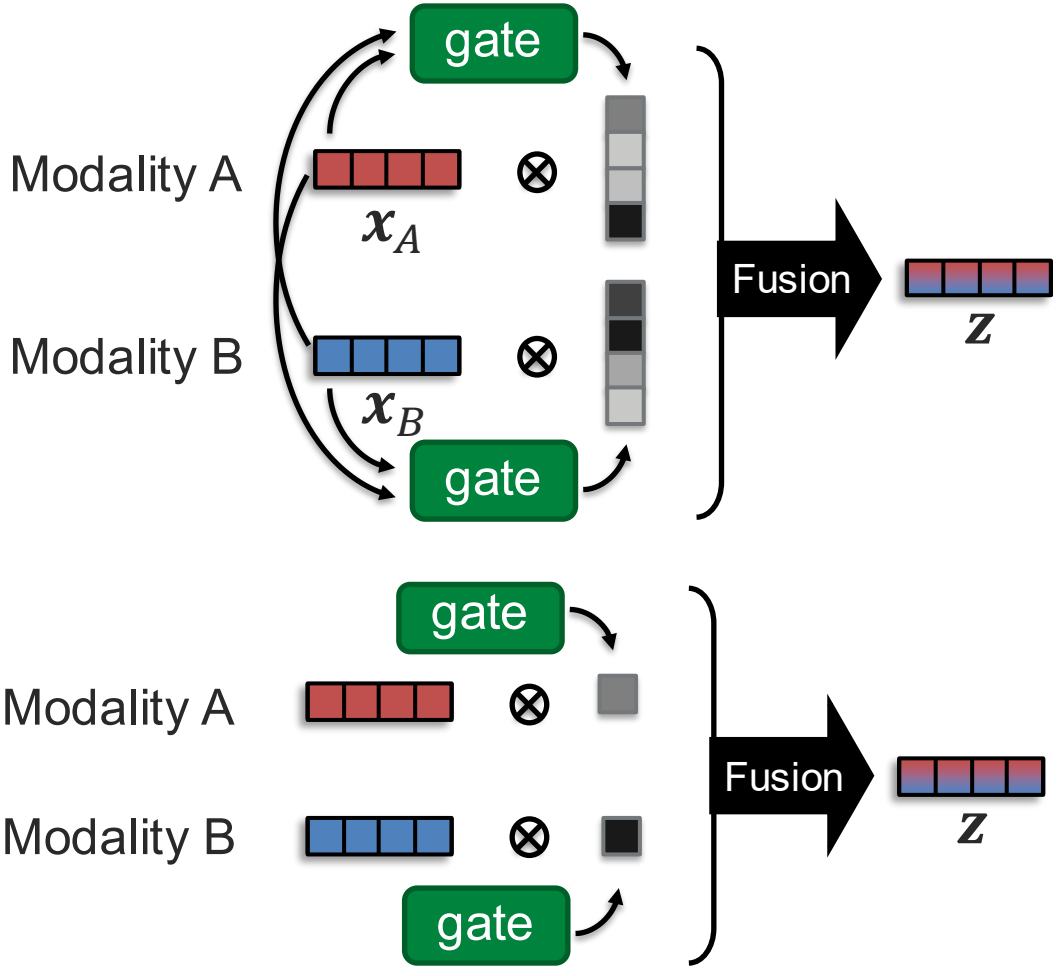


… but the weight matrix may end up quite large!

[Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017]

multisensory intelligence

Gated Fusion



Example with additive fusion:

$$z = g_A(x_A, x_B) \cdot x_A + g_B(x_A, x_B) \cdot x_B$$

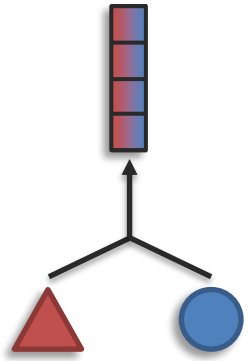
→ g_A and g_B can be seen as attention functions

→ Gating output can be one weight for the whole modality

[Arevalo et al., Gated Multimodal Units for Information Fusion. ICLR-workshop 2017]

multisensory intelligence

Summary: How To Multimodal Fusion



Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Homogenous modalities

- Late fusion
- Additive fusion
- Multiplicative fusion
- Tensor fusion
- Polynomial fusion
- Gated fusion
- Modality-shift fusion
- Nonlinear fusion
- Very early fusion
- Dynamic early fusion
- Heterogeneity-aware
- Improving optimization
- Improving robustness

Heterogenous modalities

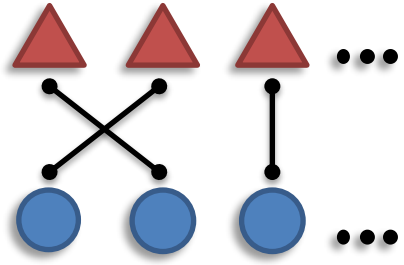
multisensory intelligence

Challenge 2: Alignment

Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

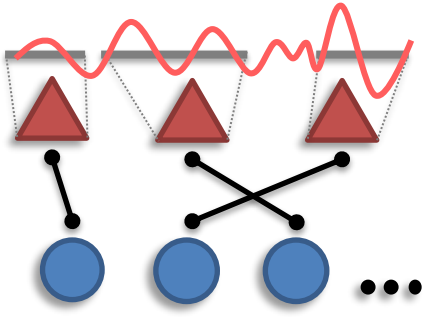
Sub-challenges:

Discrete connections



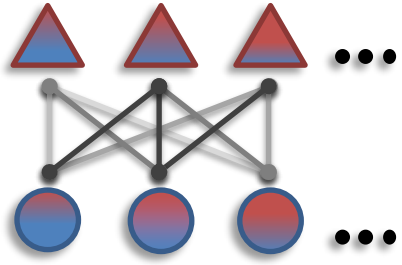
Explicit alignment
(e.g., grounding)

