Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.2: New Research Directions

Louis-Philippe Morency

* Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.

Administrative Stuff

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/30 & 9/1	**Course introduction** • Multimodal core challenges • Course syllabus	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/6 & 9/8 Read due: 9/9	**Basic concepts: neural networks** • Loss functions and neural networks • Gradient and optimization	**Unimodal representations** • Dimensions of heterogeneity • Visual representations
Week 3 9/13 & 9/15 Read due: 9/16 Proj. Due: 9/14	**Unimodal representations** • Language representations • Signals, graphs and other modalities	**Multimodal representations** • Cross-modal interactions • Multimodal fusion
Week 4 9/20 & 9/22 *Proj. due: 9/25*	**Multimodal representations** • Coordinated representations • Multimodal fission	**Multimodal alignment** • Explicit alignment • Multimodal grounding
Week 5 9/27 & 9/29 Read due: 9/30	***Project hours (Research ideas)***	**Aligned representations** • Self-attention transformer models • Masking and self-supervised learning
Week 6 10/4 & 10/6 *Proj. due: 10/9*	**Multimodal aligned representations** • Multimodal transformers • Video and graph representations	**Multimodal Reasoning** • Structured and hierarchical models • Memory models

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/11 & 10/13 Read due: 10/14	**Multimodal Reasoning** • Reinforcement learning • Discrete structure learning	**Multimodal Reasoning** • Logical and causal inference • External knowledge
Week 8 10/18 & 10/20	***Fall Break – No lectures***	
Week 9 10/25 & 10/27 *Proj. due: 10/30*	**Generation** • Translation, summarization, creation • Generative models: VAEs	**Generation** • GANs and diffusion models • Model evaluation and ethics
Week 10 11/1 & 11/3	***Project presentations (midterm)***	***Project presentations (midterm)***
Week 11 11/8 & 11/10 Read due: 11/12	**Transference** • Multi-task • Modality transfer	**Transference** • Multimodal co-learning • Co-training
Week 12 11/15 & 11/17 Read due: 11/21	**Quantification** • Heterogeneity and interactions • Biases and fairness	**New research directions** • Recent approaches in multimodal ML

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 13 11/22 & 11/24	***Thanksgiving Week – No Class –***	
Week 14 11/30 & 12/2	***Language, Vision, and Actions*** • Robots, navigation and embodied AI • Guest lecturer: Yonatan Bisk	***Multimodal Language Grounding*** • Grounded semantics and pragmatics • Guest lecturer: Daniel Fried
Week 15 12/6 & 12/8 *Proj. due: 12/11*	***Project presentations (final)***	***Project presentations (final)***

Final assignment due on Sunday 12/11

Final Project Report (Due Sunday 12/11 at 8pm)

Main goals:

1. Produce a research paper which will motivate your research problem, describe the prior work, present your research contributions, explain the details of your experiments, and discuss your results.
2. Novel research ideas (N-1 new ideas for N students)
   - Novel algorithm
   - Novel application
3. Incorporate feedback from previous milestones
4. Compare to multimodal baselines from midterm report
   1. Did the proposed ideas solve the errors highlighted in error analysis?
   2. Broader implications of proposed ideas.

Final Project Presentations (Tuesday 12/6 and Thursday 12/8)

Main objective:

- Present your research ideas to the broad community
- Focus on only one (or few) of your new research ideas
- All students should present and answer questions
- All presentations are in person (no remote presentations)
- Non-presenting students will be asked to give feedback

Presentation length:

- 30-seconds elevator pitch
- 4-minute full presentation – all students should present

- Best poster award each day! (1 extra day for final report)

Last Reading Assignment

- Four main steps for the reading assignments
  - Monday 8pm: Official start of the assignment
  - Wednesday 8pm: Select your paper
  - **Friday 8pm:** Post your summary
  - **Monday 8pm:** Post your extra comments

Advanced Topics in Multimodal ML (11-877)

Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2022 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantifications. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

- **Time:** Friday 10:10-11:30 am
- **Location:** Virtual for the first 2 weeks (find zoom link in piazza), GHC 5222 thereafter
- **Discussion and Q&A:** Piazza
- **Assignment submissions:** Canvas (for registered students only)
- **Contact:** Students should ask all course-related questions on Piazza, where you will also find announcements.

Instructor Louis-Philippe Morency
Email: morency@cs.cmu.edu

Instructor Amir Zadeh
Email: abagherz@cs.cmu.edu

Instructor Paul Liang
Email: pliang@cs.cmu.edu

1/28 Week 2: **Cross-modal interactions** [synopsis]

- What are the different ways in which modalities can interact with each other in multimodal tasks? Can we formalize a taxonomy of such cross-modal interactions, which will enable us to compare and contrast them more precisely?
- What are the design decisions (aka inductive biases) that can be used when modeling these cross-modal interactions in machine learning models?
- What are the advantages and drawbacks of designing models to capture each type of cross-modal interaction? Consider not just prediction performance, but tradeoffs in time/space complexity, interpretability, etc.
- Given an arbitrary dataset and prediction task, how can we systematically decide what type of cross-modal interactions exist, and how can that inform our modeling decisions?
- Given trained multimodal models, how can we understand or visualize the nature of cross-modal interactions?

- Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!
- What Does BERT with Vision Look At?
- Multiplicative Interactions and Where to Find Them
- Cooperative Learning for Multi-view Analysis
- Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers
- Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks

2/4 Week 3: **Multimodal co-learning** [synopsis]

- What are the types of cross-modal interactions involved to enable such co-learning scenarios where multimodal training ends up generalizing to unimodal testing?
- What are some design decisions (inductive bias) that could be made to promote transfer of information from one modality to another?
- How do we ensure that during co-learning, only useful information is transferred, and not some undesirable bias? This may become a bigger issue in low-resource settings.
- How can we know if co-learning has succeeded? Or failed? What approaches could we develop to visualize and probe the success of co-learning?
- How can we formally, empirically, or intuitively measure the additional information provided by auxiliary modality? How can we design controlled experiments to test these hypotheses?
- What are the advantages and drawbacks of information transfer during co-learning? Consider not just prediction performance, but also tradeoffs with complexity, interpretability, fairness, etc.

- Multimodal Prototypical Networks for Few-shot Learning
- SMIL: Multimodal Learning with Severely Missing Modality
- Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions
- Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision
- What Makes Multi-modal Learning Better than Single (Provably)
- Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities
- Zero-Shot Learning Through Cross-Modal Transfer
- 12-in-1: Multi-Task Vision and Language Representation Learning
- A Survey of Reinforcement Learning Informed by Natural Language

https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

New Course: Artificial Social Intelligence (11-866)

- Seminar-style course (reading discussions)
- Fridays 3pm
- Two versions:
  - 6-credit version: reading discussions only
  - 12-credit version: + independent study (team course project)
- Open to all students (but only registered students)

- More details in the coming weeks…

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.2: New Research Directions

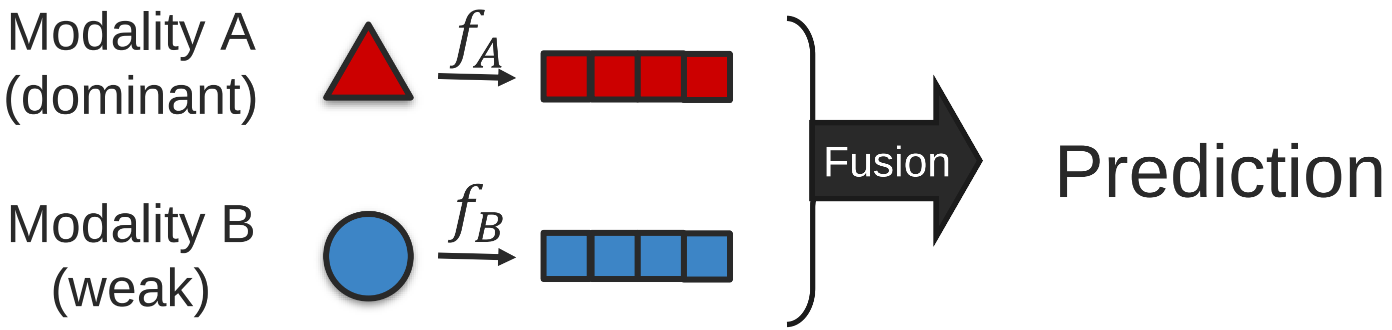
Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Representation Fusion

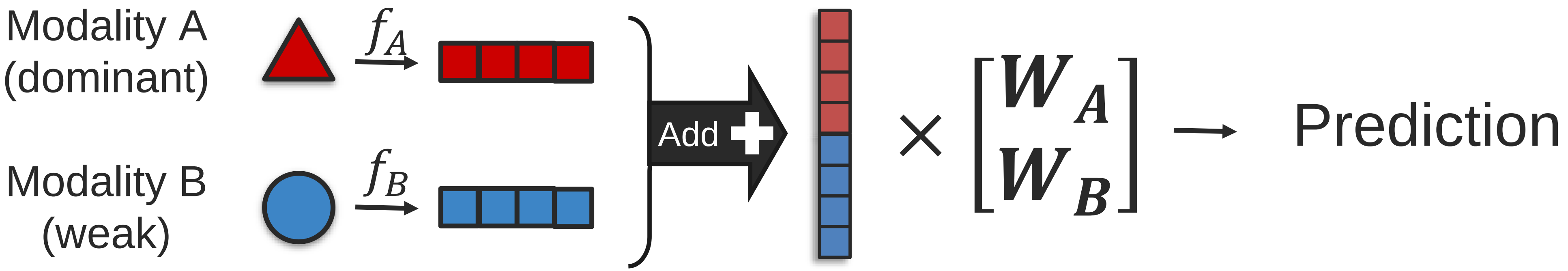
Balanced Multimodal Learning via On-the-fly Gradient Modulation

Modality dominance Under-optimized unimodal representations
(even when multimodal model performs better)



Balanced Multimodal Learning via On-the-fly Gradient Modulation

Modality dominance Under-optimized unimodal representations
(even when multimodal model performs better)



$$\hat{\boldsymbol{y}} = f(x_A, x_B) = \boldsymbol{w}_A \cdot f_A(\boldsymbol{x}_A) + \boldsymbol{w}_{\boldsymbol{B}} \cdot f_B(\boldsymbol{x}_B) + b$$

Problem: The dominant modality (with largest weights $\boldsymbol{W_A}$ or $\boldsymbol{W_B}$) gets most of the gradient updates

Balanced Multimodal Learning via On-the-fly Gradient Modulation

Problem: The dominant modality (with largest weights $\boldsymbol{W_A}$ or $\boldsymbol{W_B}$) gets most of the gradient updates

$$W_A^{t+1} = W_A^t - \eta \cdot \nabla_{W_A} L$$

$$= W_A^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^i)} \frac{\partial f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^i)}{W_A}$$

The gradient for each modality is weighted by the joint discriminative loss

$$= W_A^t - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \boxed{\frac{\partial L}{\partial f\left(\boldsymbol{x}_A^i, \boldsymbol{x}_B^i\right)}} f_A(\boldsymbol{x}_A)$$

This joint discriminative loss is dependent on the weights $\boldsymbol{W_A}$ or $\boldsymbol{W_B}$

$$\hat{\boldsymbol{y}} = f(x_A, x_B) = \boldsymbol{w}_A \cdot f_A(\boldsymbol{x}_A) + \boldsymbol{w_B} \cdot f_B(\boldsymbol{x}_B) + b$$

Balanced Multimodal Learning via On-the-fly Gradient Modulation

Problem: The dominant modality (with largest weights $\boldsymbol{W_A}$ or $\boldsymbol{W_B}$) gets most of the gradient updates

Solution: Weight the gradient based on its contribution to the learning objective

$$W_A^{t+1} = W_A^t - \eta \cdot k_t^A \cdot \nabla_{W_A} L$$

$$\downarrow$$

On-the-fly Gradient Modulation (OGM)

monitor discrepancy of each modality's contribution to the objective

Balanced Multimodal Learning via On-the-fly Gradient Modulation

Solution: Weight the gradient based on its contribution to the learning objective

$$W_A^{t+1} = W_A^t - \eta \cdot k_t^A \cdot \nabla_{W_A} L$$

Hyper-parameter

Ratio computed over a batch B_t

$$\rho_t^v = \frac{\sum_{i \in B_t} s_i^v}{\sum_{i \in B_t} s_i^a}.$$

for modality u,

$$k_t^u = \begin{cases} 1 - \tanh(\alpha \cdot \rho_t^u) & \rho_t^u > 1 \\ 1 & \text{others,} \end{cases}$$

contribution from a

$$\overbrace{s_i^a = \sum_{k=1}^{M} 1_{k=y_i} \cdot \text{softmax}(W_t^a \cdot \varphi_t^a(\theta^a, x_i^a) + \frac{b}{2})_k,}^{f_A}$$

contribution from v

$$\underbrace{s_i^v = \sum_{k=1}^{M} 1_{k=y_i} \cdot \text{softmax}(W_t^v \cdot \varphi_t^v(\theta^v, x_i^v) + \frac{b}{2})_k,}_{f_B}$$

