Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 9.1: Generation 1 – Generation and Generative Models

Paul Liang

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

Administrative Stuff

Midterm Project Report (Due Monday 10/31 at 8pm)

Main goals:

1. Experiment with state-of-the-art approaches
   - Run on your own dataset state-of-the-art models
     - Teams of 3 or 4 students: 2 state-of-the-art models
     - Teams of 5 or 6 students: 3 state-of-the-art models
2. Perform a detailed error analysis
   - Visualize the errors made by the state-of-the-art models
   - Discuss how you could address these issues
3. Update your research ideas
   - You should have N-1 research ideas (N=number of teammates)
   - Your ideas should center around multimodal challenges
     - At most 1 idea can be unimodal in nature

Midterm Project Report (Due Monday 10/31 at 8pm)

Some suggestions:

- You do not need to re-implement state-of-the-art models
  - But you need to rerun them yourself on your own data
- You may want to fine-tune your baseline models on your data
- If your dataset is too large:
  - You can use a subset of your data.
  - But be consistent between experiments
- The most important part is the discussion
  - How is your error analysis affecting your proposed research ideas?

Midterm Project Presentations (Tuesday 11/1 and Thursday 11/3)

Main objective:

- Present your research ideas and get feedback from classmates

Presentation length:

- Teams with 3 students: 4 minutes
- Teams with 4 students: 5 minutes
- Teams with 5 students: 6 minutes
- Teams with 6 students: 7 minutes

- Following each presentation, audience will be asked to share feedback

Midterm Project Presentations (Tuesday 11/1 and Thursday 11/3)

- Administrative guidelines
  - All presentations will be done from the same laptop
    - Google Drive directory will be shared to host your presentation
    - Preferred option: Google Slides
    - Second option: Microsoft Powerpoint
  - Be sure to be on time! We have many presentations each day ☺
  - All presentations are in person (no remote presentations
    - The schedule will be shared soon
      - Half the teams on Tuesday and second half on Thursday
      - We will use the opposite order for the final presentations
    - Audience students should plan to be in person
      - Because of room capacity constrained, a few students will be asked to be remote

Midterm Project Presentations (Tuesday 11/1 and Thursday 11/3)

- Some suggestions:
  - Do not present your results from state-of-the-art baseline models
    - Only exception: if the result directly justifies one of your research ideas
  - The focus of your presentation should be about your research ideas
    - Plan about 1 minute for each research idea
    - Present the ideas at the high-level, so that audience understands it
  - Only 1 minute (or less) for the intro (dataset, task)
  - All teammates should be included in the presentation
  - Be as visual as possible in your slides

Midterm Project Presentations (Tuesday 11/1 and Thursday 11/3)

- Grading guidelines for presentations (4 points)
  - Quality of the slides (incl. images, videos and clear explanations)
  - Good motivation and explanation of the problem
  - Future research ideas (describe their future research directions)
  - Presentations skills (incl. explanations, voice and body posture)

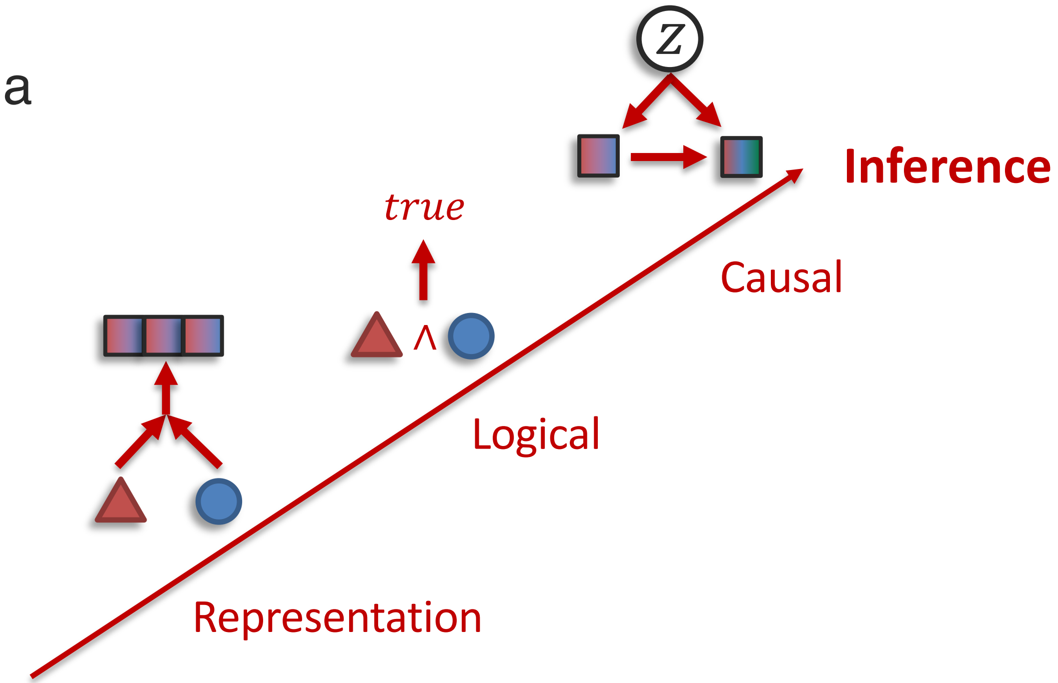
- Grade will also be given for audience feedback (1 point)
  - You should plan to give feedback for at least 6 teams
  - Try to be constructive in your feedback
  - Sharing pointer to relevant papers is quite helpful

Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Towards explicit inference paradigms:

1. Logical inference
2. Causal inference: how can one determine the actual **causal** effect of a variable in a larger system?