Continuous alignment



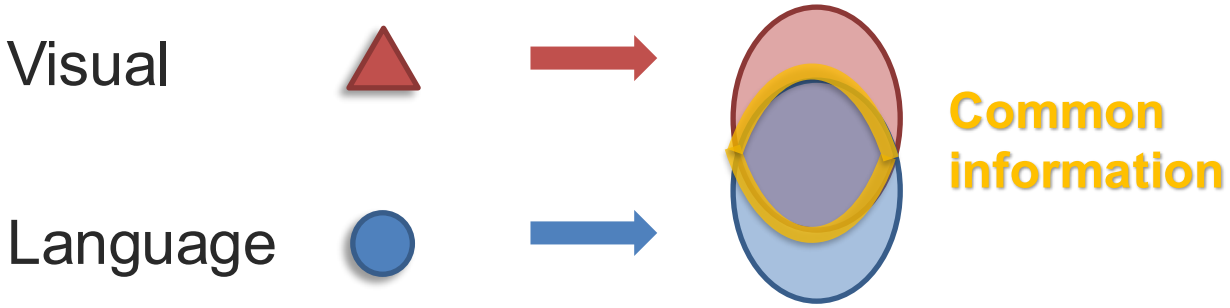
Granularity of individual elements

Contextualized representation

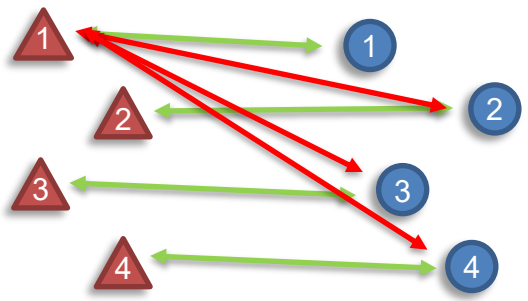


Implicit alignment + representation

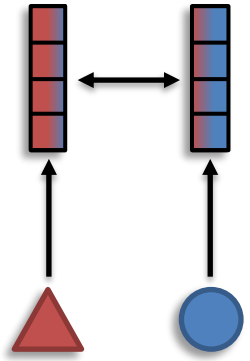
Modality Connections



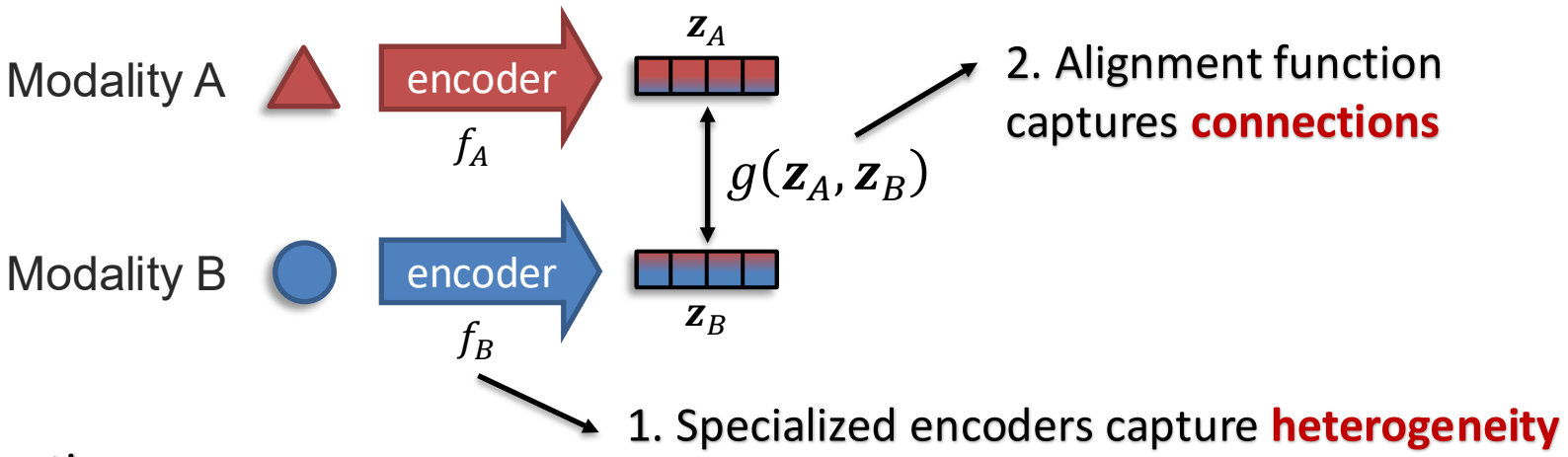
Supervision: Paired data



Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



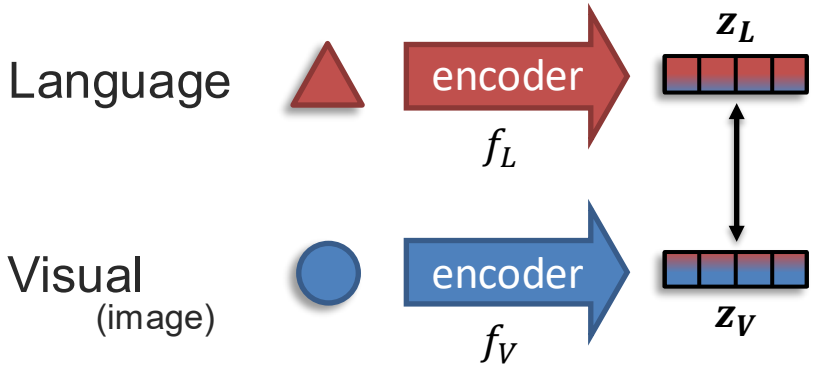
Learning with alignment function:

$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

multisensory intelligence

Contrastive Language Image Pretraining



Popular contrastive loss: InfoNCE

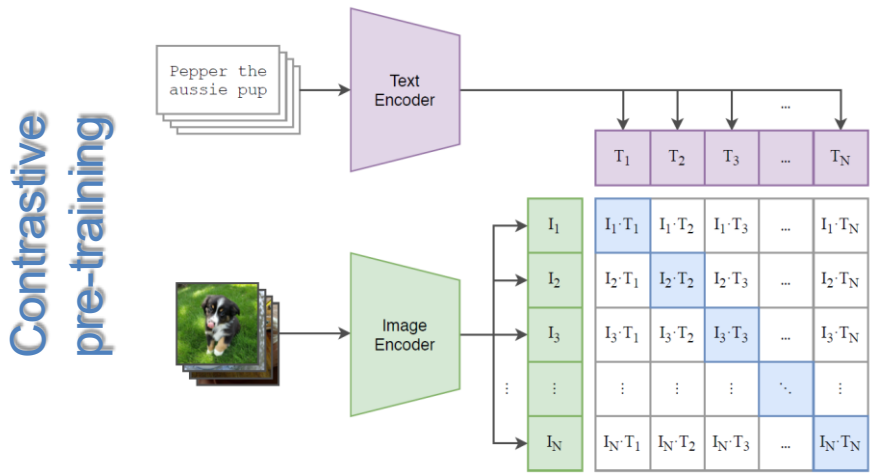
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\text{sim}(\boldsymbol{z}_A^l, \boldsymbol{z}_B^l)}{\sum_{i=1}^{N} \text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

Positive and negative pairs:



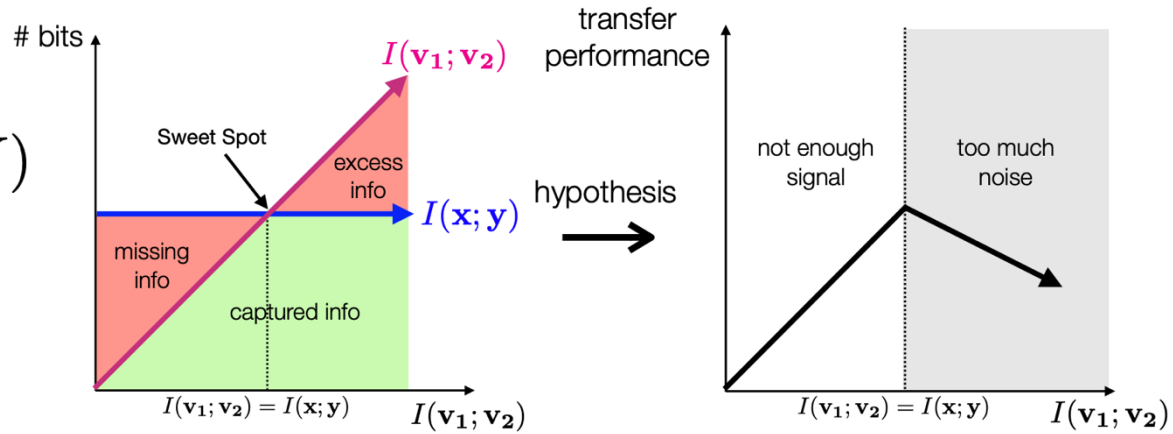
- CLIP encoders (f_L and f_V) are great for language-vision tasks
- z_L and z_V are coordinated but not identical representation spaces

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

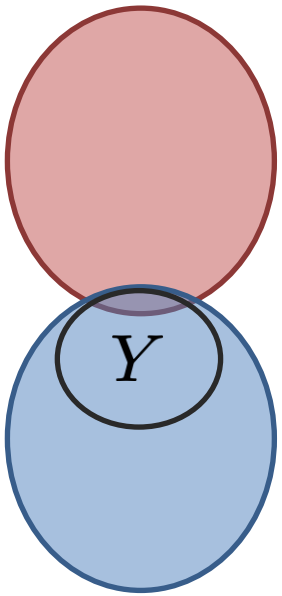
Multiview Redundancy and Contrastive Learning

How much information should be shared?

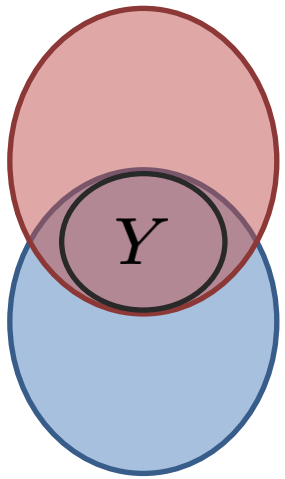
Multi-view redundancy: $I(X_1; X_2) = I(X_1; Y)$