Balanced Multimodal Learning via On-the-fly Gradient Modulation

Dataset	CREMA-D		VGGSound	
Method	Acc	mAP	Acc	mAP
Audio-only	52.5	54.2	44.3	48.4
Visual-only	41.9	43.0	31.0	34.3
Baseline	50.8	52.6	48.4	51.7
Concatenation	51.7	53.5	49.1	52.5
Summation	51.5	53.5	49.1	52.4
FiLM [32]	50.6	52.1	48.5	51.6
Baseline†	54.4	56.2	50.1	53.5
Concatenation†	61.9	63.9	**50.6**	**53.9**
Summation†	**62.2**	**64.3**	50.4	53.6
FiLM†	55.6	57.4	50.0	52.9

OGM-GE

Achieve considerable improvement over common fusion methods on different multimodal tasks

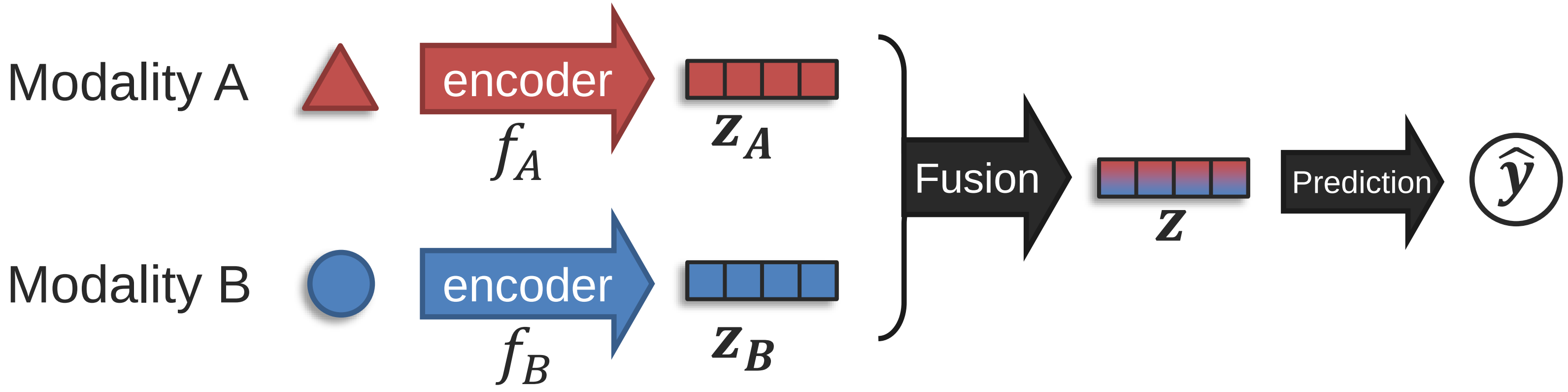
Extra contribution:

Generalization Enhancement (GE)

introduce extra dynamic Gaussian noise to avoid generalization drop

Representation Fusion

Geometric Multimodal Contrastive Representation Learning

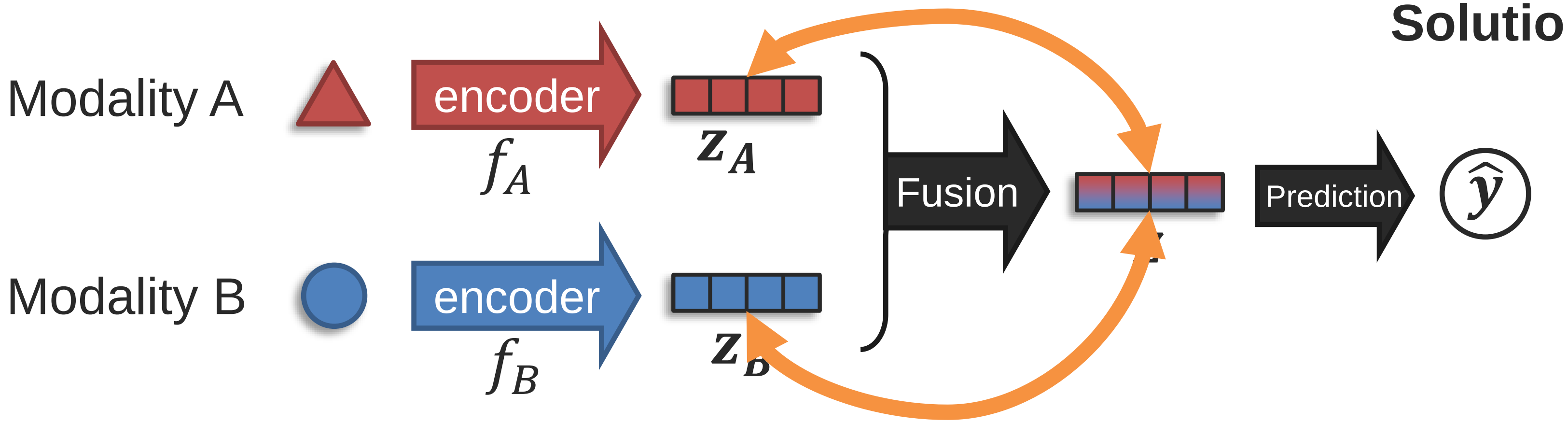


Challenge: To help with robustness, we would like the unimodal representations (z_A and z_B) to be close to the multimodal representation z

But in practice, they end up not being aligned!

(related to the "heterogeneity" gap)

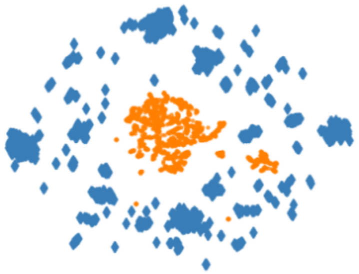
Geometric Multimodal Contrastive Representation Learning



Solution: Align the unimodal and multimodal representations through contrastive learning

Geometrically misaligned

MVAE

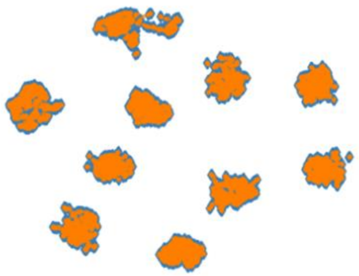


MFM

Complete z

Visual z_A

Aligned

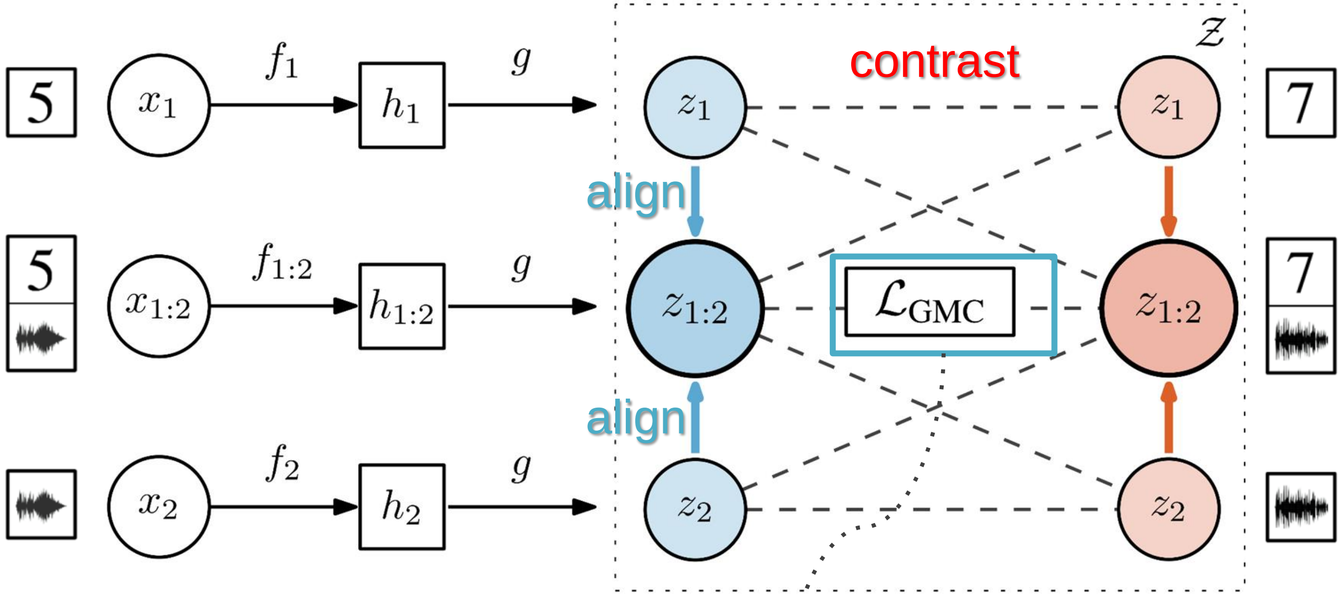


GMC

Proposed approach

Geometric Multimodal Contrastive Representation Learning

Geometric Multimodal Contrastive (GMC) learning:



positive pairs

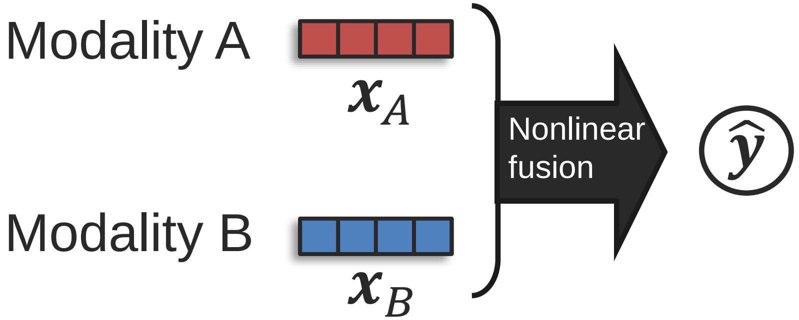
$$\mathcal{L}_{\text{GMC}}(\mathcal{B}) = \sum_{m=1}^{M}\sum_{i=1}^{B} -\log \frac{s_{m,1:M}(i,i)}{\Omega_m(i)}$$

$$\sum_{i\neq j}\left(s_{m,1:M}(i,j) + s_{m,m}(i,j) + s_{1:M,1:M}(i,j)\right)$$

negative pairs

Representation Fusion

Measuring Non-Additive Interactions



Nonlinear fusion:

$$\hat{y} = f(\boldsymbol{x}_A, \boldsymbol{x}_B)$$

Projection?

Additive fusion:

$$\hat{y}' = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$$

Projection from nonlinear to additive (using EMAP):

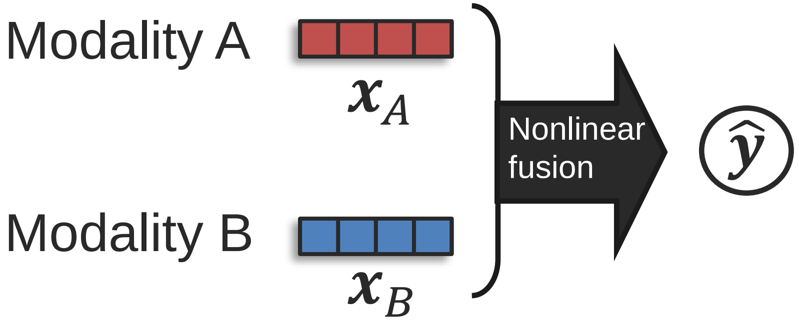
$$\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \underbrace{\mathbb{E}_{\boldsymbol{x}_B}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_A(\boldsymbol{x}_A)} + \underbrace{\mathbb{E}_{\boldsymbol{x}_A}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_B(\boldsymbol{x}_B)}$$

Modality A + Modality B

Additive fusion (approximation)

Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020 ⟶ introduced the EMAP method

Measuring Non-Additive Interactions



Nonlinear fusion:

$$\hat{y} = f(\boldsymbol{x}_A, \boldsymbol{x}_B)$$

EMAP projection

Additive fusion:

$$\hat{y}' = \hat{f}_A(\boldsymbol{x}_A) + \hat{f}_B(\boldsymbol{x}_B)$$

		I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Nonlinear	Neural Network	90.4	69.2	78.5	51.1	63.5	71.1	79.9
Polynomial	Polykernel SVM	**91.3**	**74.4**	**81.5**	50.8	–	72.1	**80.9**
Nonlinear	FT LXMERT	83.0	68.5	76.3	**53.0**	63.0	66.4	78.6
Nonlinear	↳ + Linear Logits	89.9	73.0	80.7	**53.4**	**64.1**	**75.5**	80.3
Additive	Linear Model	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
	Best Model	**91.3**	**74.4**	**81.5**	**53.4**	***64.2***	**75.5**	**80.9**
Additive	↳ + EMAP	**91.1**	**74.2**	**81.3**	51.0	***64.1***	**75.9**	**80.7**

Always a good baseline!

Differences are small!!!

Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020 ⟶ introduced the EMAP method

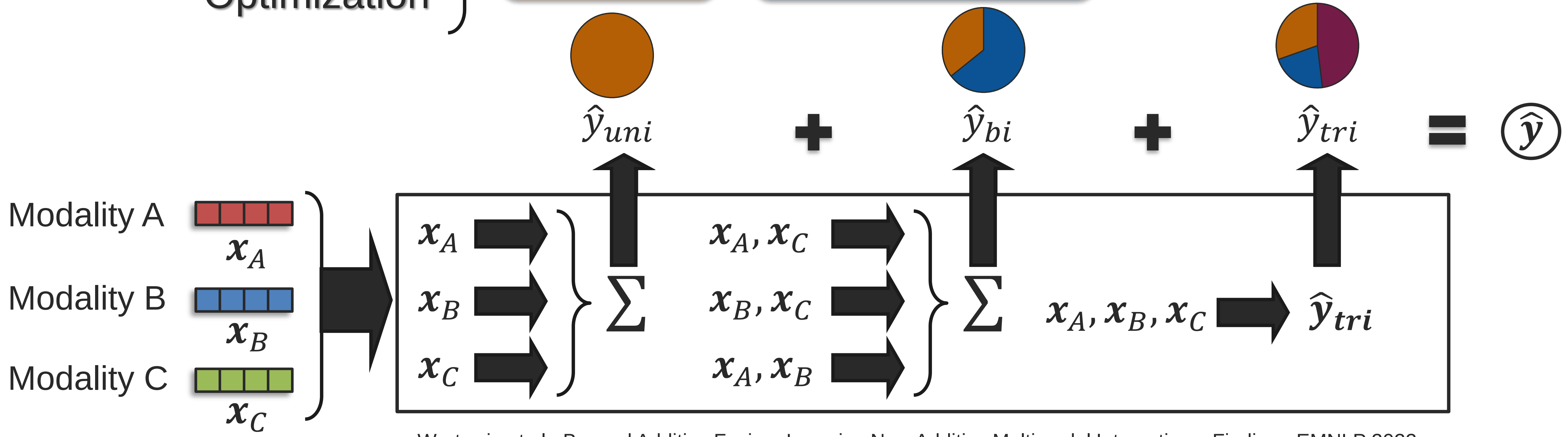
Learning Non-additive Bimodal and Trimodal Interactions

Idea: prioritize simpler interactions

Unimodal (additive) — residual — **Bimodal** (non-additive) — residual — **Trimodal** (non-additive)

Multimodal Residual Optimization

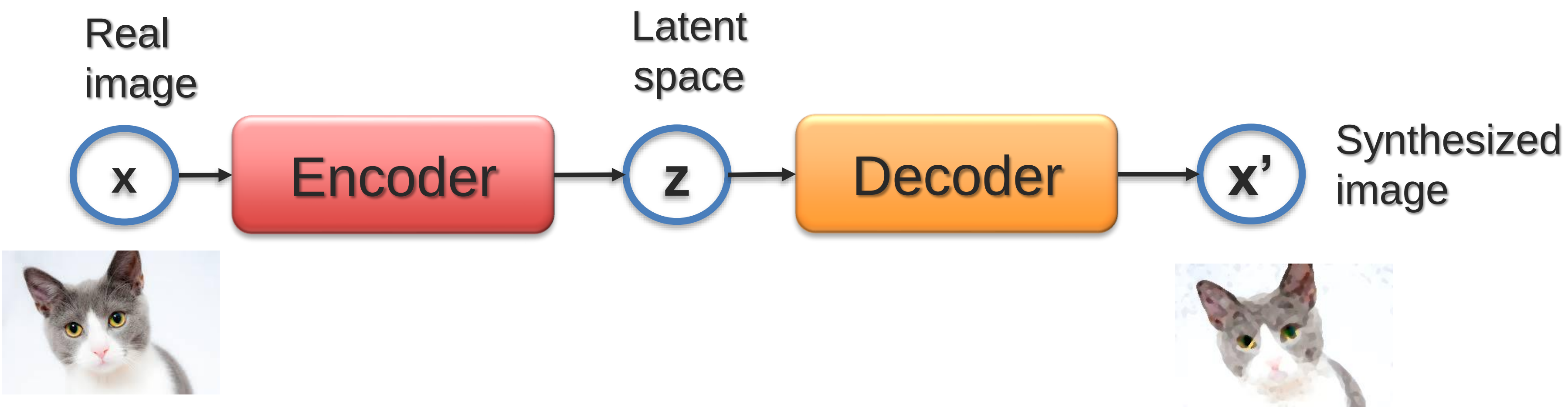
$$\boxed{\mathcal{L}(y, \hat{y}_{uni})} + \boxed{\mathcal{L}(\boxed{y - \hat{y}_{uni}}, \hat{y}_{bi})} + \mathcal{L}(\boxed{y - \hat{y}_{uni} - \hat{y}_{bi}}, \hat{y}_{tri})$$



Wortwein et al., Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions, Findings-EMNLP 2022

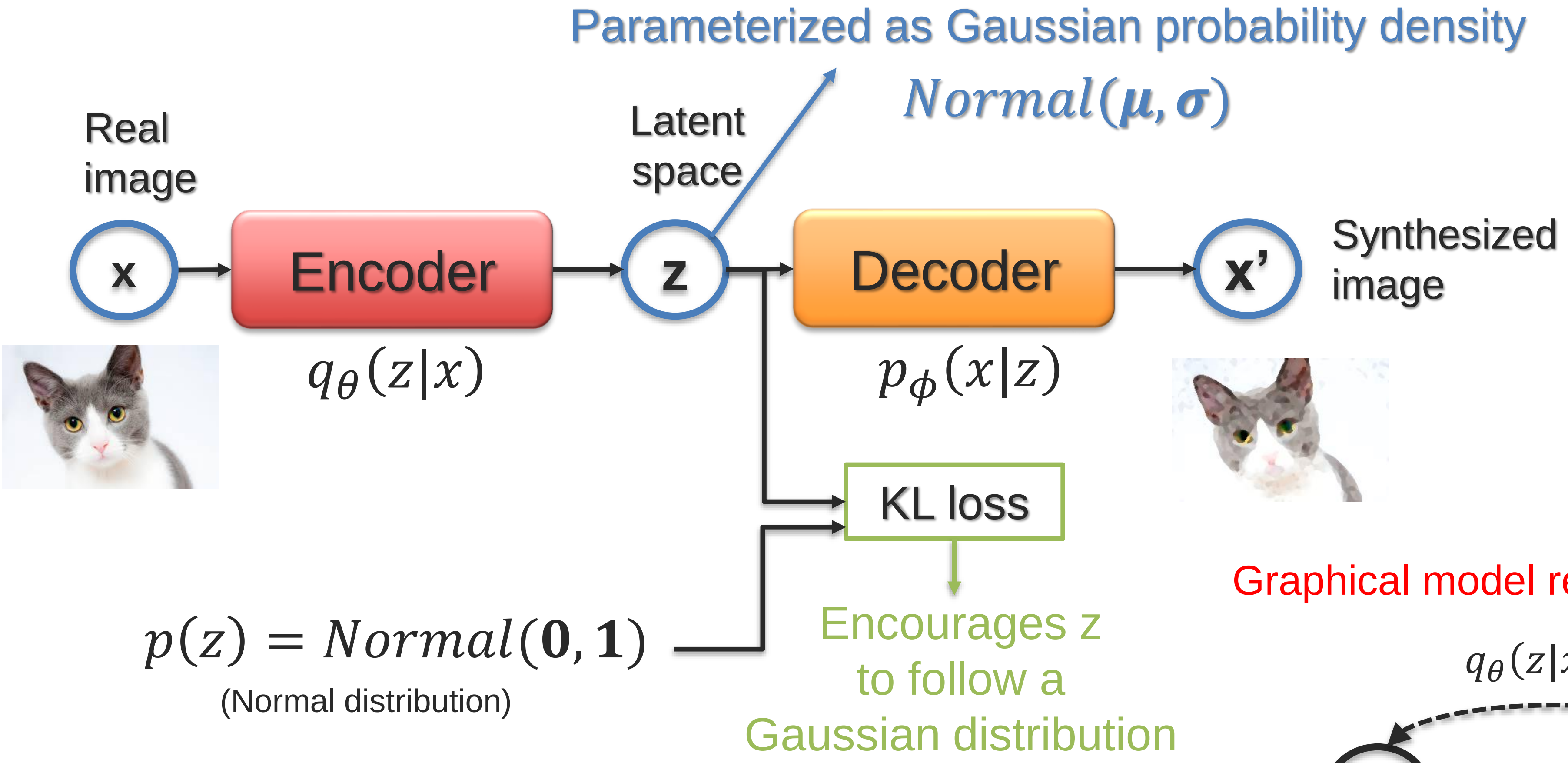
Representation Fusion, Transference and Generation

Auto-encoder

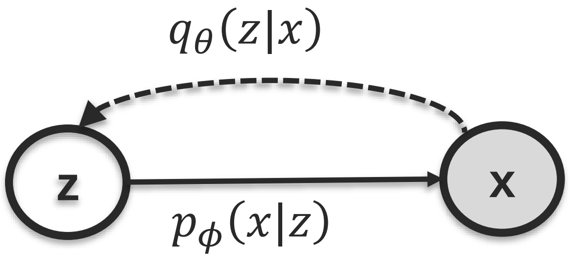


After learning this autoencoder,
can I input any z vector in the decoder?

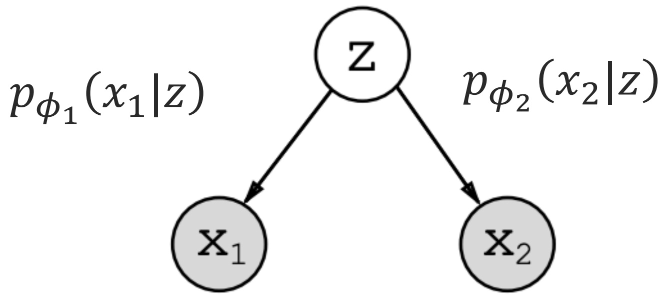
Variational Autoencoder



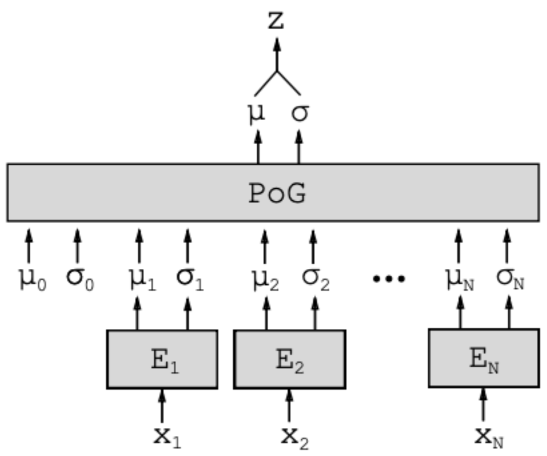
Graphical model representation?



Multimodal VAE (MVAE)



What will be the encoder?



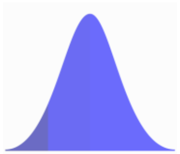
Product of expert (PoG) to combine the variational parameters from the unimodal encoders

[Wu, Mike, and Noah Goodman. “Multimodal Generative Models for Scalable Weakly-Supervised Learning.”, NIPS 2018]

Learning Multimodal VAEs

Joint distribution

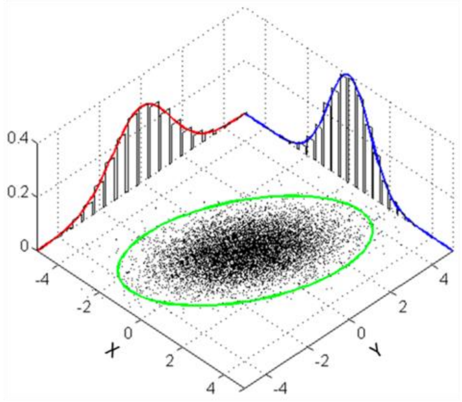
Encoded modality 1



⋮

Encoded modality *2*

Some combination (*e.g.* product, concatenation)



But what if one of the modalities is only **partially observed** at train/test time ?

What if we allow **each modality to help model the other(s)** ?