Causal Inference

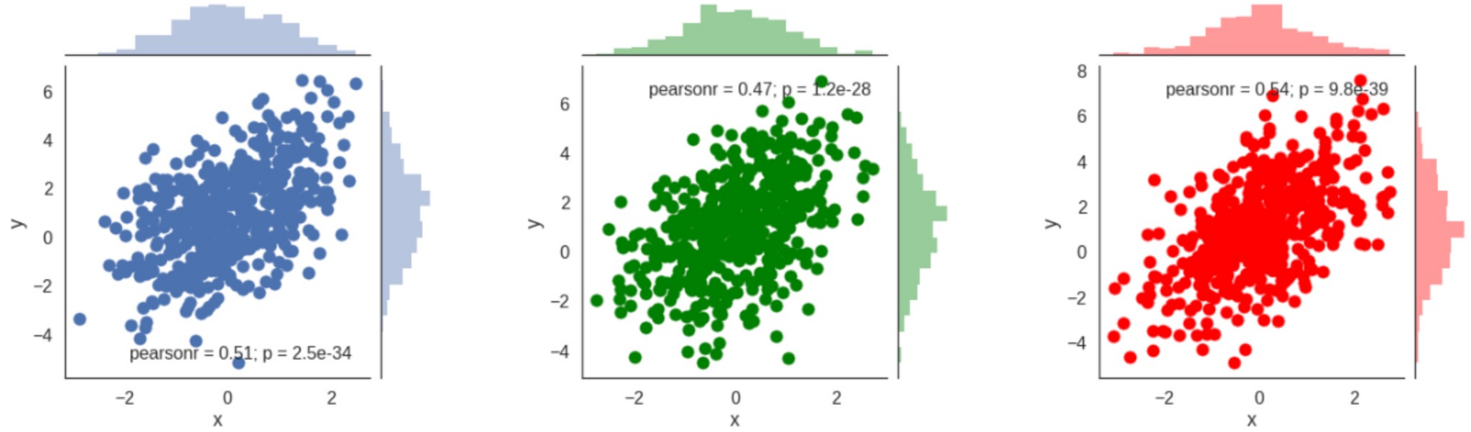
Intervention

Causal inference is reliant on the idea of interventions —what outcome might have occurred if X happened (an intervention), possibly contrary to observed data.