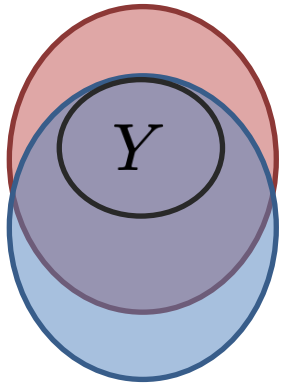
Not enough signal



Just right



Too much noise

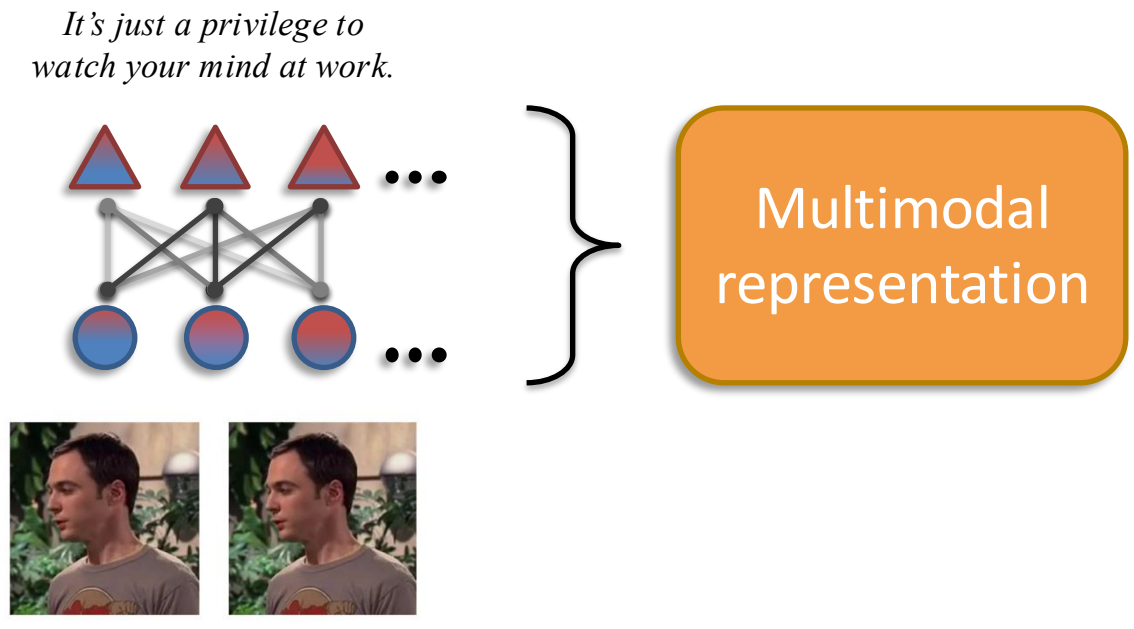


Multi-view redundancy may not hold for multimodal problems!

[Tian et al., What makes for Good Views for Contrastive Learning? NeurIPS 2020]
[Tosh et al., Contrastive Learning, Multi-view Redundancy, and Linear models. ALT 2021]

Large Multimodal Models

Part 1: Multimodal foundation model representations of text, video, audio



Multimodal Transformers

It's just a privilege to watch your mind at work.

privilege
mind

vocal
emphasis
eye-roll

vocal
emphasis
eye-roll

Vision-to-language attention

	vocal			eye-roll
privilege	0.7			0.3
mind	1.0			

3 x 4
(weighted) outer product

(row) normalize to 0-1

3 x d d x 4

$$h = \text{softmax}\left(\frac{X_1 W_q W_k^T X_2^T}{\sqrt{d}}\right) X_2 W_v$$

4 x d

normalize wrt dimension *d*

privilege
mind

New **language** representation
contextualized with vision

Sarcasm

[Liang et al., Multimodal Language Analysis with Recurrent Multistage Fusion. EMNLP 2018]
[Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL 2019]

multisensory intelligence

Visual-and-Language Transformer (ViLT)

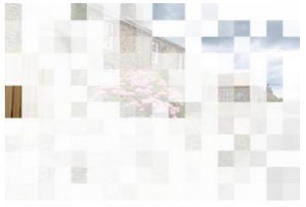
Example of alignment between modalities:

a display of flowers growing out and over the retaining wall in front of cottages on a cloudy day.

flowers

wall

cottages



cloudy

a room with a rug, a chair, a painting, and a plant.

rug

chair

painting

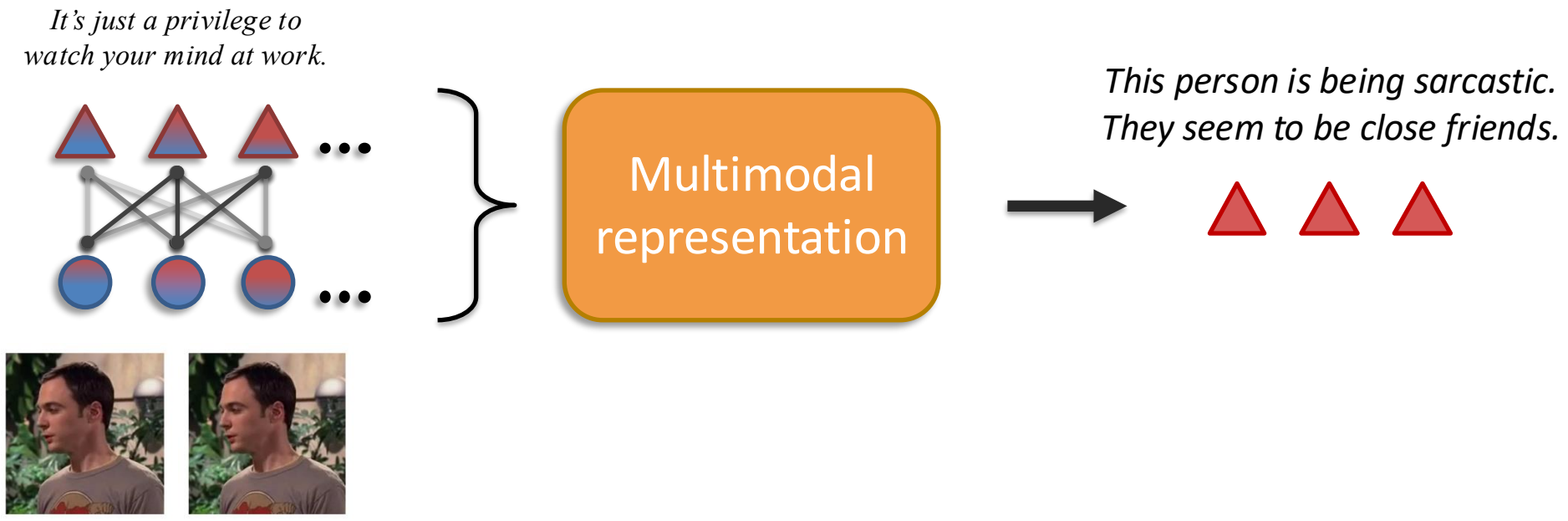
plant

[Kim et al., ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML 2021]