Learning Multimodal VAEs Through Mutual Supervision

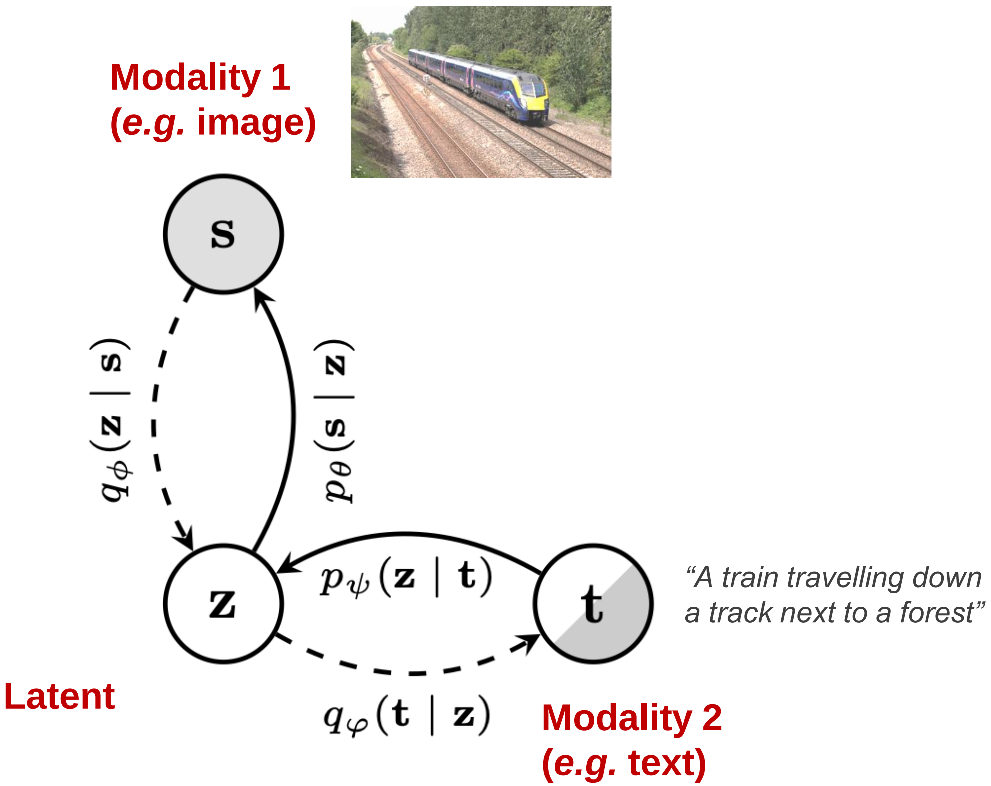
Cue in **semi-supervised VAEs**:

Generation (solid arrows):
$p(\mathbf{s},\mathbf{z},\mathbf{t}) = p(\mathbf{s}|\mathbf{z})\, p(\mathbf{z}|\mathbf{t})\, p(\mathbf{t})$

Latent representation of **s** (image) is supervised by **t** (caption)

Captioning (dashed arrows):
$p(\mathbf{t},\mathbf{z}|\mathbf{s}) = p(\mathbf{t}|\mathbf{z})\, p(\mathbf{z}|\mathbf{s})$

Latent representation of **t** (caption) is supervised by the **s** (image)



Information flows in **both directions:**
s (image) → **z** (latent)→ **t** (caption)
t (caption) → **z** (latent) → **s** (image)

Joy *et al.*, "Learning Multimodal VAEs Through Mutual Supervision", ICLR 2022

Learning Multimodal VAEs Through Mutual Supervision

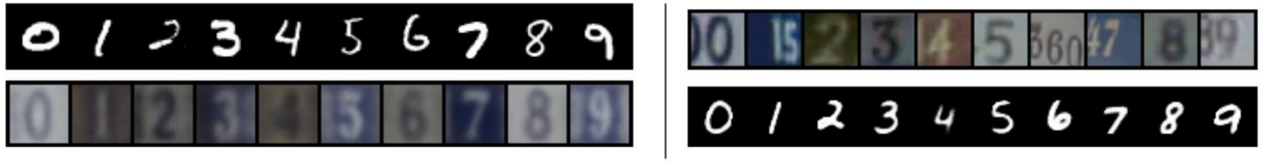
Cross-modal **generation** results

Datasets

MNIST-SVHN

Input

Output



CUB

Input

Output

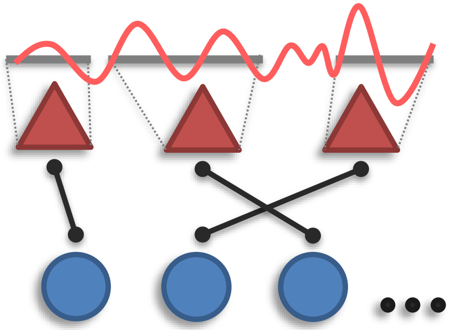
Input

Output

Joy *et al.*, "Learning Multimodal VAEs Through Mutual Supervision", ICLR 2022

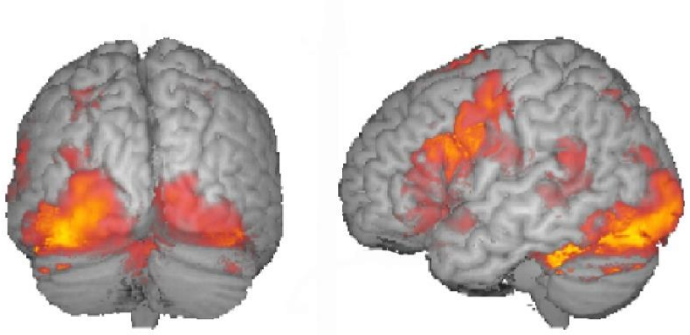
Representation Coordination

Discretization (aka Segmentation)

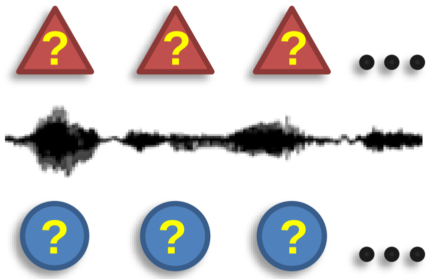


Common assumptions: ① Segmented elements

Examples:



Medical imaging

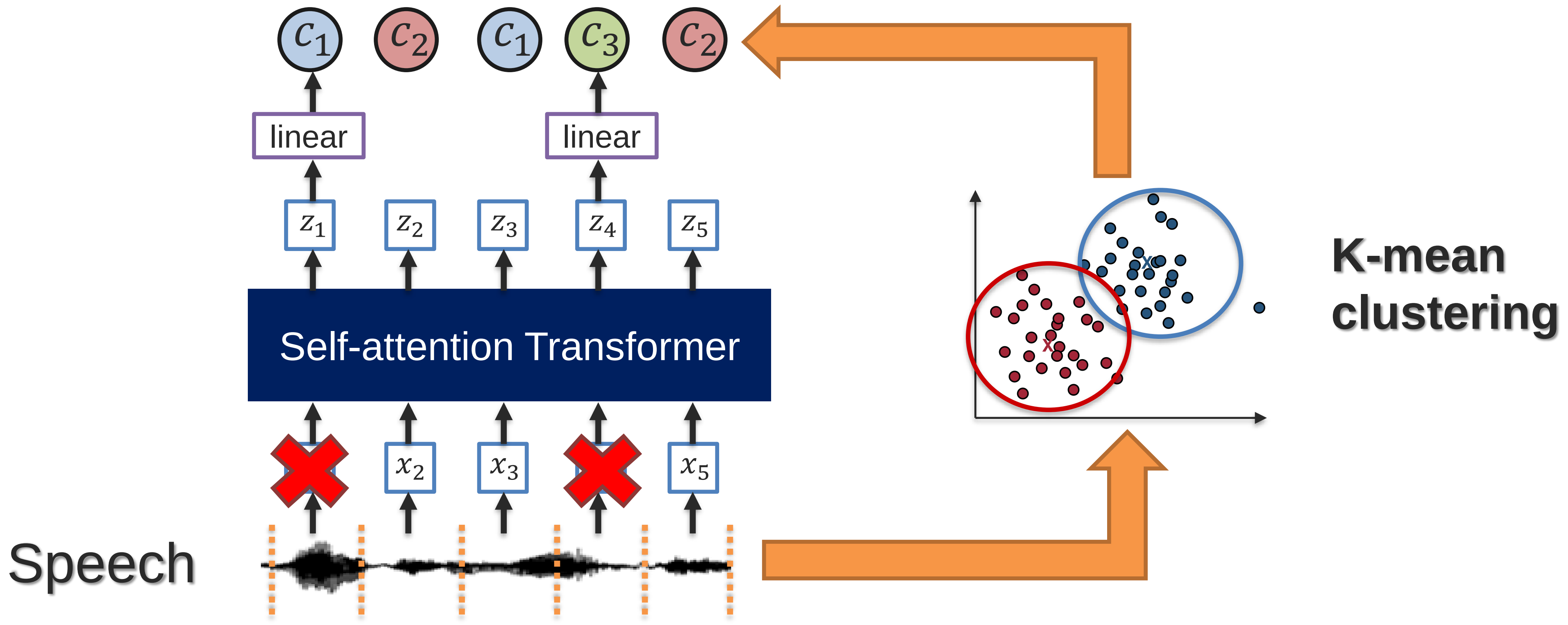


Signals

Images

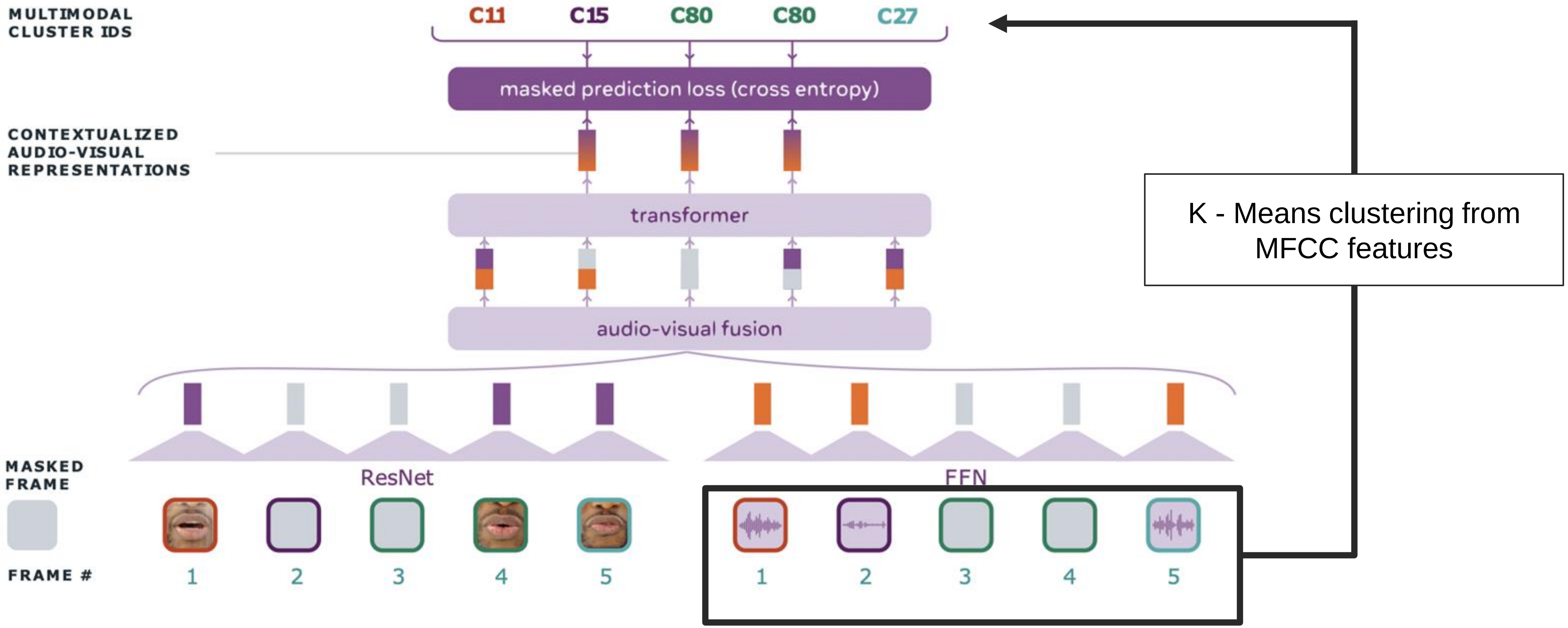
Discretization and Representation – Cluster-based Approaches