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

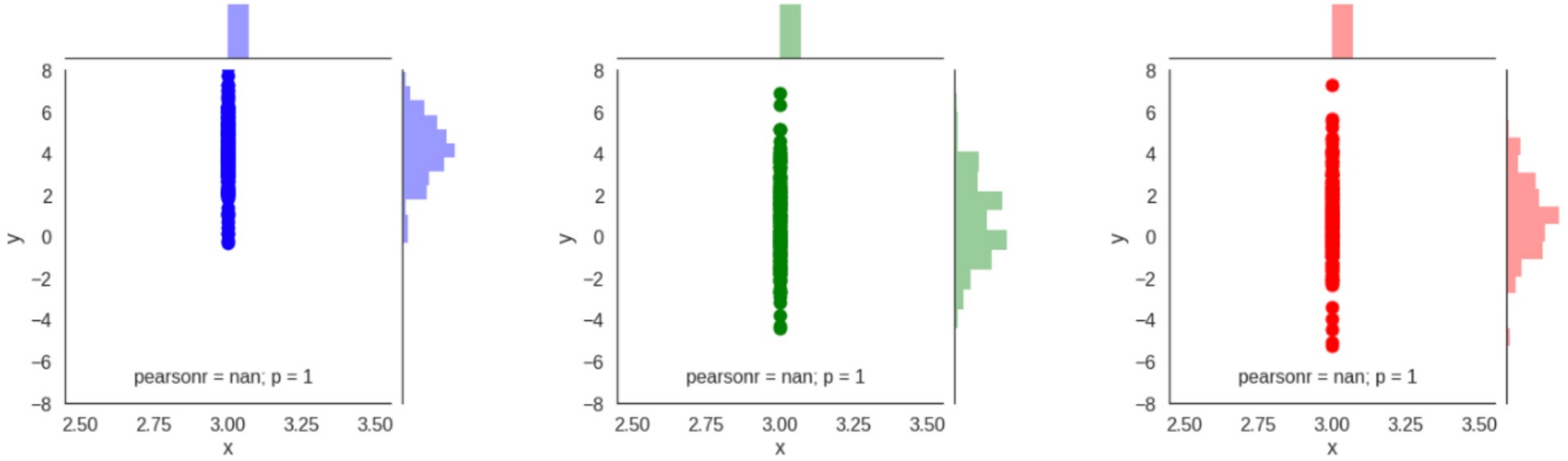
Intervention

Let’s say I really want to set the value of *x* to 3. What happens to *y*?

```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```

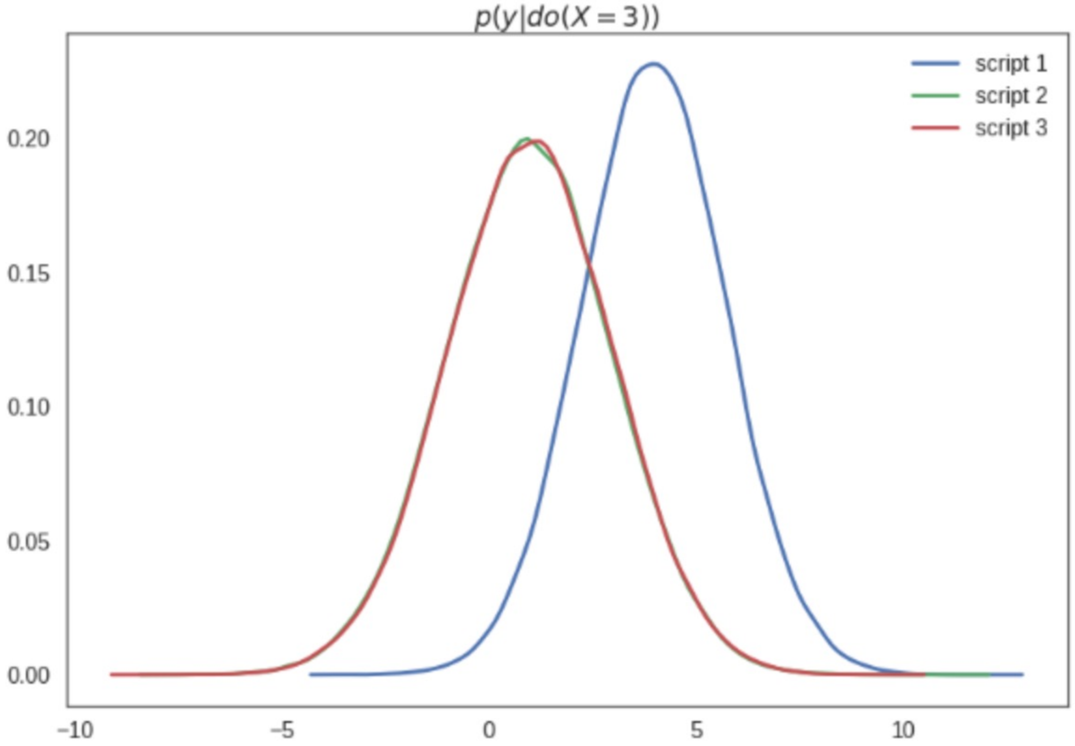


[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

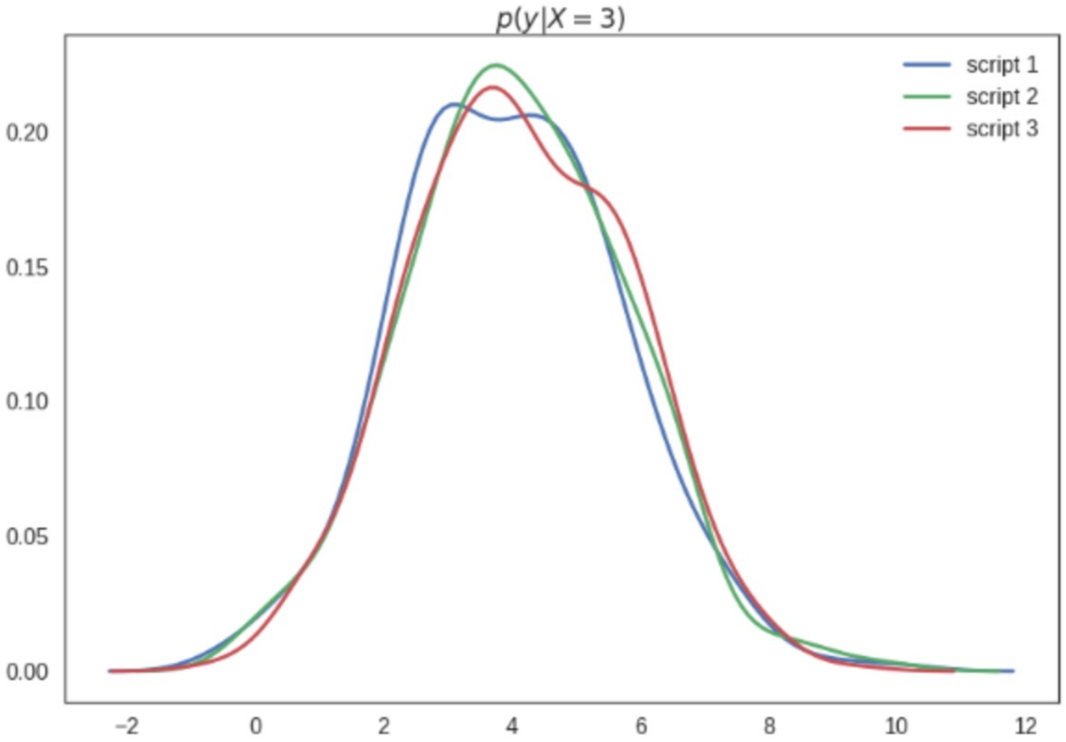
Causal Inference

Intervention

The marginal distribution of *y*: p(y | do(x=3)).



The marginal distribution of *y*: p(y | x=3).



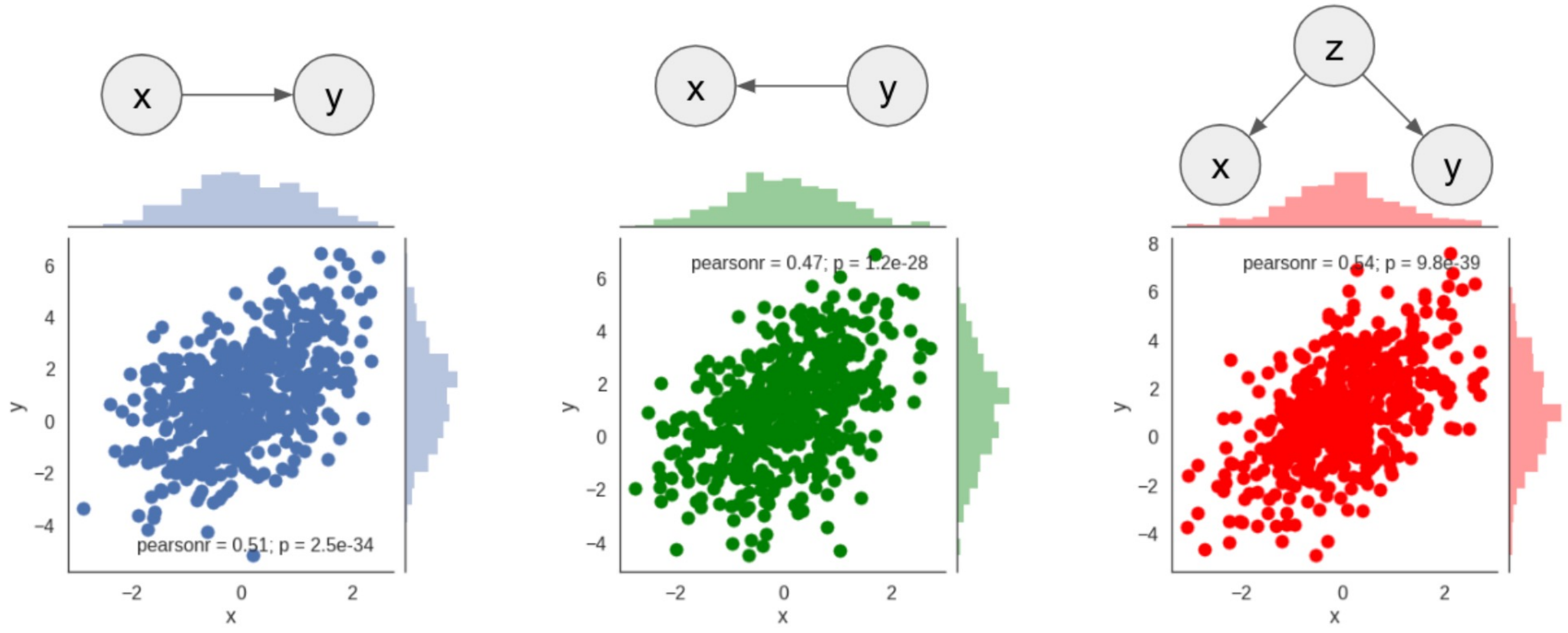
The joint distribution of data alone is insufficient to predict behavior under interventions.

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

Causal diagrams: arrow pointing from cause to effect.

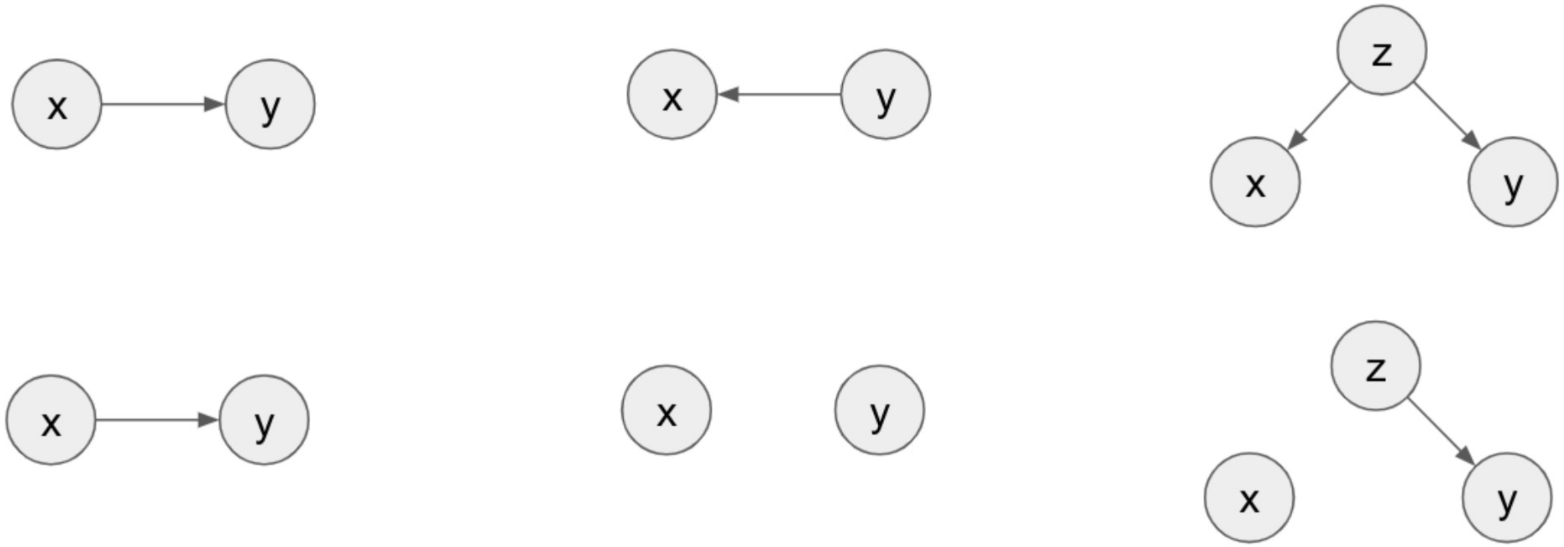
x = randn() y = x + 1 + sqrt(3)*randn()	y = 1 + 2*randn() x = (y-1)/4 + sqrt(3)*randn()/2	z = randn() y = z + 1 + sqrt(3)*randn() x = z



[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

Intervention mutilates the graph by removing all edges that point into the variable on which intervention is applied (in this case *x*).



$P(y|do(X)) = p(y|x)$ $\qquad$ $P(y|do(X)) = p(y)$ $\qquad$ $P(y|do(X)) = p(y)$