multisensory intelligence

Large Multimodal Models

Part 2: Adapting large language models for multimodal text generation

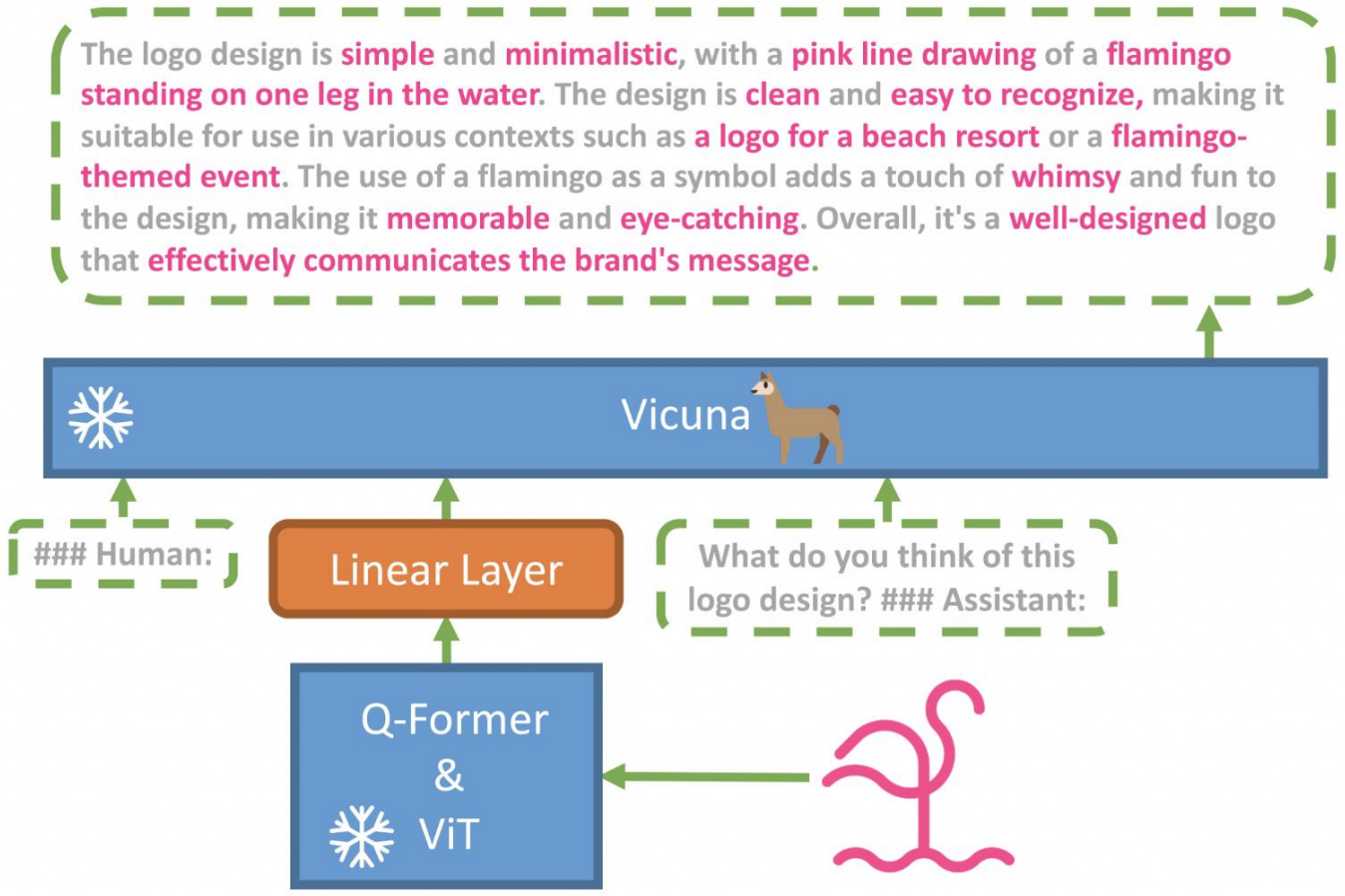


Adapting Large Multimodal Models

Mini-GPT4

Stage 1: **Alignment** using paired image-text data.

Stage 2: **Instruction tuning** using image + text instructions and example completions.



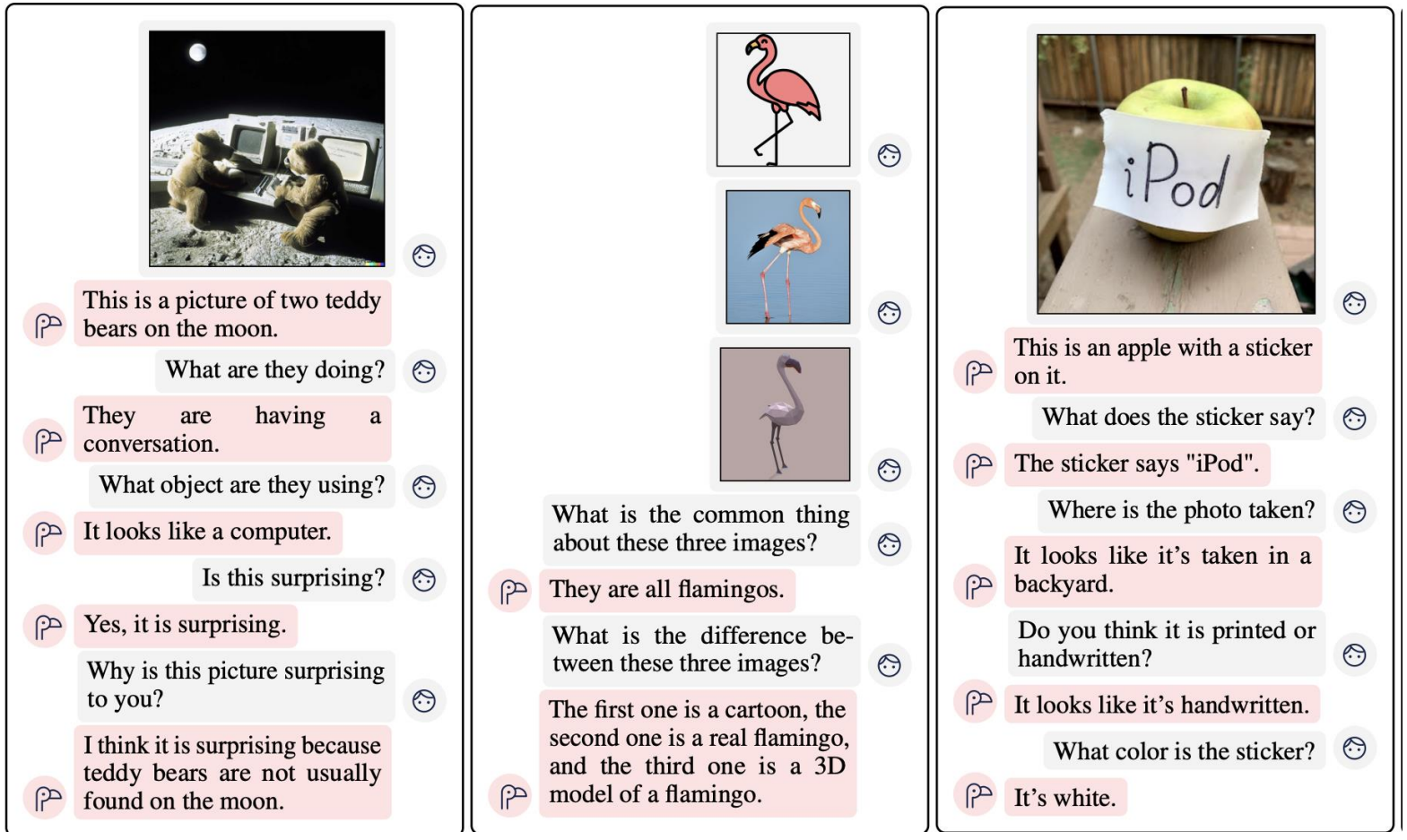
The architecture of MiniGPT-4.

[Zhu et al., MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models. 2023]

multisensory intelligence

Adapting Large Language Models

Flamingo

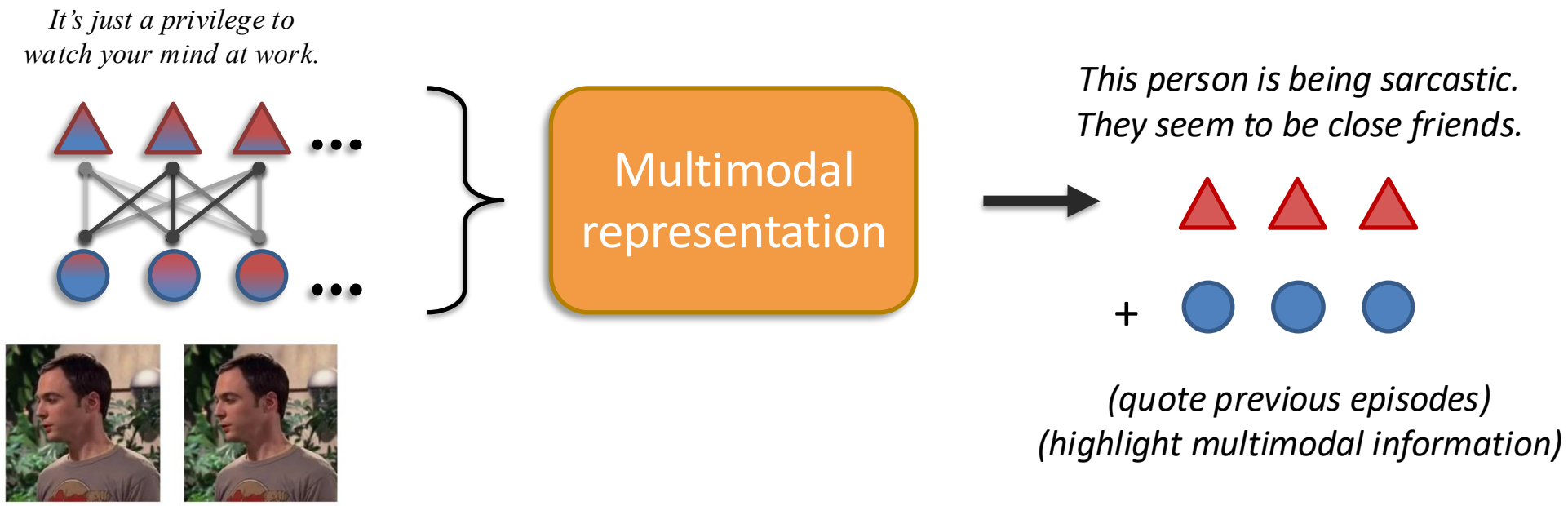


[Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022]