HUBERT: Hidden-Unit BERT

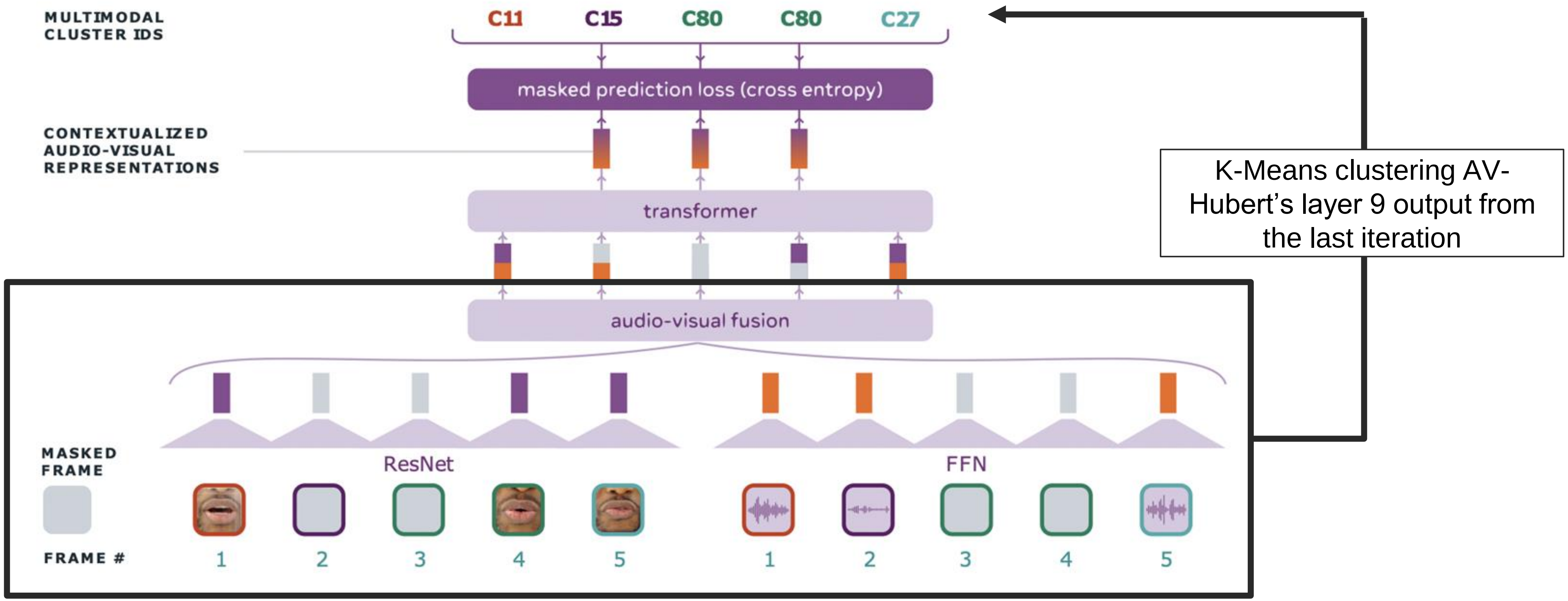


Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, arxiv 2021

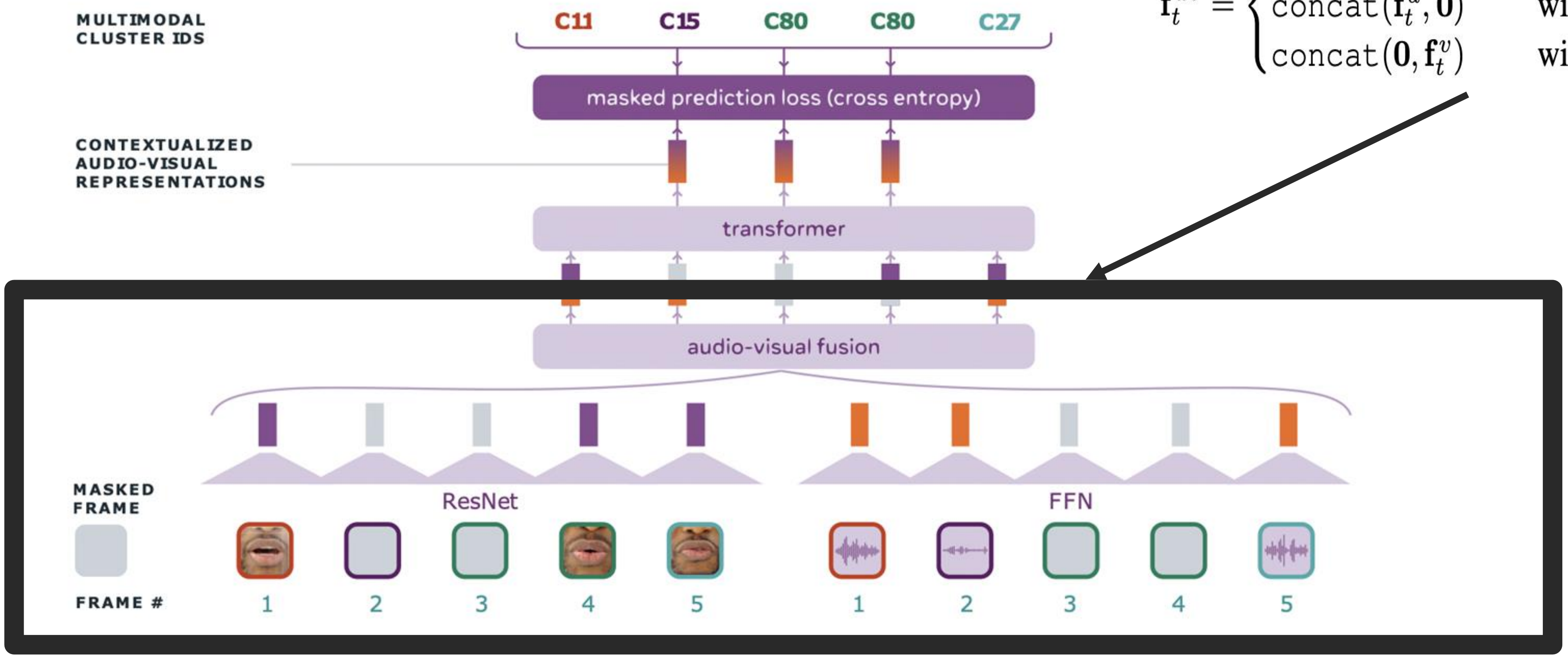
How do we get target cluster IDs ? (Iteration 1)



How do we target cluster IDs in a Multimodal way? (Iteration 2)



LEARNING AUDIO-VISUAL SPEECH REPRESENTATION BY MASKED MULTIMODAL CLUSTER PREDICTION

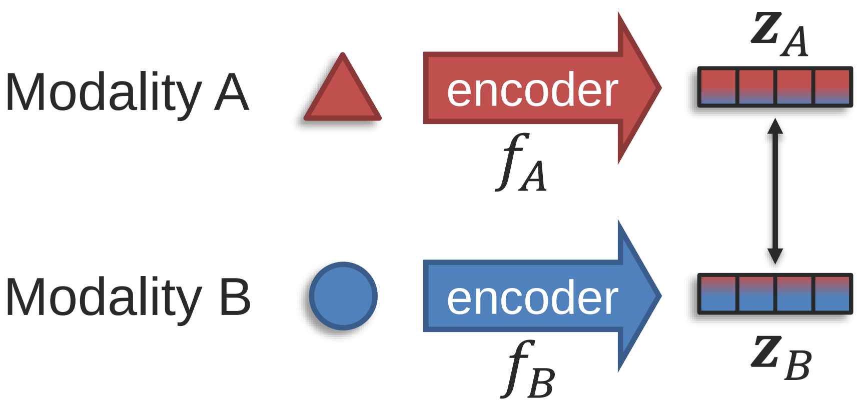


Modality dropout:

$$\mathbf{f}_t^{av} = \begin{cases} \texttt{concat}(\mathbf{f}_t^a, \mathbf{f}_t^v) & \text{with } p_m \\ \texttt{concat}(\mathbf{f}_t^a, \mathbf{0}) & \text{with } (1-p_m)p_a \\ \texttt{concat}(\mathbf{0}, \mathbf{f}_t^v) & \text{with } (1-p_m)(1-p_a) \end{cases}$$

Representation Coordination

Coordination with Contrastive Learning



Contrastive loss:

brings positive pairs closer and pushes negative pairs apart

Popular contrastive loss: InfoNCE

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

positive pairs

Similarity function can be cosine similarity

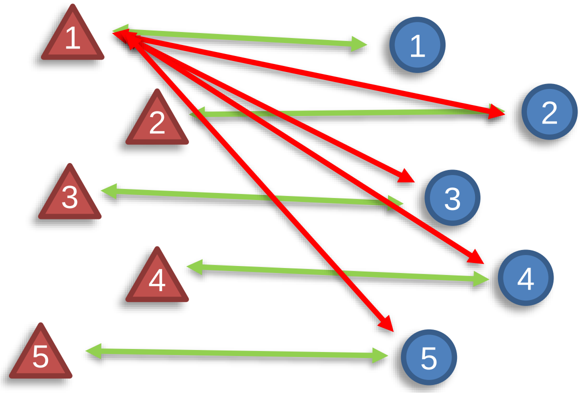
negative pairs and positive pairs

$$sim(\boldsymbol{z_A}, \boldsymbol{z_B}) = e^{(z_A \cdot z_B/\tau)}$$

temperature

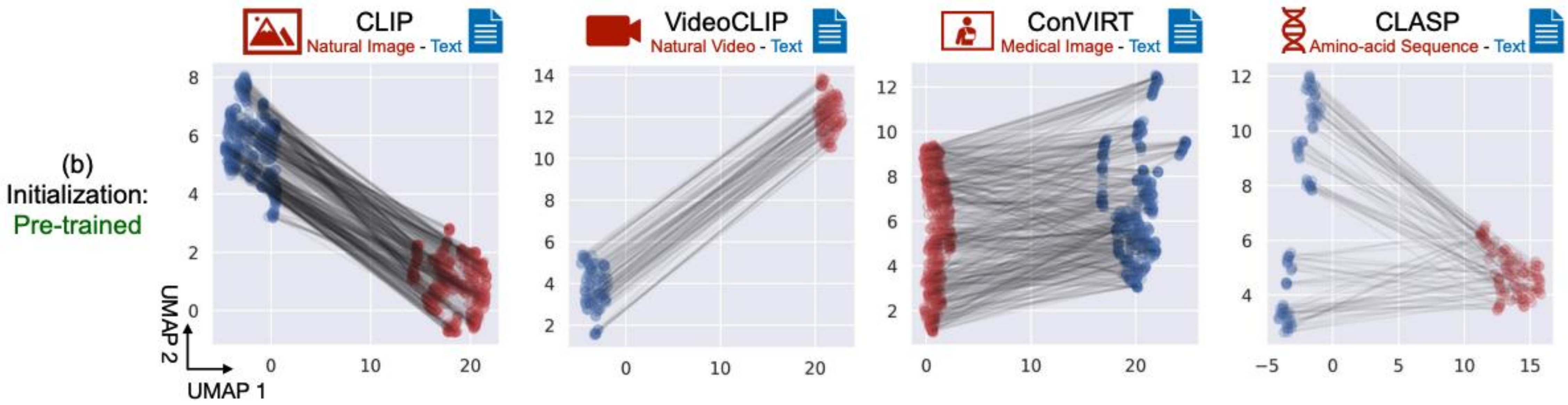
Paired data: { △ , ● }

(e.g., images and text descriptions)



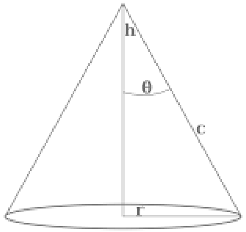
Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

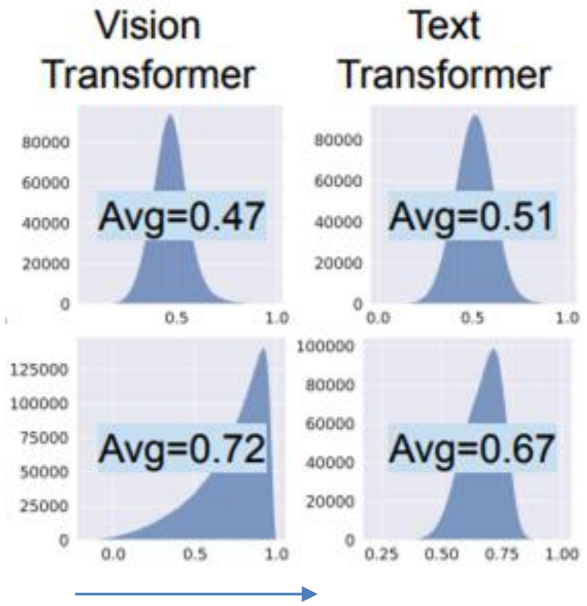
Modality Gap embeddings of different modalities are projected to completely separate regions of the embedding space



Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

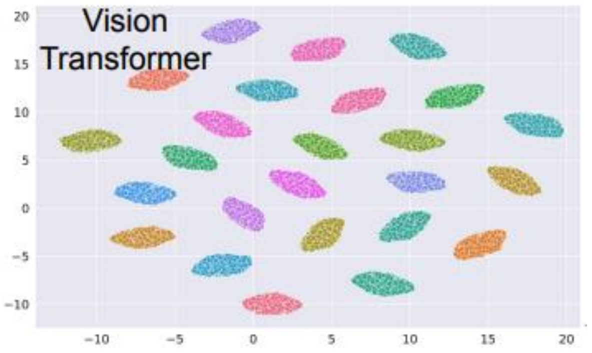
Cone effect

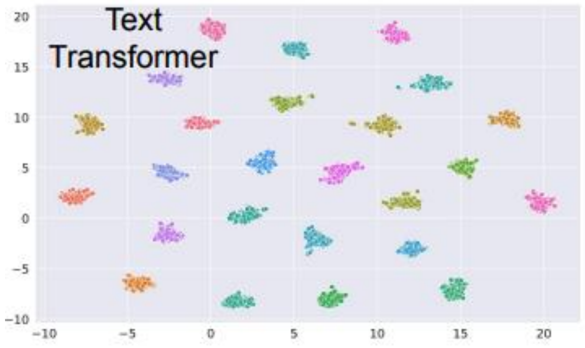




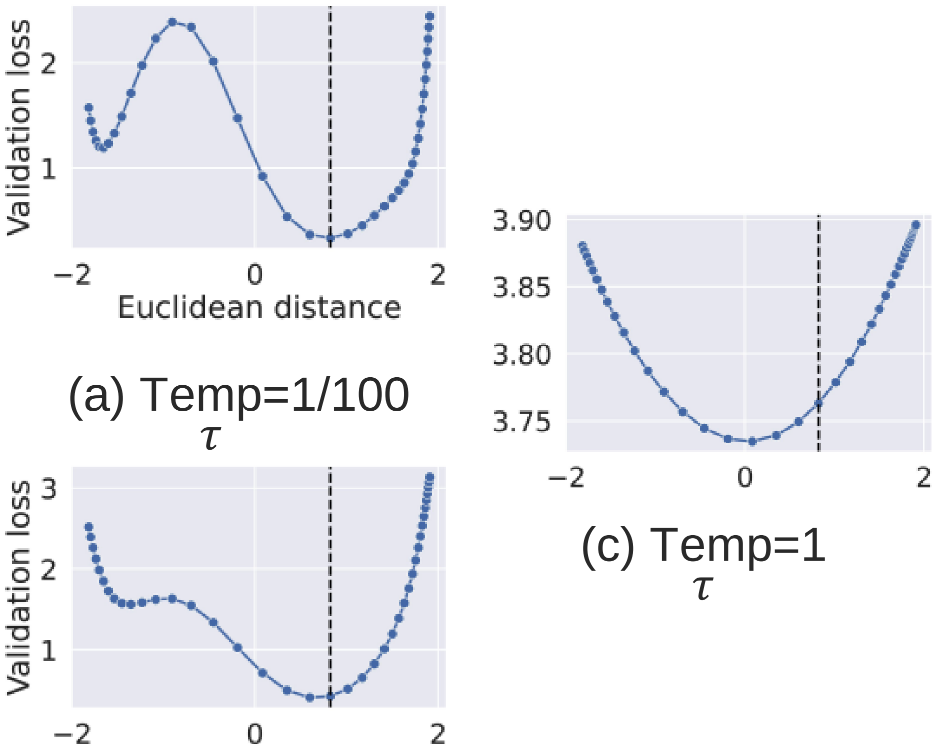
Cosine similarity between all pairs

Different random Initializations





Contrastive learning Optimizations



(a) Temp=1/100
τ

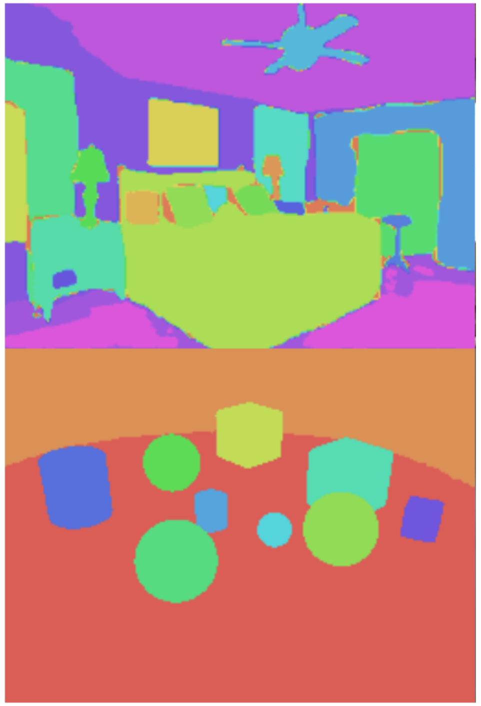
(b) Temp=1/50
τ

(c) Temp=1
τ

Generation

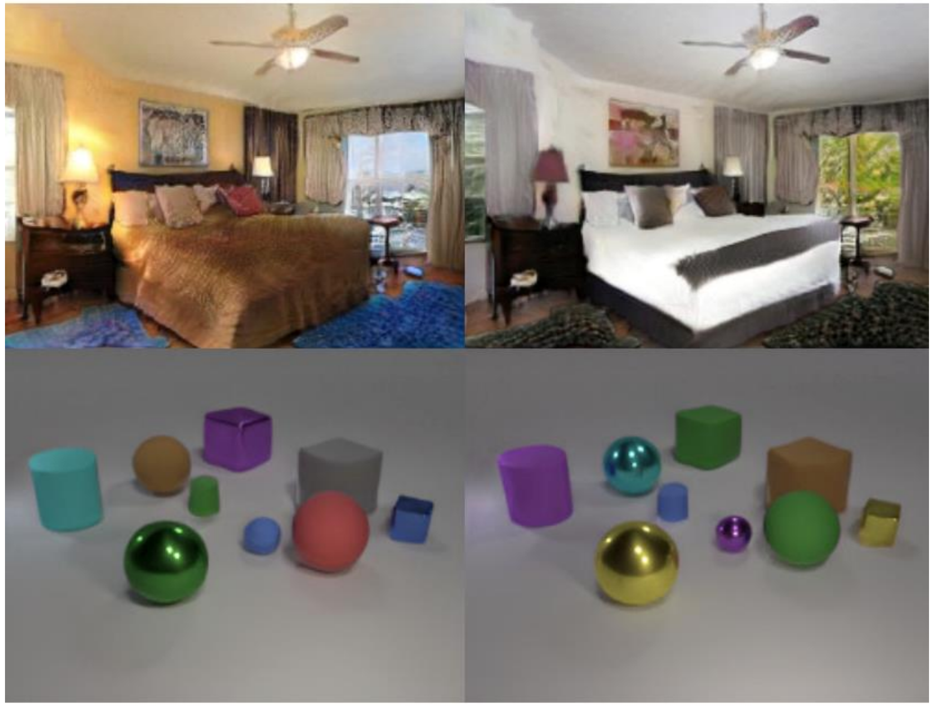
Controllability during Generation

Layout



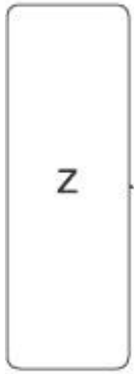
Generation

Different styles



Hudson & Zitnick, “Compositional Transformers for Scene Generation”. Neurips, 2022

Image Generation



Hudson & Zitnick, “Compositional Transformers for Scene Generation”. Neurips, 2022