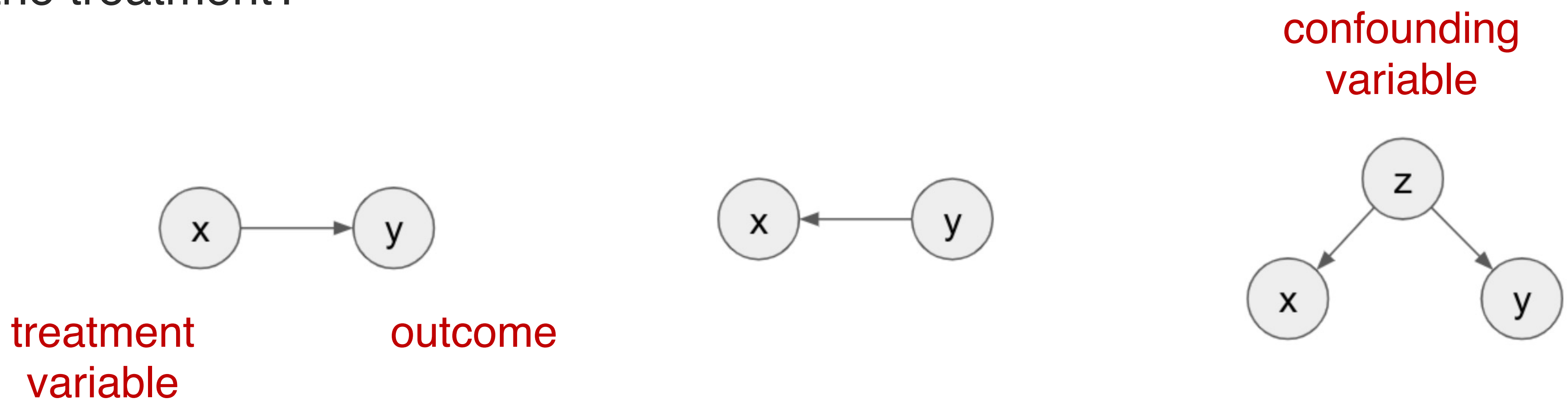
[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

Intervention in real-life is typically very hard!

E.g., does treatment x treat disease y?

Can I estimate the intervention p(y|do(X=x))?
Requires answering: all else being equal, what would be the patient’s outcome if they had not taken the treatment?



Lots of work, see Judea Pearl, The Book of Why

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

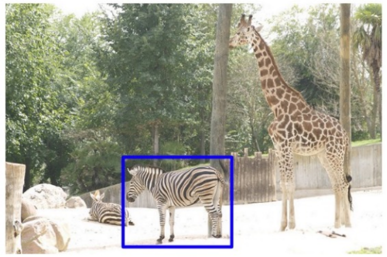
Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

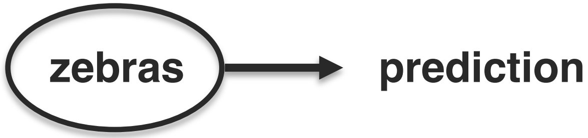
Target object in question

Q: How many zebras are there in the picture?
A: 2



Baselines: **2**

i.e., treatment variable



BUT: correlation or causation?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Recall error analysis!

Causal Inference

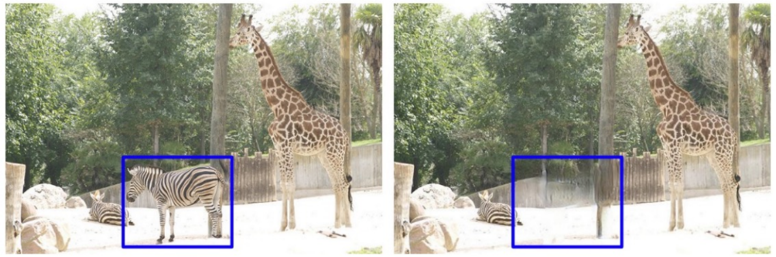
Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

Target object in question

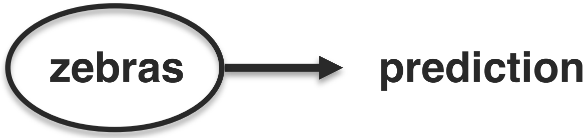
Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1



Baselines: 2 2

i.e., treatment variable



Interventional conditional: $p(y|do(zebras = 1))$

Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