multisensory intelligence

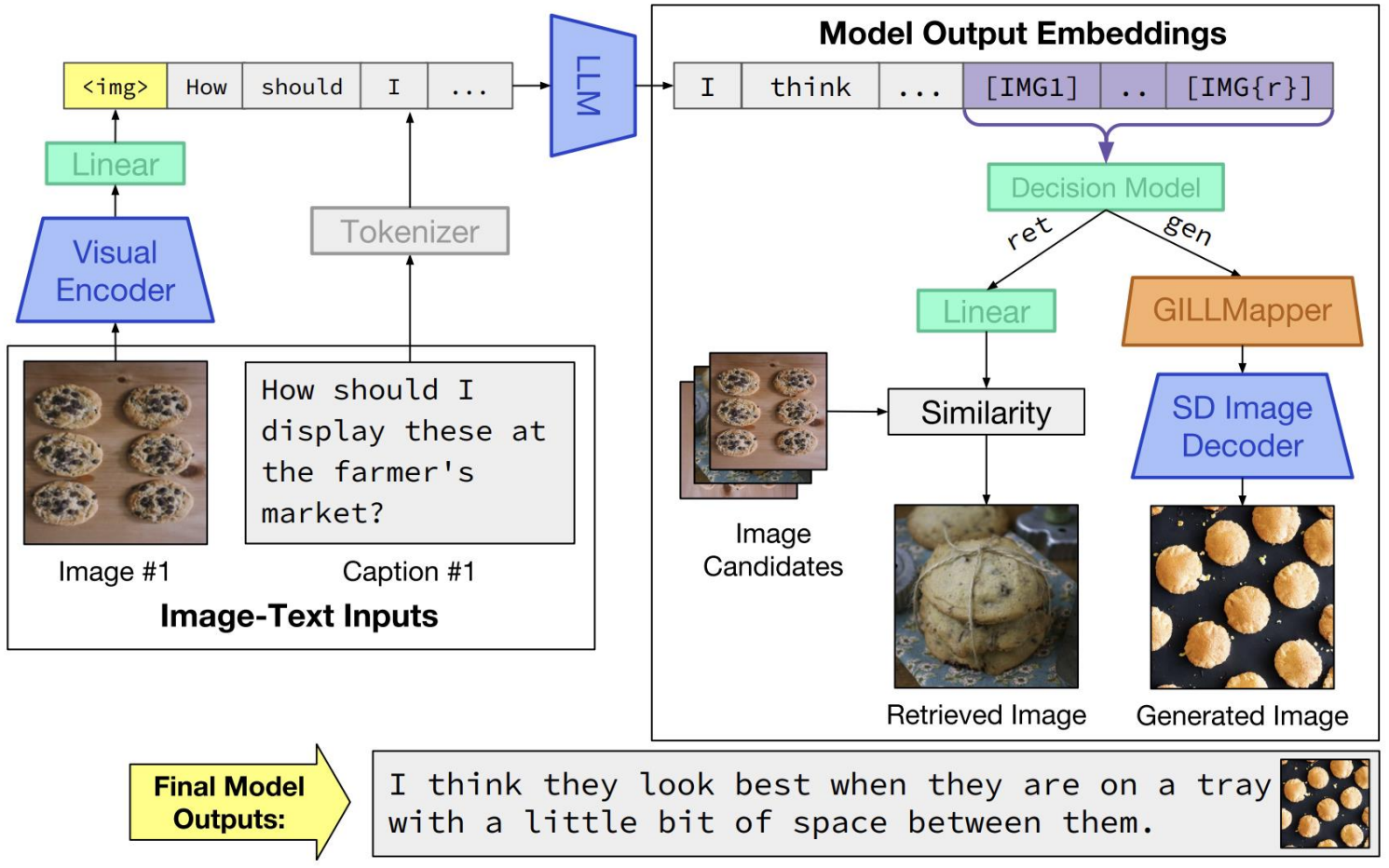
Large Multimodal Models

Part 3: Enabling text and image generation



Grounding LMs for Multimodal Generation

Large multimodal models with image generation



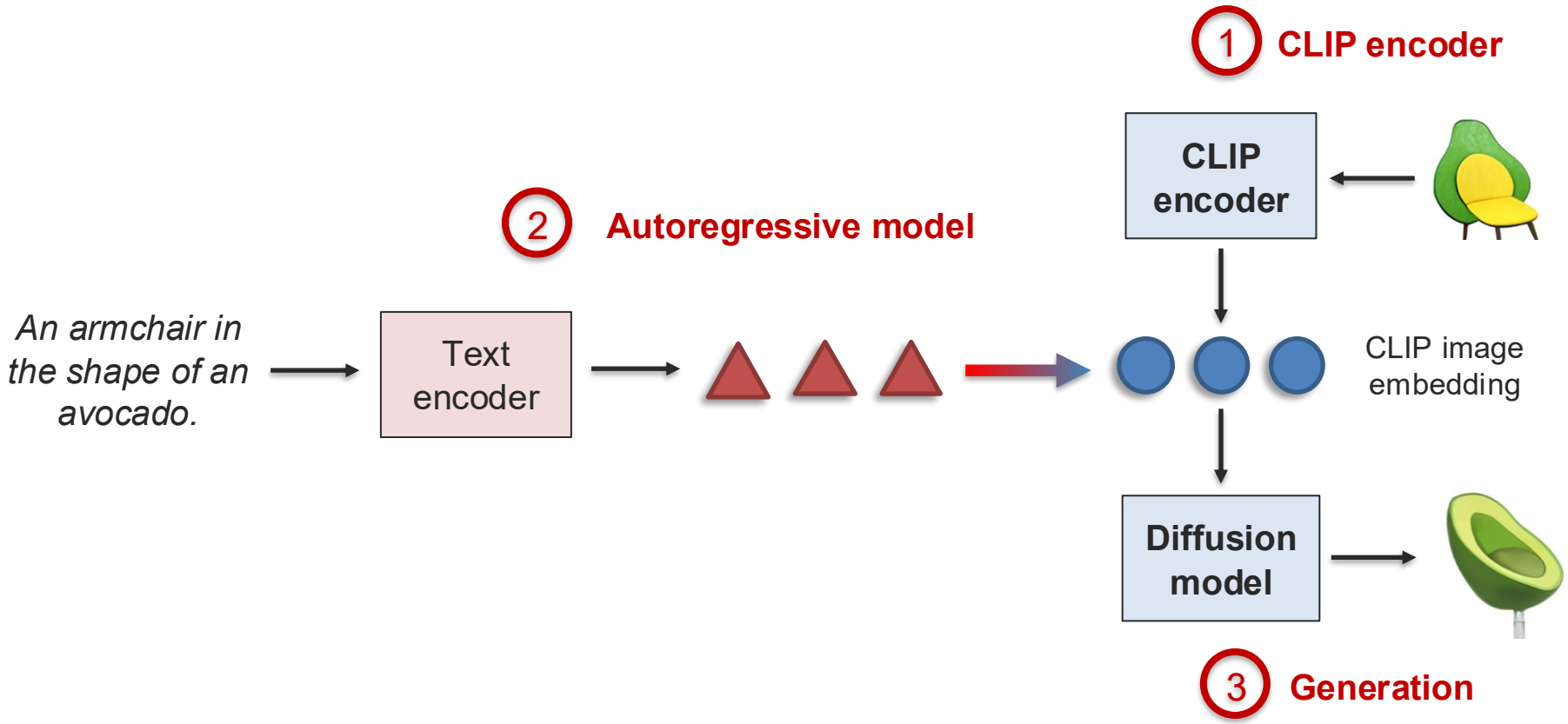
[Koh et al., Generating Images with Multimodal Language Models. NeurIPS 2023]

multisensory intelligence

From Text to Multimodal Generation

Directly training diffusion models with conditional information

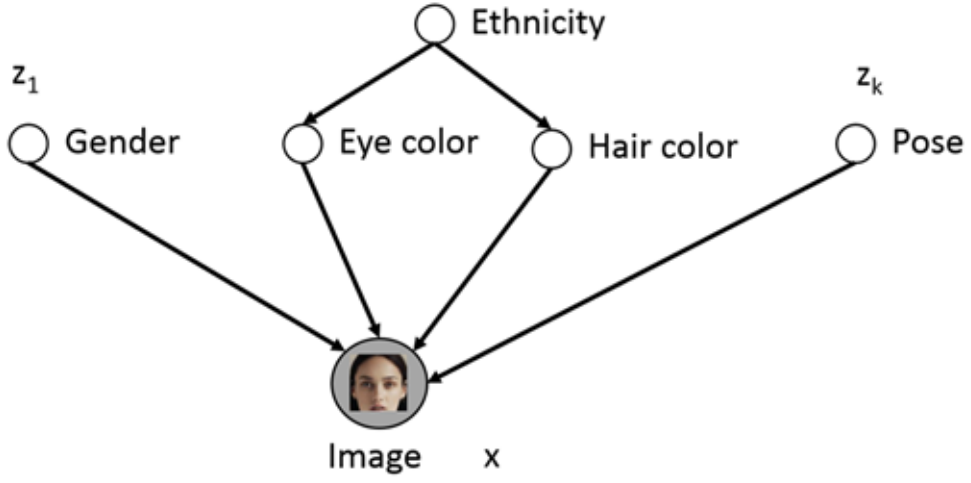
Conditional latent variables are pretrained CLIP embeddings, then diffusion model to generate image.