Image Generation

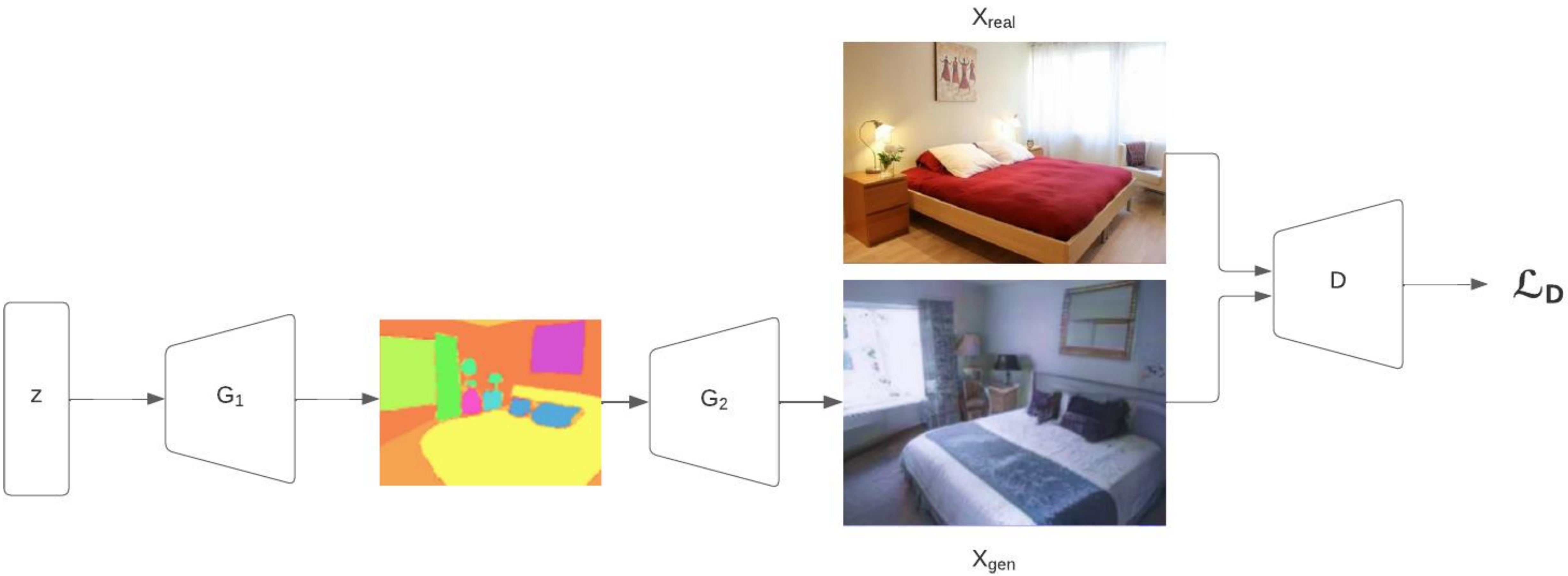


Image Generation

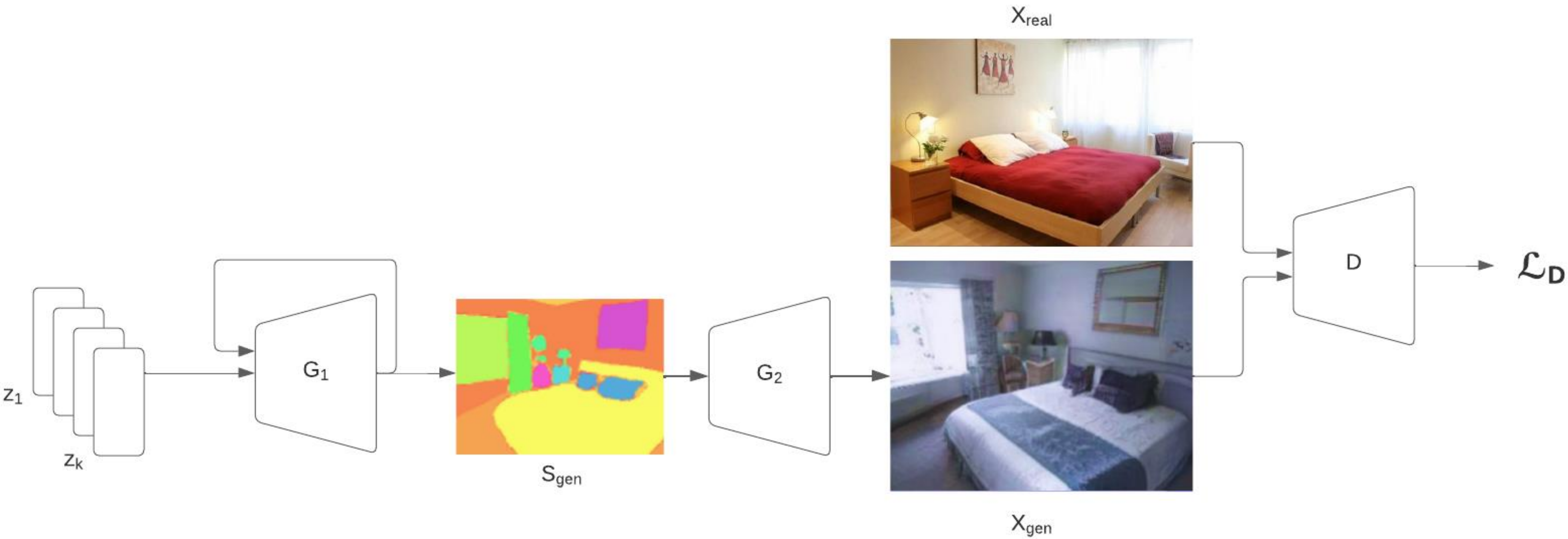


Image Generation

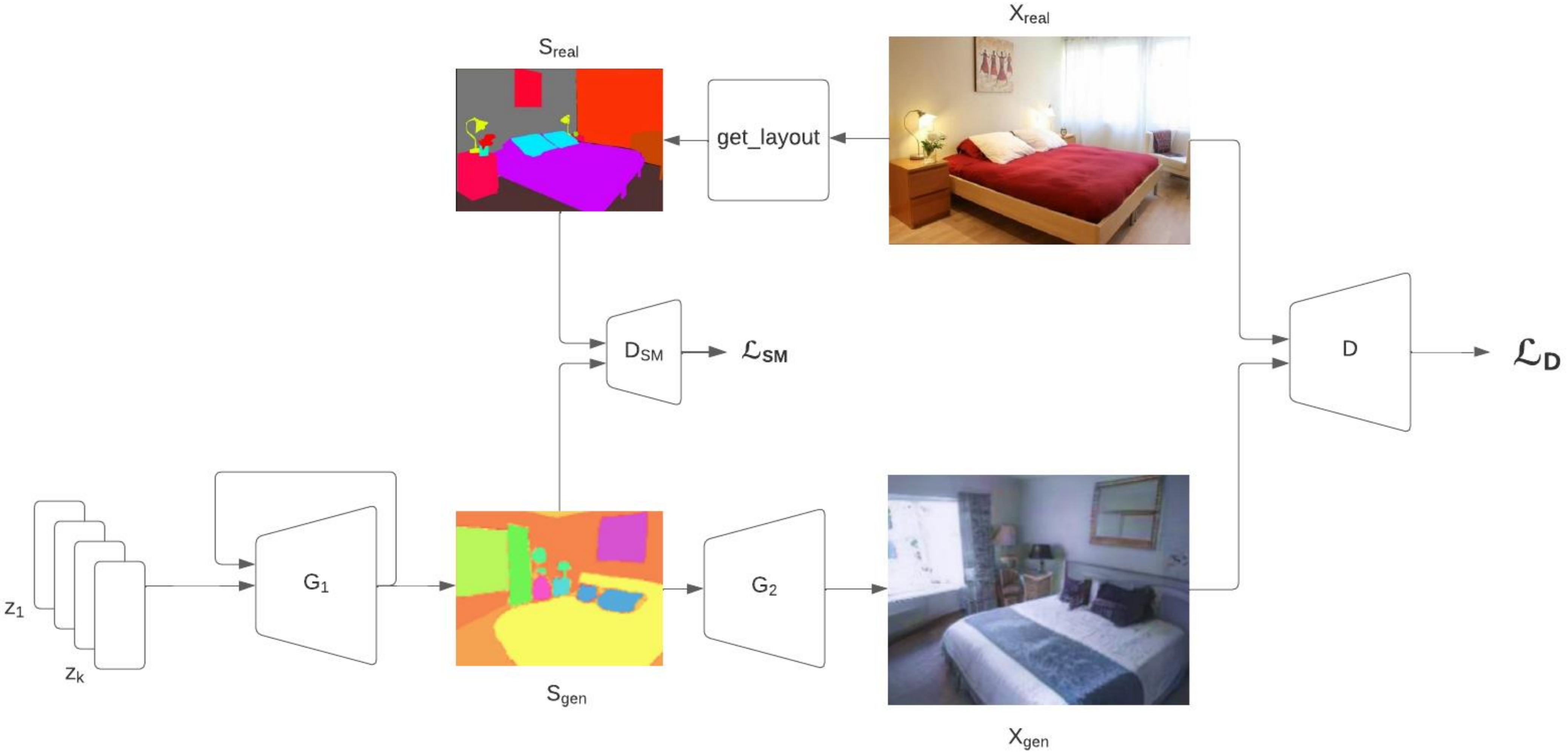
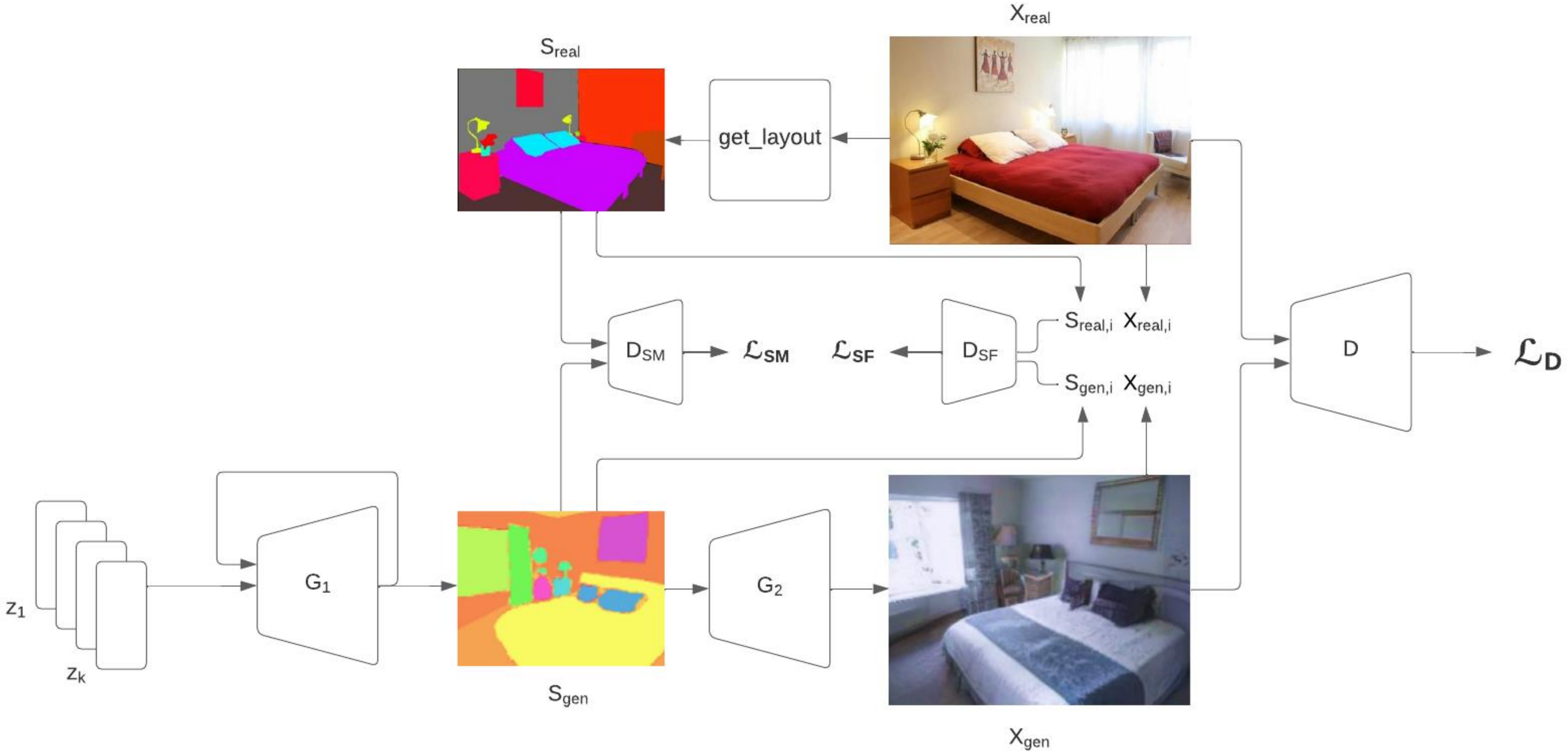
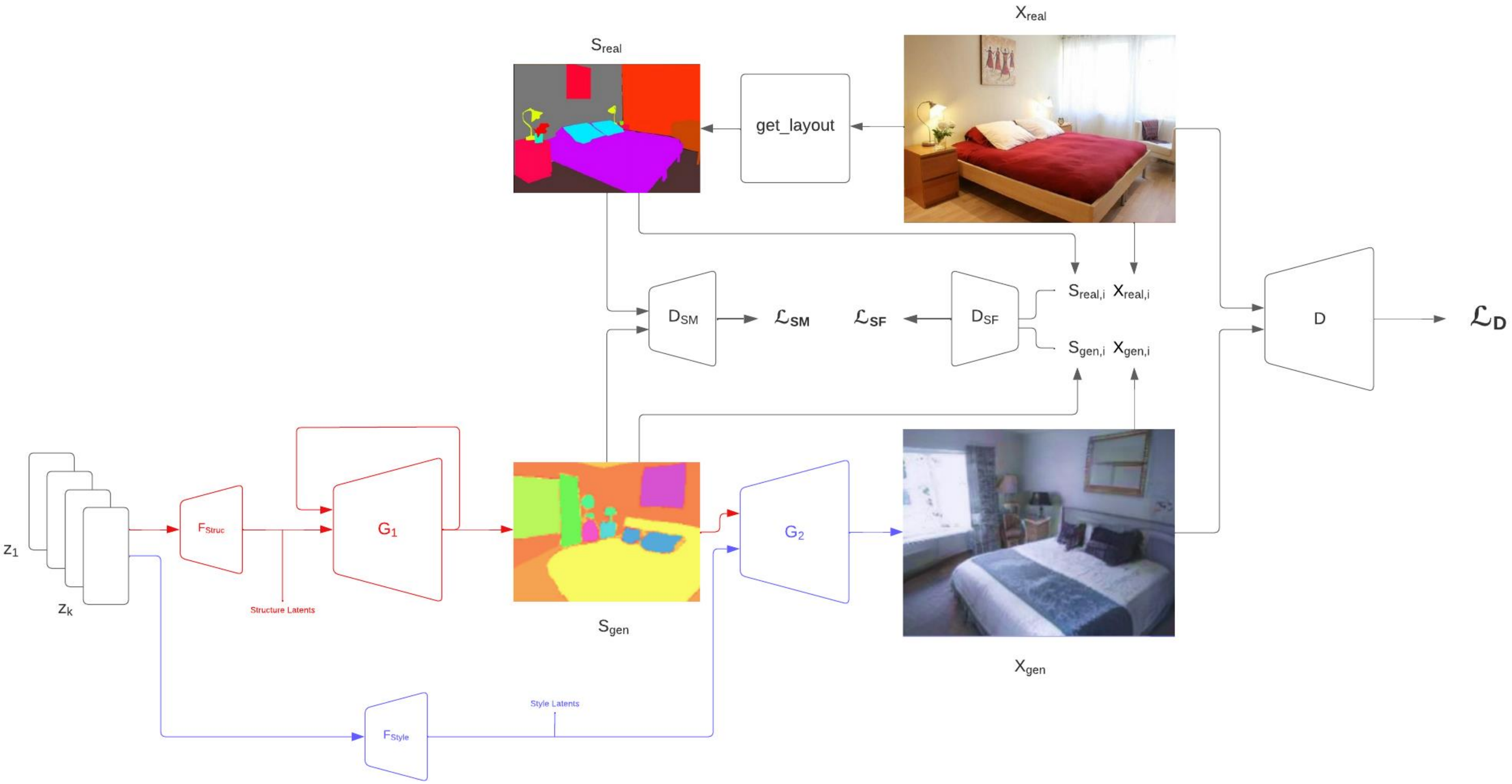


Image Generation



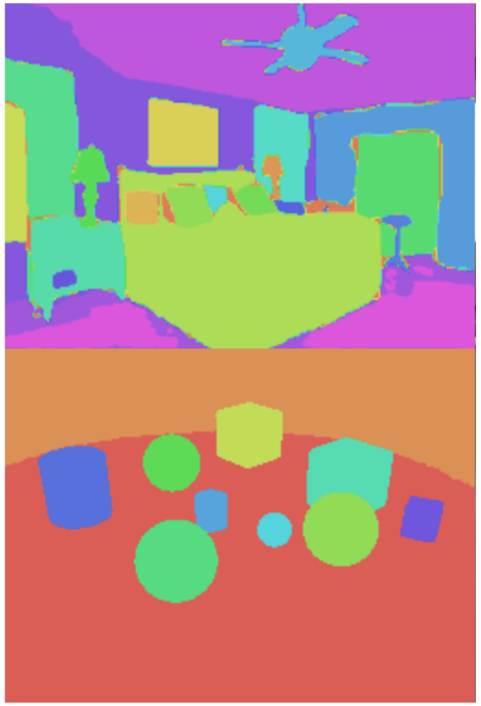
GANFormer 2.0



Hudson & Zitnick, “Compositional Transformers for Scene Generation”. Neurips, 2022

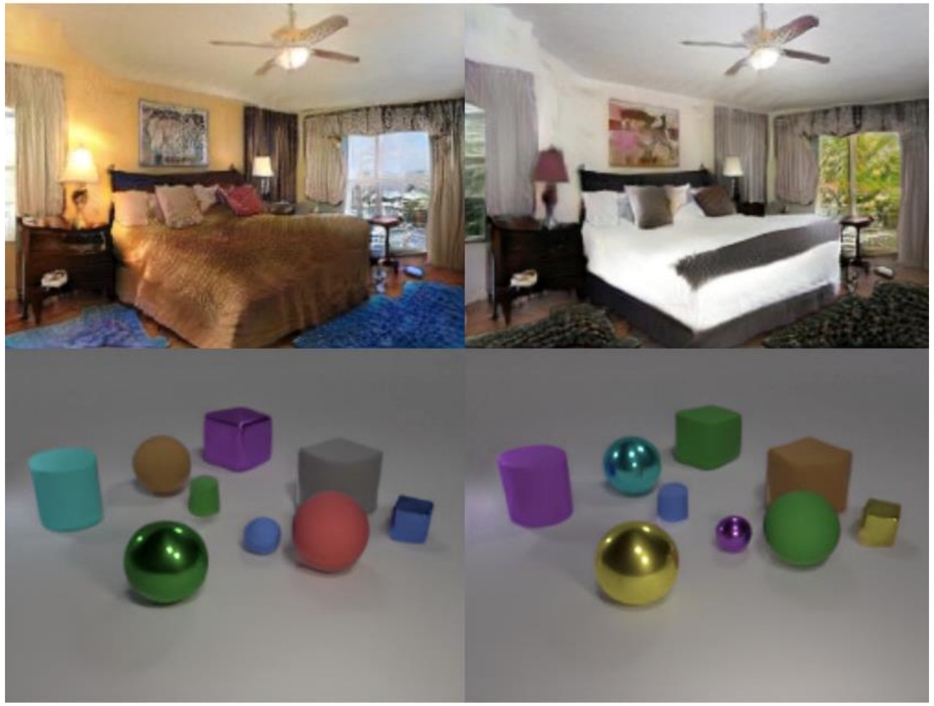
Controllability during Generation

Layout



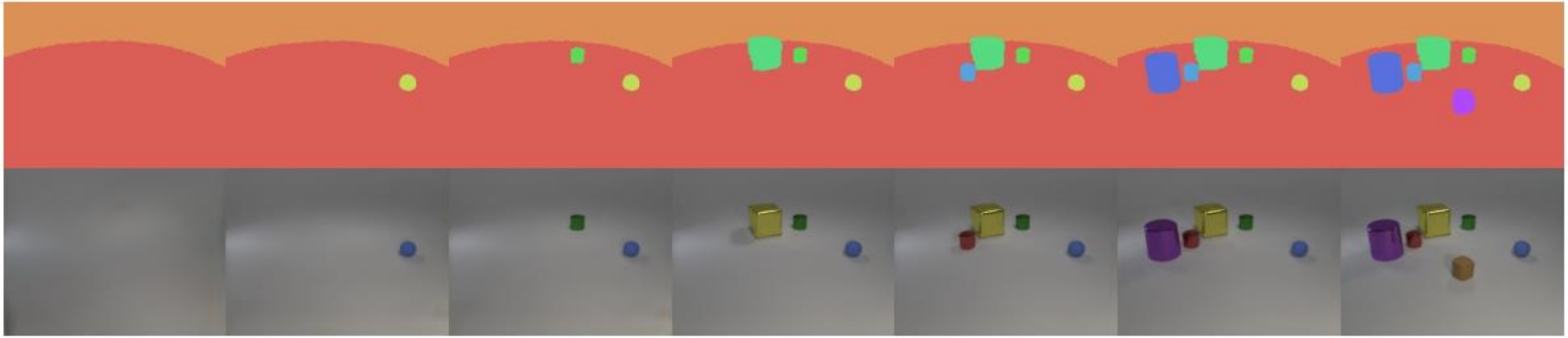
Generation

Different styles



Hudson & Zitnick, “Compositional Transformers for Scene Generation”. Neurips, 2022