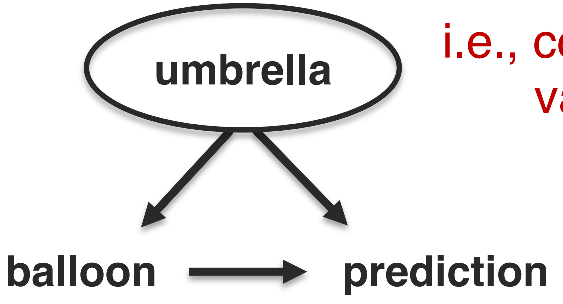
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red

Baselines: **pink**



i.e., confounding variable

Is my model picking up irrelevant objects?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Recall error analysis!

Causal VQA: does my multimodal model capture causation or correlation?

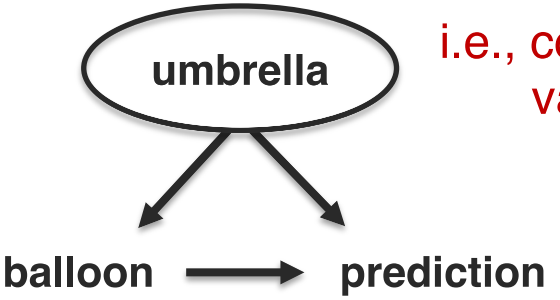
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red *umbrellas removed*; A: red

Baselines: **pink** **red**



i.e., confounding variable

Interventional conditional: $p(y|do(no\ umbrella))$

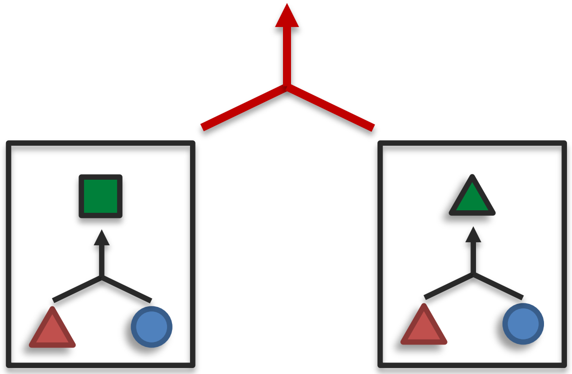
Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

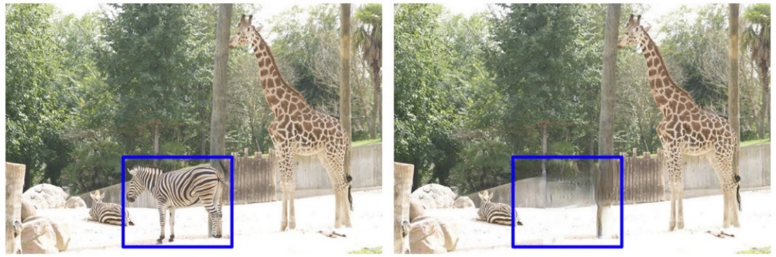
Causal Inference

Causal inference via data augmentation

Covariance
(targeted changes to answer)

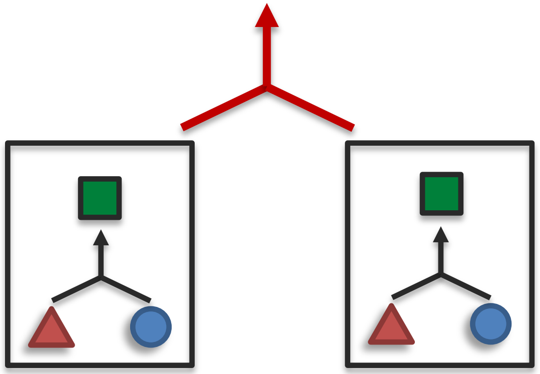


Q: How many zebras are there in the picture?
A: 2 *zebra removed* A: 1



With **relevant** object | Without **relevant** object

Invariance
(answer stays the same)



Q: What color is the balloon?
A: red *umbrellas removed*; A: red

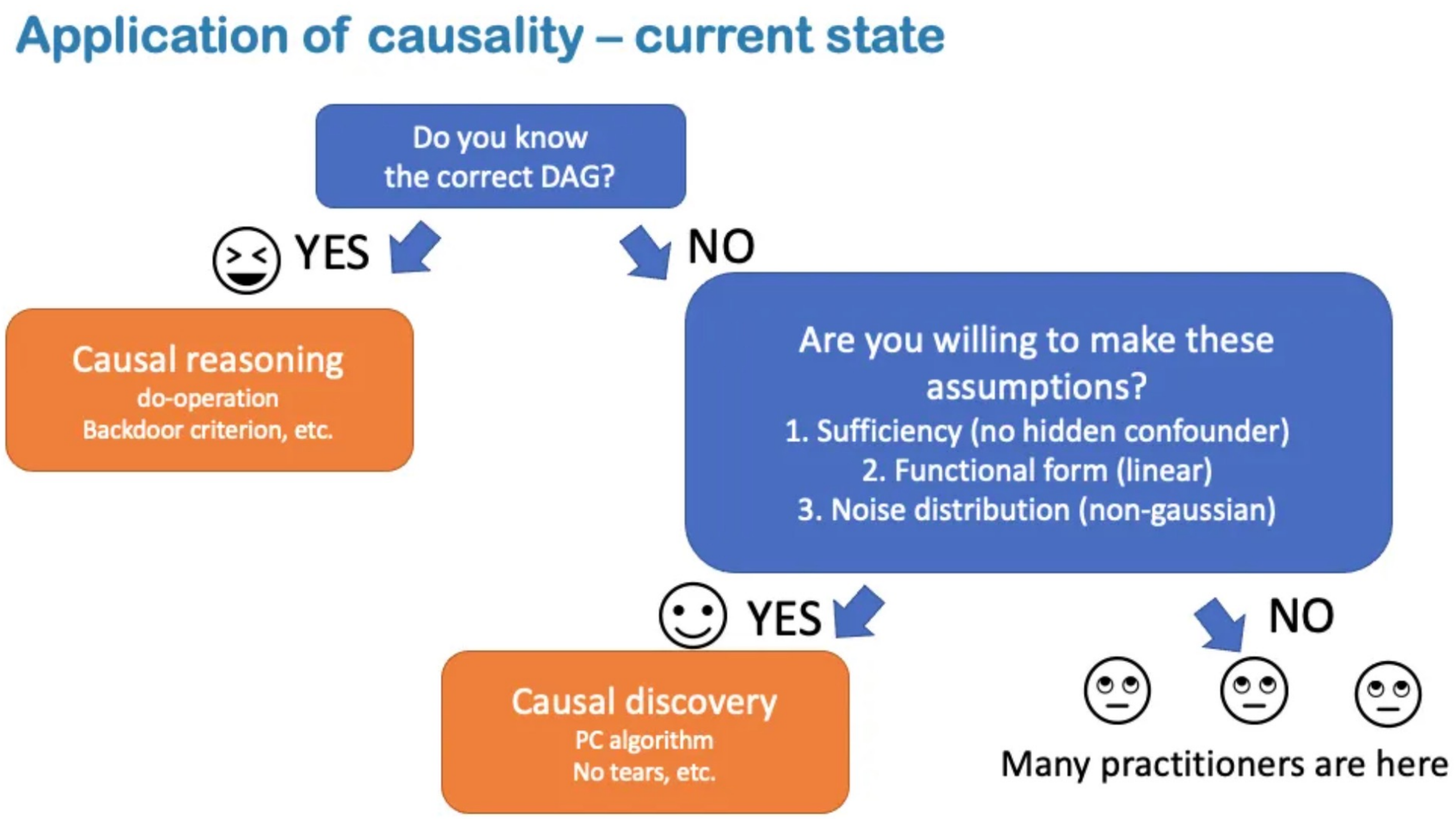
With **irrelevant** object | Without **irrelevant** object