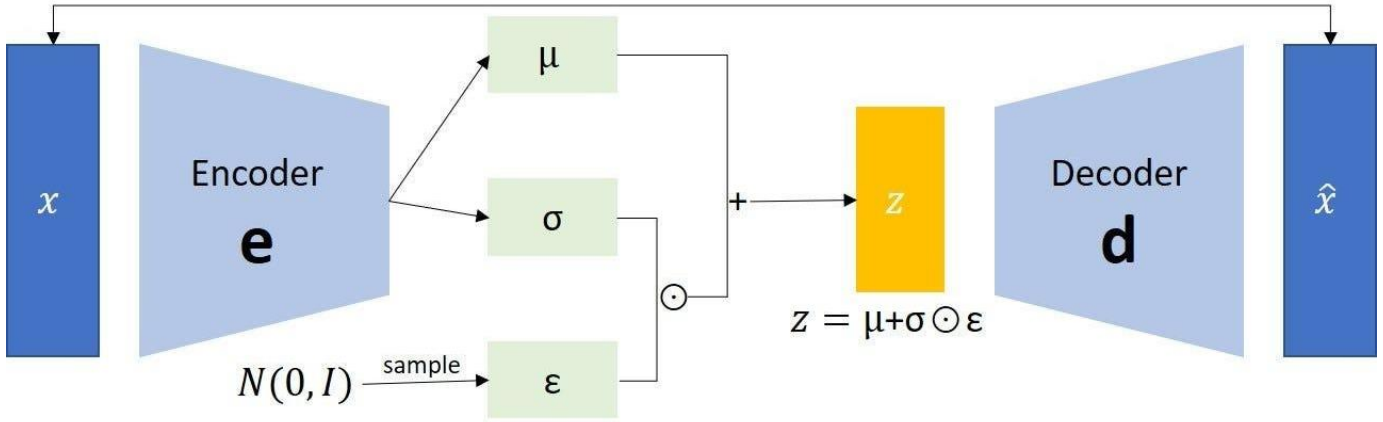
[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

Latent Variable Generative Models

- Lots of variability in images **x** due to gender, eye color, hair color, pose, etc.
- However, unless images are annotated, these factors of variation are not explicitly available (latent).
- Idea: explicitly model these factors using latent variables **z**



Learning Parameters of VAEs



Generative/decoder

$p(x|z;\theta)$

$q(z|x;\phi)$

Inference/encoder

1. Take a datapoint x_i.
2. Map it to μ, σ using $q_\phi(z|x_i)$. **encoder**
3. Sample $\epsilon \sim N(0, I)$ and compute $\hat{z} = \mu + \sigma\epsilon$. **re-parameterize**
4. Reconstruct $\hat{x}$ by sampling from $p(x|\hat{z};\theta)$. **decoder**

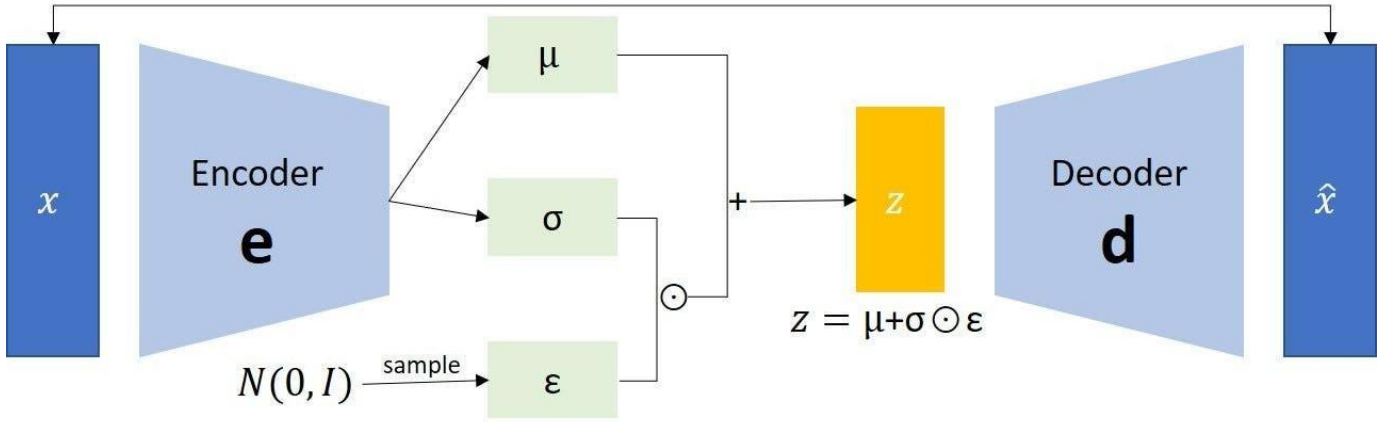
$$\mathcal{L}(\mathbf{x};\theta,\phi) = \underbrace{E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z};\theta)]}_{\text{reconstruction}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))}_{\text{prior over z, standard Gaussian}}$$

reconstruction
x -> z -> x

prior over z,
standard Gaussian

Prior on z: $\mathbf{z} \sim \mathcal{N}(0, I)$

Learning Parameters of VAEs

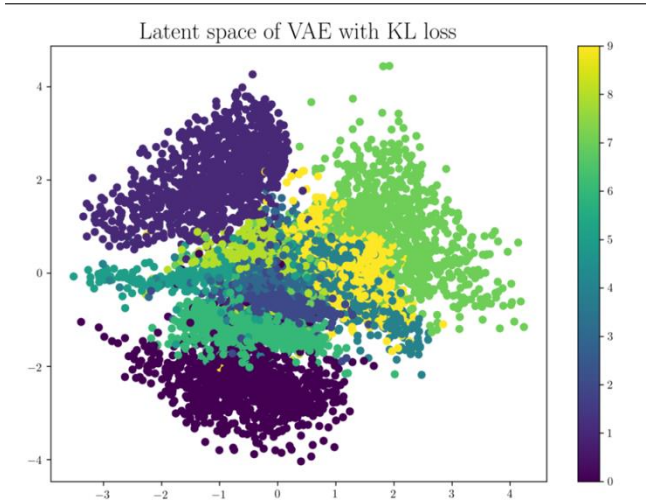


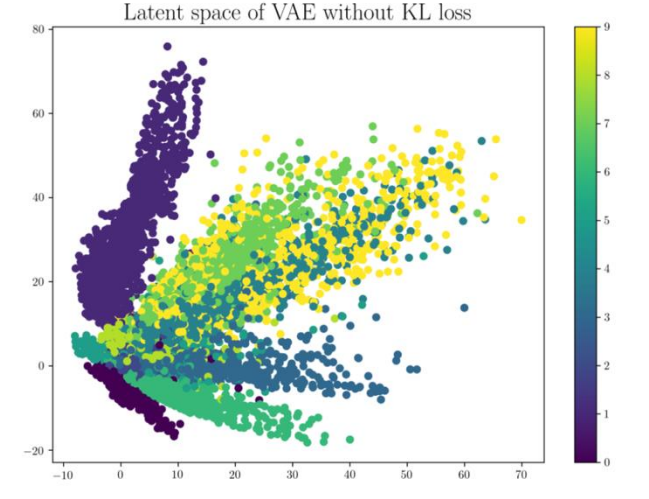
1. Take a datapoint x_i.
2. Map it to μ, σ using $q_\phi(z|x_i)$. **encoder**
3. Sample $\epsilon \sim N(0, I)$ and compute $\hat{z} = \mu + \sigma\epsilon$. **re-parameterize**
4. Reconstruct $\hat{x}$ by sampling from $p(x|\hat{z}; \theta)$. **decoder**