Transparency and Interpretability



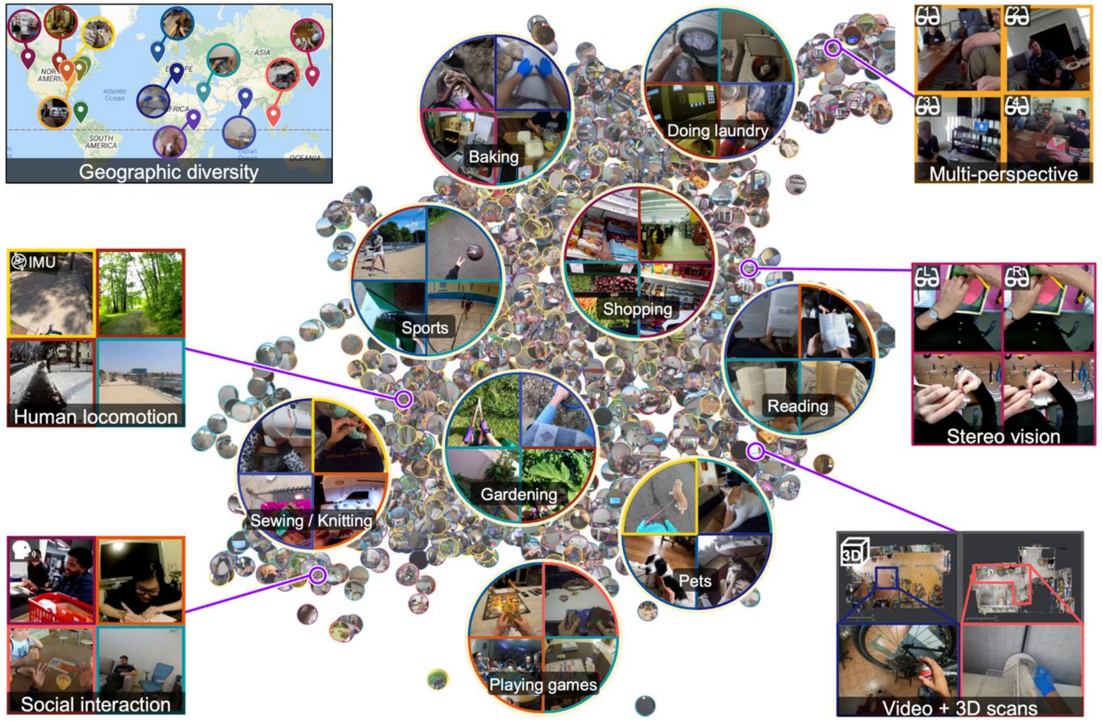
Hudson & Zitnick, “Compositional Transformers for Scene Generation”. Neurips, 2022

Multimodal Benchmarks

Ego4D: Around the World in 3,000 Hours of Egocentric Video

Ego4D: New **in-the-wild** benchmark-suite with 3,670 hours of **egocentric video**

Eye gaze	Video	3D meshes	Stereo	Audio

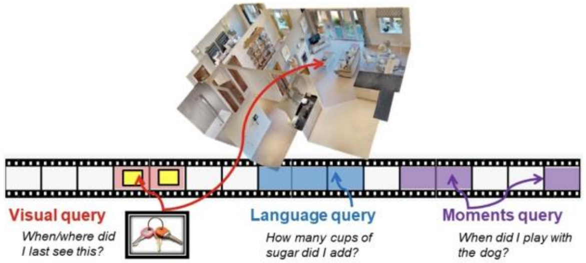


Grauman *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video", CVPR 2022

Ego4D: Around the World in 3,000 Hours of Egocentric Video

Benchmarks

Input	Benchmark
Video + Query →	**Episodic Memory**
Video →	**Hands and Objects** (identify object state-changes)
Audio + video →	**Audio-visual diarization** (Transcription, speaker detection, tracking)
Audio + video →	**Forecasting** (Locomotion, hand-movement)



State-change: Plant removed from ground

Grauman *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video", CVPR 2022

Multimodal Benchmarks

Learning to Explain: Multimodal Reasoning via Thought Chain for Science Question Answering

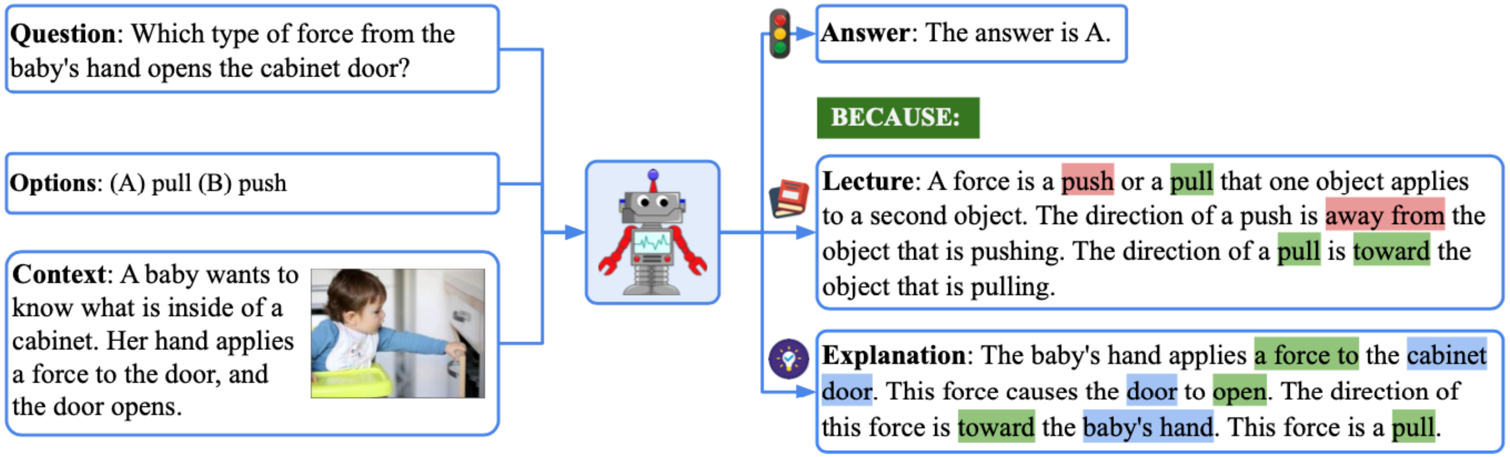
QA process

Human: utilize the information available across different modalities to synthesize a consistent and complete *chain of thought* (CoT)

DL models: black box like large-scale language models

ScienceQA

A new benchmark



Pan Lu et al., "Learning to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering" NeurIPS 2022

Learning to Explain: Multimodal Reasoning via Thought Chain for Science Question Answering

Language Models with Chain of Thought

A *chain of thought* refers to coherent flow of sentences that reveals the premises and conclusion of a reasoning problem.

Question: `question` : I_i^{ques}
Options: (A) `option` : I_{i1}^{opt} (B) `option` : I_{i2}^{opt} (C) `option` : I_{i3}^{opt}
Context: `context` : I_i^{cont}
Answer: The answer is `answer` : I_i^{a}. BECAUSE: `lecture` : I_i^{lect} `explanation` : I_i^{exp}

Question: `question` : I_t^{ques}
Options: (A) `option` : I_{t1}^{opt} (B) `option` : I_{t2}^{opt} (C) `option` : I_{t3}^{opt} (D) `option` : I_{t4}^{opt}
Context: `context` : I_t^{cont}
Answer:

Pan Lu et al., "Learning to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering" NeurIPS 2022

Learning to Explain: Multimodal Reasoning via Thought Chain for Science Question Answering

Predicted Answer along with the Chain-of-Thought from GPT-3

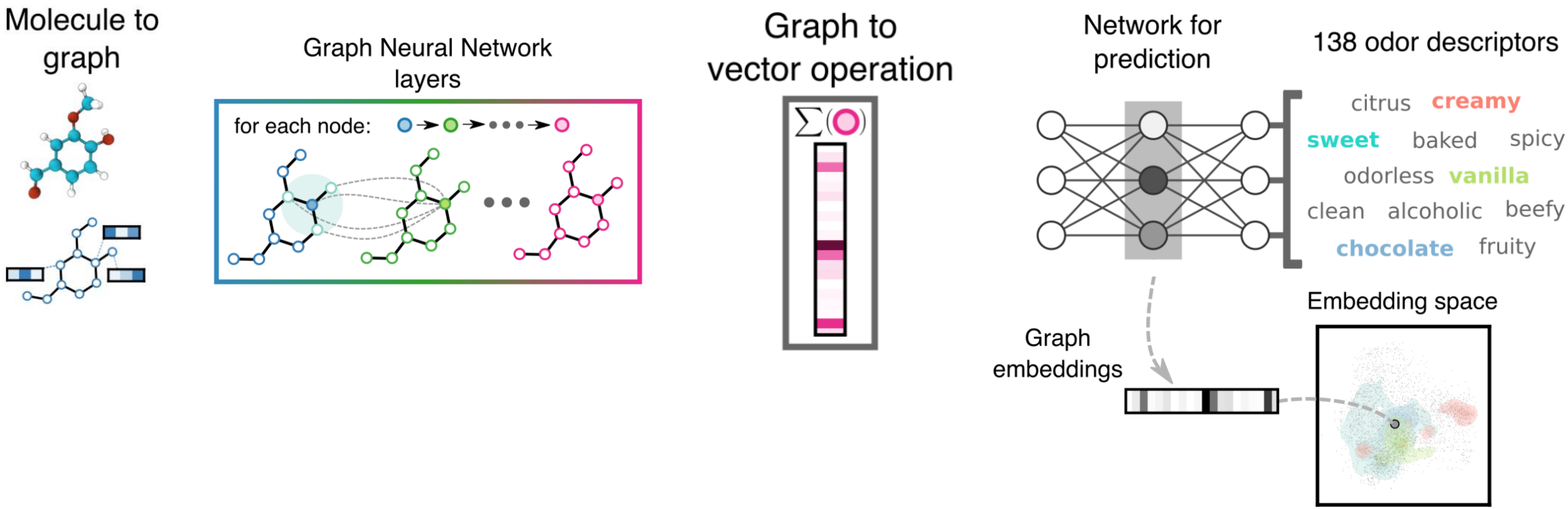
Question: What do these two changes have in common?\n
bleaching clothes\n
a piece of apple turning brown\n
Options: (A) Both are only physical changes. (B) Both are caused by cooling. (C) Both are chemical changes. (D) Both are caused by heating.\n
Answer: **The answer is C. BECAUSE:** A **physical change** is a change to a substance that does not change what the substance is. A **physical change** only changes the substance's physical properties, such as its size, shape, or color.\n
A **chemical change** is a change to a substance that changes what the substance is. A **chemical change** changes the substance's chemical properties, such as its chemical structure.\n
The changes that you listed are both chemical changes.\n
When you **bleach clothes**, the chemical structure of the fabric changes. This change makes the fabric less likely to hold color.\n
When **a piece of fruit turns brown**, the chemical structure of the fruit changes. This change makes the fruit taste different.

Pan Lu et al., "Learning to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering" NeurIPS 2022

Digitizing Smell

The Principle Odor Map

Learning an Embedding Space for Scent



Sanchez-Lengeling, Benjamin, et al. "Machine learning for scent: Learning generalizable perceptual representations of small molecules." arXiv 2019

Reasoning

Cross-modal Map Learning for Vision and Language Navigation

Common SOTA approach for Vision and Language navigation:

Natural language instruction
"Turn left after the couch and head towards the kitchen"

End-to-end pipeline

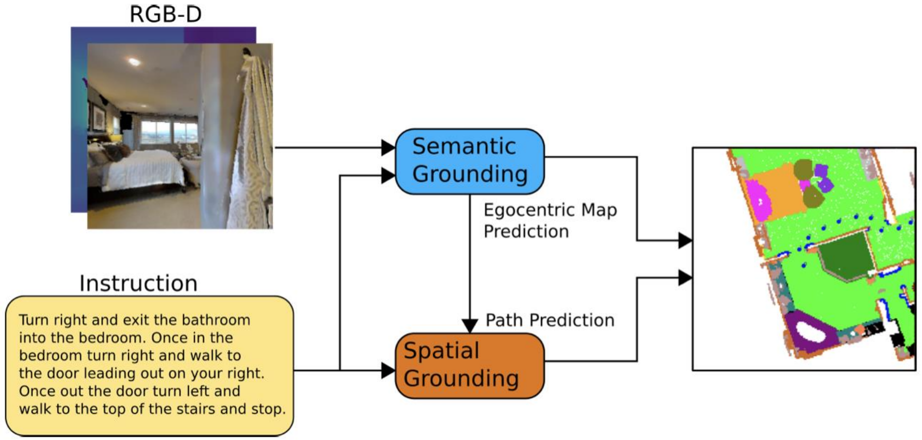
Actions
walk_forward > turn_left > ...

Can we expect this module to learn **mapping**, **planning** and **control**?

Proposed cross-modal map learning:
Two multimodal soft-dot attention modules

Instruction + RGB-D: **predict the egocentric semantic map** (out-of-view regions)

Language + predicted egocentric map: **localize the next waypoint**



Georgakis *et al.*, "Cross-modal Map Learning for Vision and Language Navigation", CVPR 2022