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference Challenges

Open challenges

Many open directions



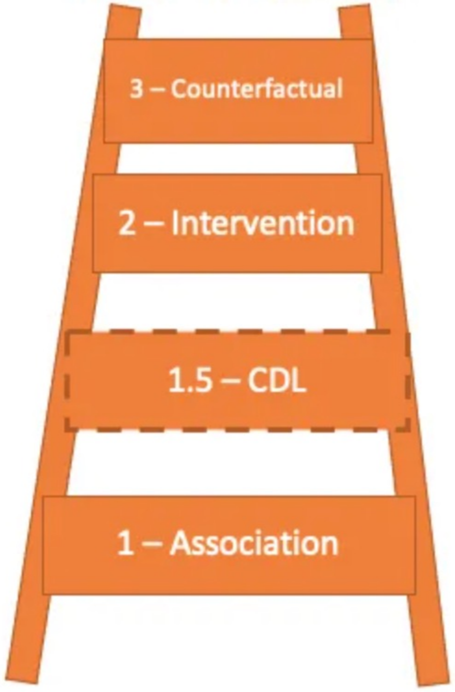
Causal deep learning, see https://www.vanderschaar-lab.com/causal-deep-learning/

Causal Inference Challenges

Open challenges

Many open directions

Ladder of causation



The space between association and intervention

Many interesting ML problems lie in Rung 1.5

- Robustness
  - Distribution shift
  - Adversarial attack
- Generalization
  - Domain adaptation
  - Transfer learning
  - Meta-learning
  - Few-shot learning
- Other potential areas
  - Fairness
  - Data augmentation
  - Etc.

1. Empirically verifiable
2. "Good enough"

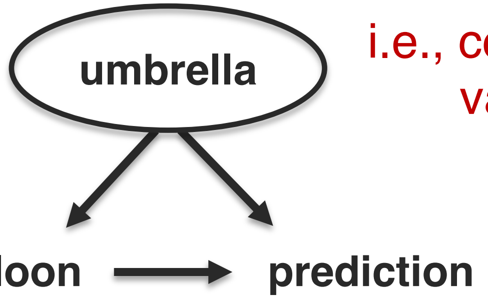
Causal deep learning, see https://www.vanderschaar-lab.com/causal-deep-learning/

Sub-Challenge 3c: Inference Paradigm

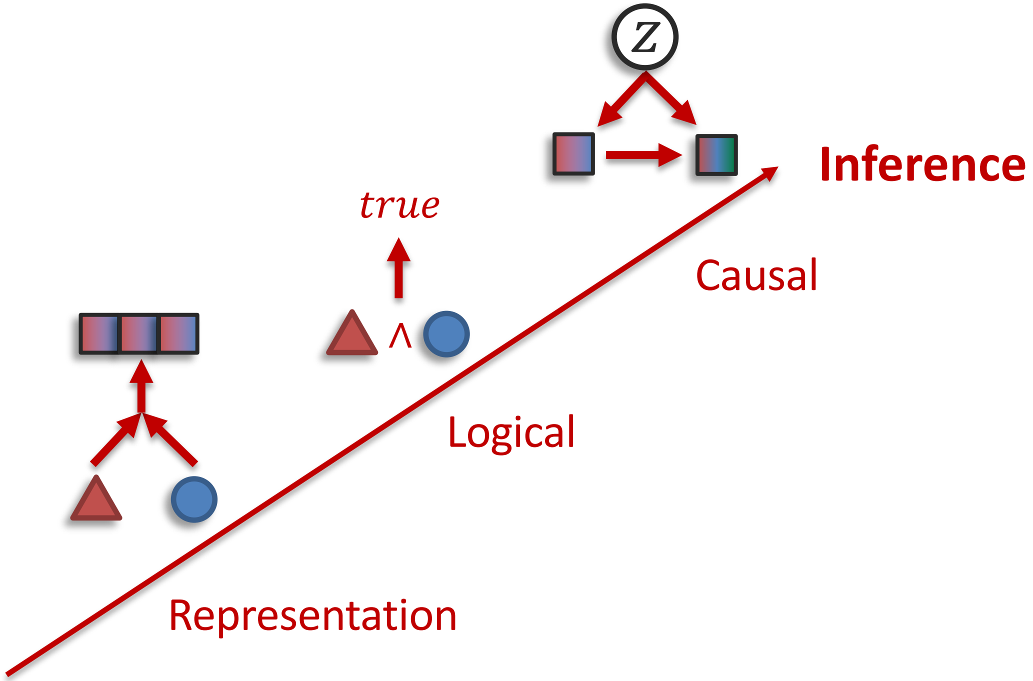
Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Towards explicit inference paradigms:
1. Logical inference
2. Causal inference

Nice, but you don't get these for free!



i.e., confounding variable



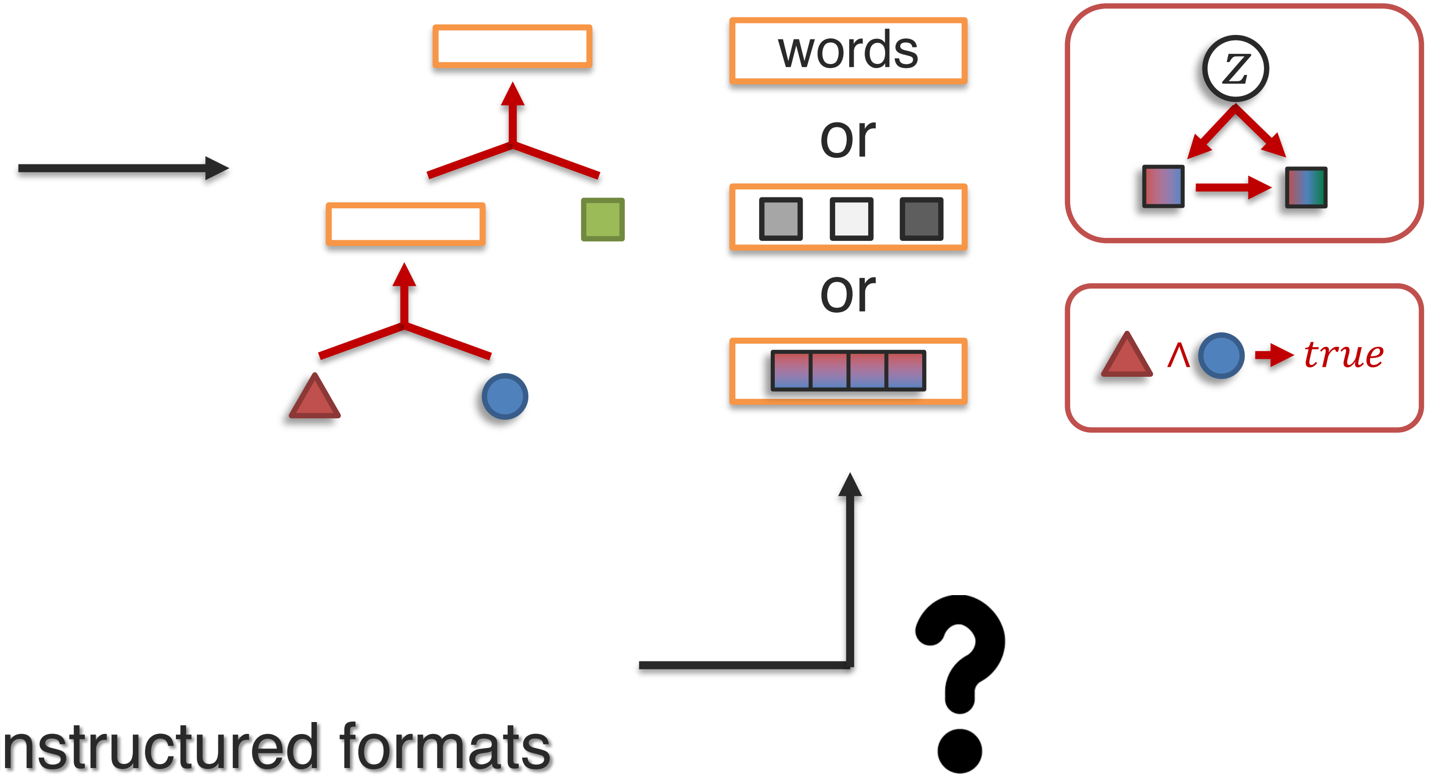
Sub-Challenge 3d: Knowledge