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \underbrace{E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}; \theta)]}_{\text{reconstruction}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))}_{\text{prior over z, standard Gaussian}}$$

reconstruction
x -> z -> x

prior over z,
standard Gaussian



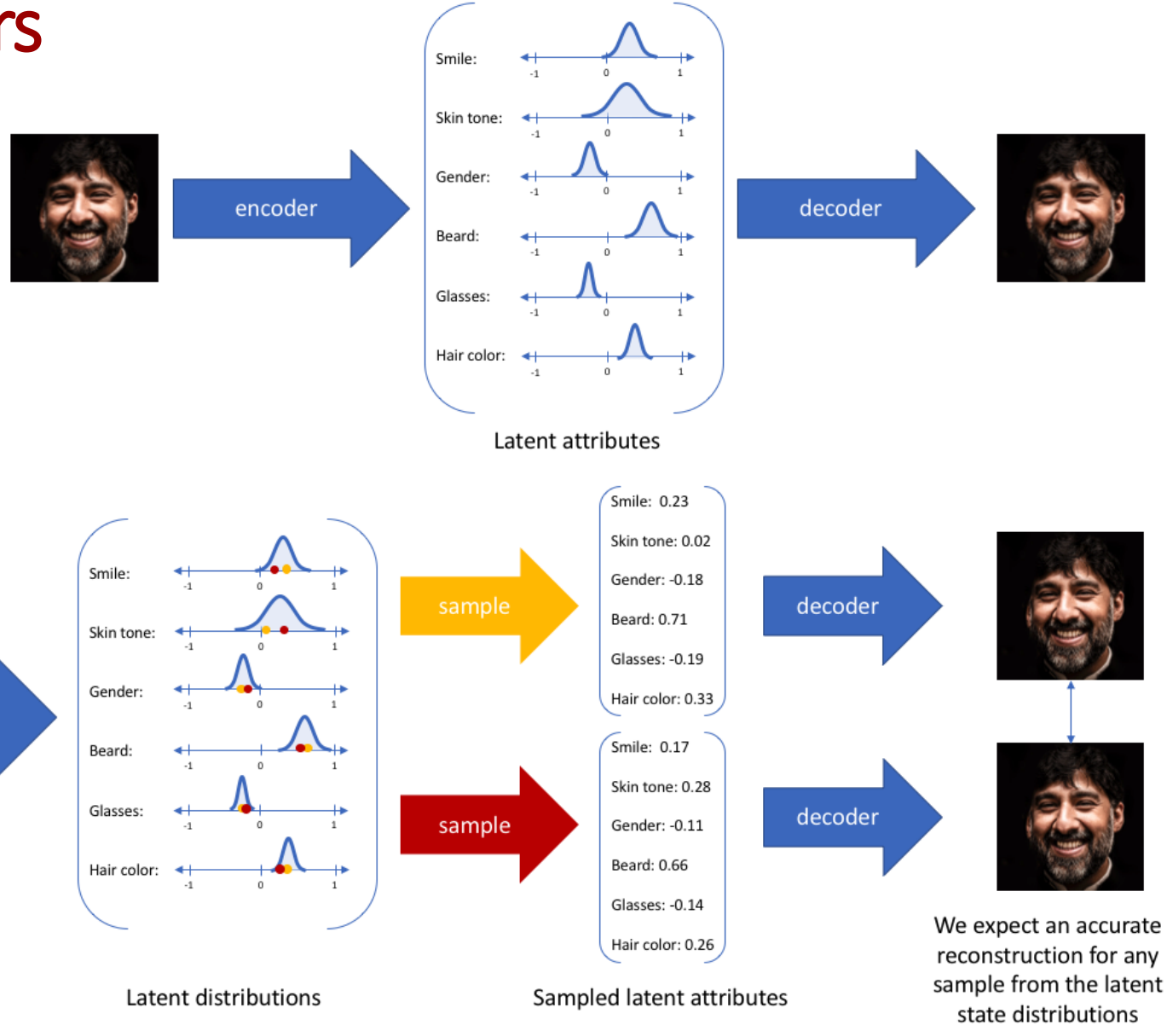


Prior on z: $\mathbf{z} \sim \mathcal{N}(0, I)$

Variational Autoencoders

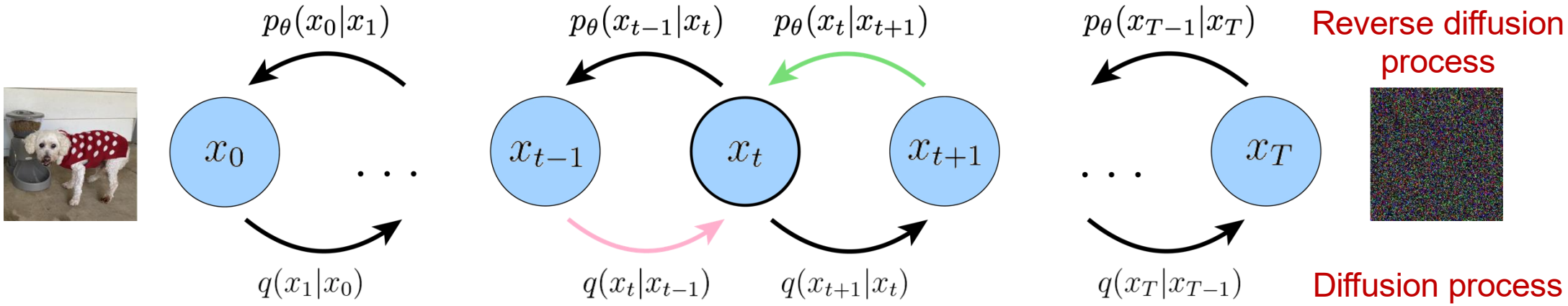
Key ideas:

1. Encoder decoder structure.
2. Simple latent variable z.
3. Complex p(x|z) decoder via neural networks.
4. Reconstruction objective.
5. Prior over latent variable z:
   -- smoother latent space permits sampling
   -- disentangles different dimensions of z



From VAEs to Diffusion Models

Generative modeling via denoising



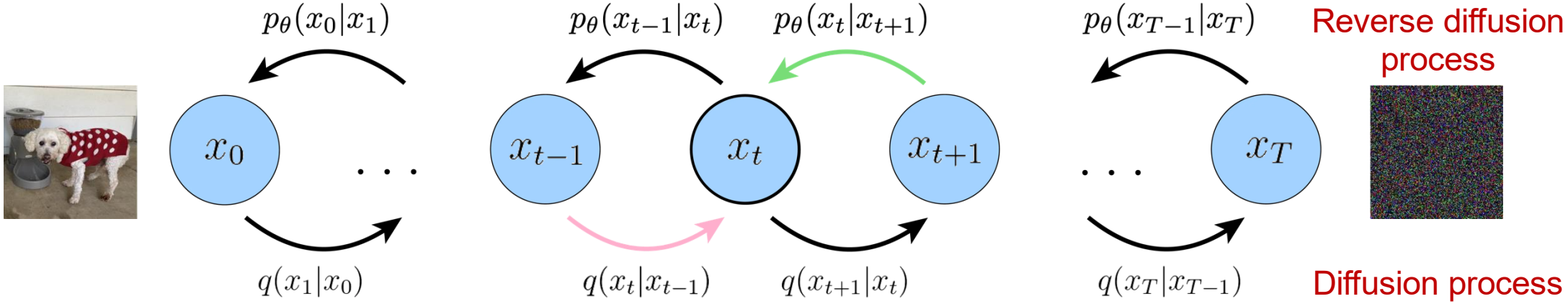
Similar to variational autoencoder, but:

1. The latent dimension is exactly equal to the data dimension.
2. Encoder q is not learned, but pre-defined as a Gaussian distribution centered around the output of previous timestep.
3. Gaussian parameters of latent encoders vary over time such that distribution of final latent is a standard Gaussian.

[Tutorial by Calvin Luo and Yang Song]

Learning Diffusion Models

Key idea: use variational inference



$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \,||\, p(\boldsymbol{x}_T))}_{\text{prior matching term}}$$

Multi-level VAE!