Definition: The derivation of knowledge in the study of inference, structure, and reasoning.

Domain knowledge

Knowledge graphs

Knowledge in other unstructured formats



External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources

What kind of board is this?

Requires knowledge of water sports, sports equipment, etc.

Existing models struggle when external knowledge is needed.
How can we leverage external knowledge?

[Marino et al., OK-VQA: A visual question answering benchmark requiring external knowledge. CVPR 2019]

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources

What kind of board is this?

Concepts: interpretable language

Object detector

Wikidata

- ***Wakeboard boat**: boat designed to create a wake…*
- ***Wakeboarder**: …*
- ***Kitesurfer**: …*
- ***Skiboarding**: …*
- ***Boardsport**: …*

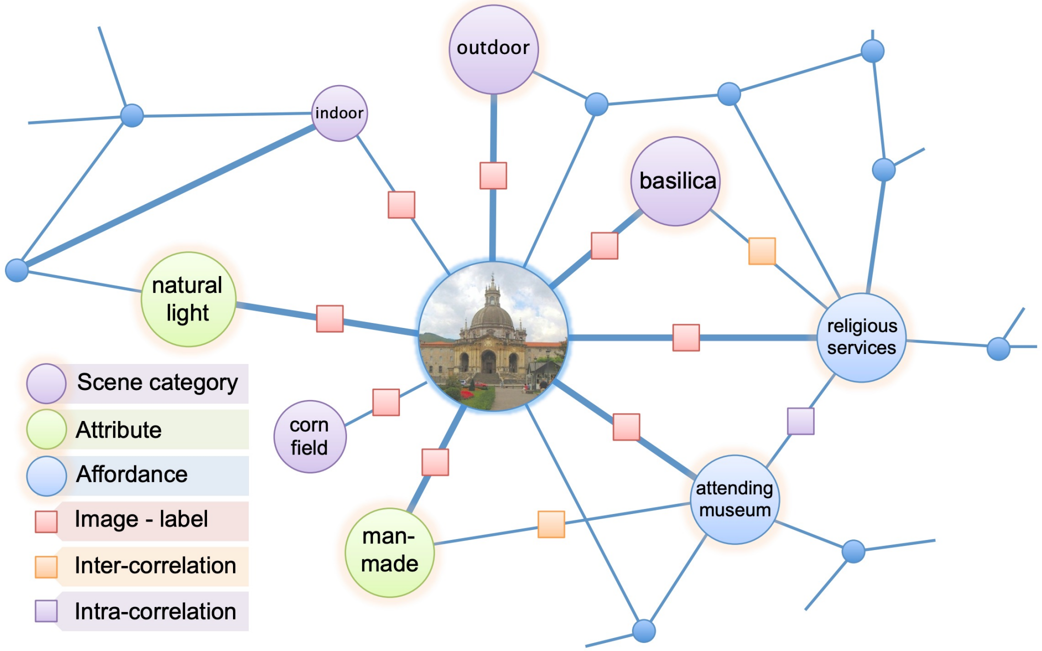
Language model → ***surfboard***

Structure: multi-step retrieval **Composition: neural**

[Gui et al., KAT: A Knowledge Augmented Transformer for Vision-and-Language. NAACL 2022]

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources



Class **auditorium**

Affordances community and social work, taking class for personal interest, religious practices, waiting, attending the performing arts