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

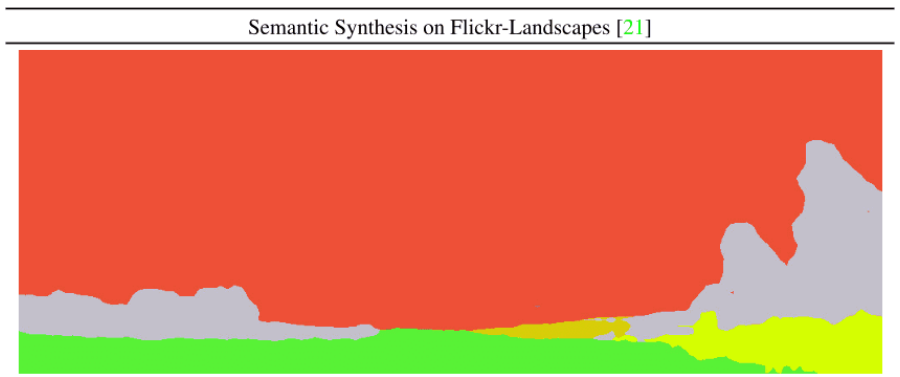
[Tutorial by Calvin Luo and Yang Song]

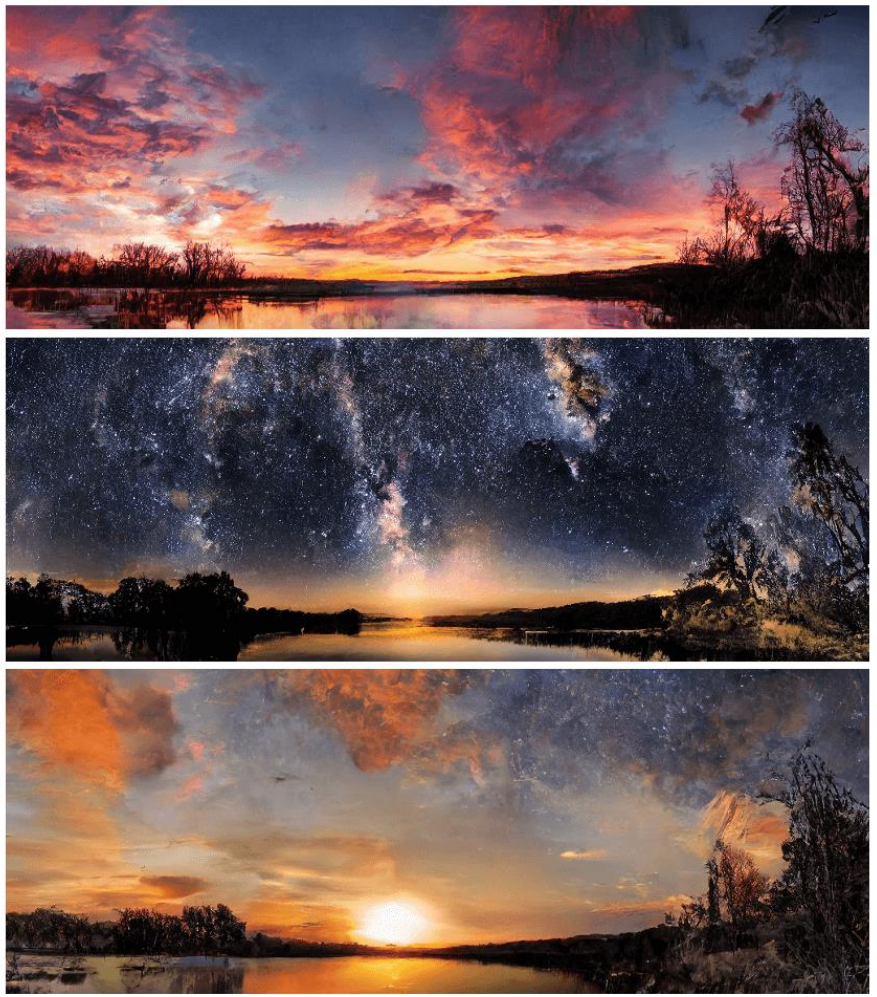
multisensory intelligence

Comparisons of Generative Models

Feature	VAE	Diffusion Models	Flow Matching
Core Idea	Encode/decode with latent noise	Add noise and learn to reverse it	Learn a continuous flow from noise
Training Objective	Minimize reconstruction + KL loss	Learn the score (gradient) of data	Match a vector field (ODE-based)
Noise Handling	Noise in latent space	Progressive noise over time	Start from noise, smooth transform
Sampling Speed	Very fast (one pass)	Slow (many denoising steps)	Faster (solving an ODE)
Advantages	Simple, fast, interpretable	Very high-quality outputs	High quality + faster than diffusion
Disadvantages	Blurry samples, limited expressiveness	Expensive, slow sampling	Newer, still developing
Key Examples	VAE (2013), β-VAE	DDPM, Stable Diffusion	Flow Matching (2023), Rectified Flow

Text-to-Image Generation with Latent Diffusion



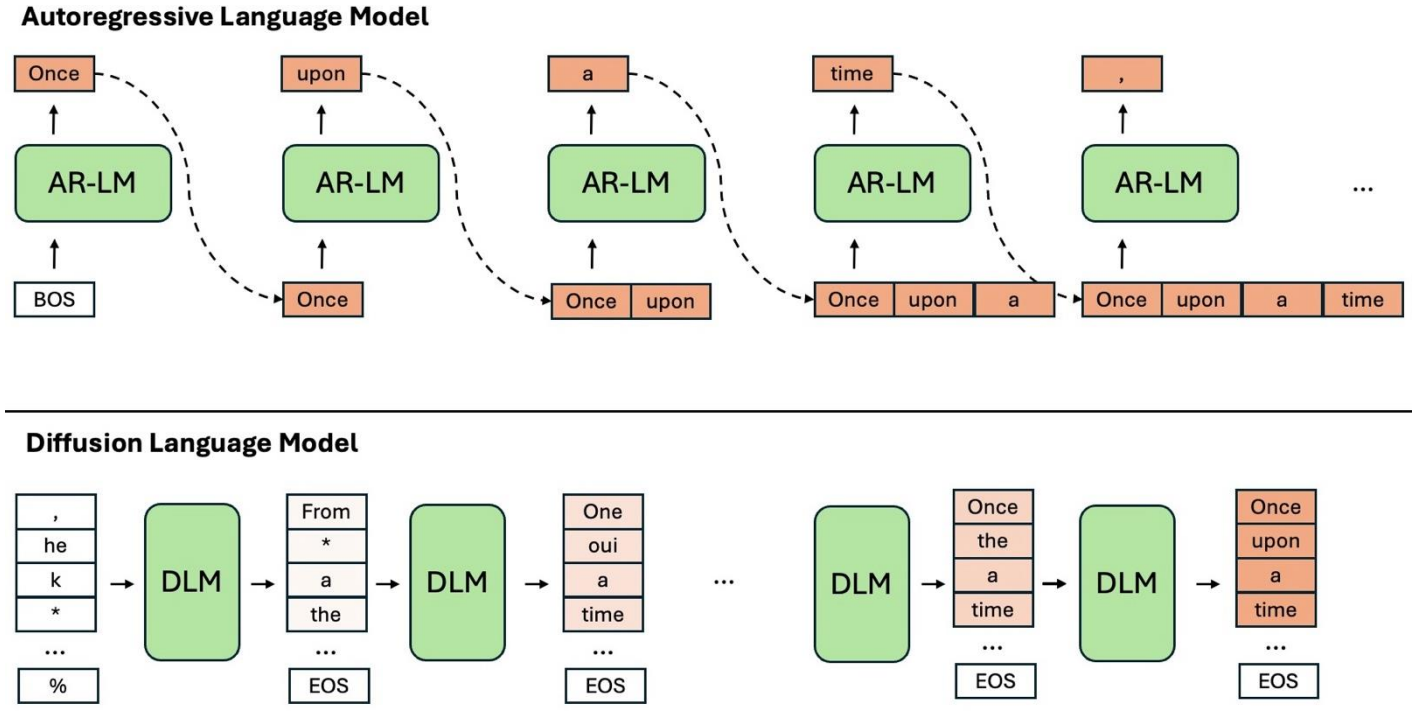


[Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022]

Diffusion Models for Text

Diffusion Language Models: Instead of going token by token, they iteratively refine and predict the *whole* sequence from a noise following a non-autoregressive decoding process.

Key advantages:
- Can fix earlier mistakes
- More controllable
- More diversity
- Faster



[Zhu et al., What are Diffusion Language Models? 2025]

multisensory intelligence

Midterm Structure

6 problems, 60 minutes.

5 problems:

1. MCQ (10, 2 points each = 20 points)
2. Short answers (4, 5 points each = 20 points)
3. Multimodal Fusion (20 points)
4. Multimodal LLMs (20 points)
5. Multimodal Generation (20 points)
6. Bonus open questions (10 points)

Assignments for This Coming Week

HW3 due today.

Project proposals will be graded and released this week.

Project midterm instructions will be out this week.

Midterm this Thursday.