Attributes congregating, indoor lighting, spectating, enclosed area, glossy

Concepts: interpretable

Structure: multi-step inference

Composition: graph-based

[Zhu et al., Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. arXiv 2015]

External Knowledge Challenges

Open challenges

Yejin Choi

Brett Helsel Professor
Office: 578 Allen Center
Fax: 206-685-2969
email: yejin@cs.washington.edu

Paul G. Allen School of Computer Science & Engineering
University of Washington
Box 352350
185 E Stevens Way NE
Seattle, WA 98195-2350

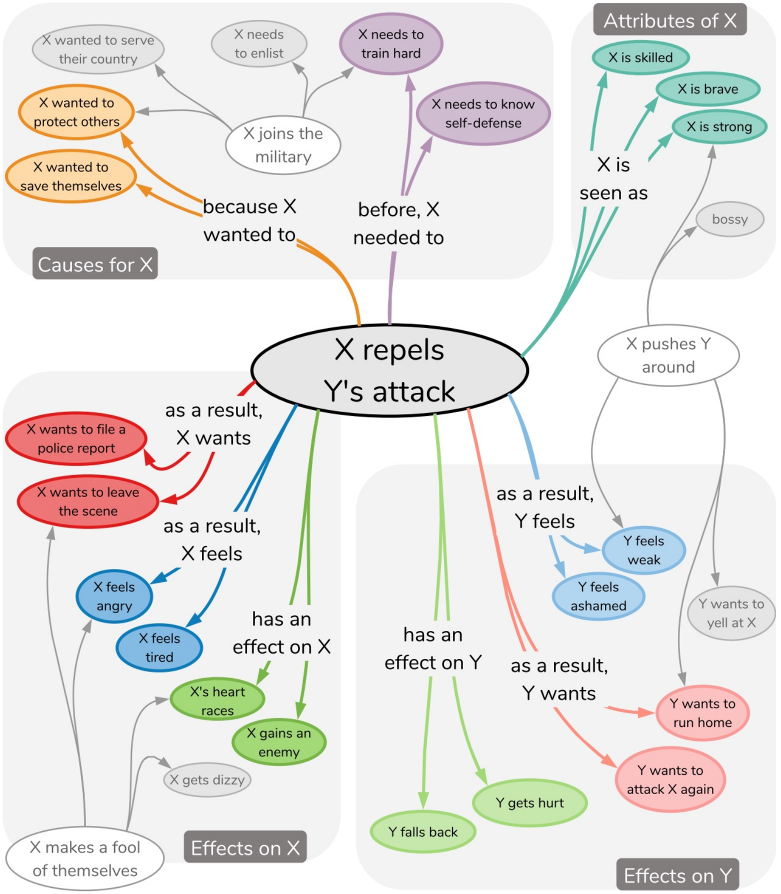
Allen Institute for Artificial Intelligence
2157 N Northlake Way, Suite 110
Seattle, WA 98103

Photo credit: Bruce Hemingway

MOSAIC COMMONSENSE

News:

- Outstanding Paper Award at ICML 2022
- Best Paper Award at NAACL 2022
- Keynote at ACL: ***"2082: An ACL Odyssey:** The Dark Matter of Language and Intelligence"* along with a fireside chat on *"The Trajectory of ACL and the Next 60 years"*
- An invited article, *"The Curious Case of Commonsense Intellgience"* for the Daedalus's special issue on AI & Society
- A podcast interview with the Gradient on commonsense and morality

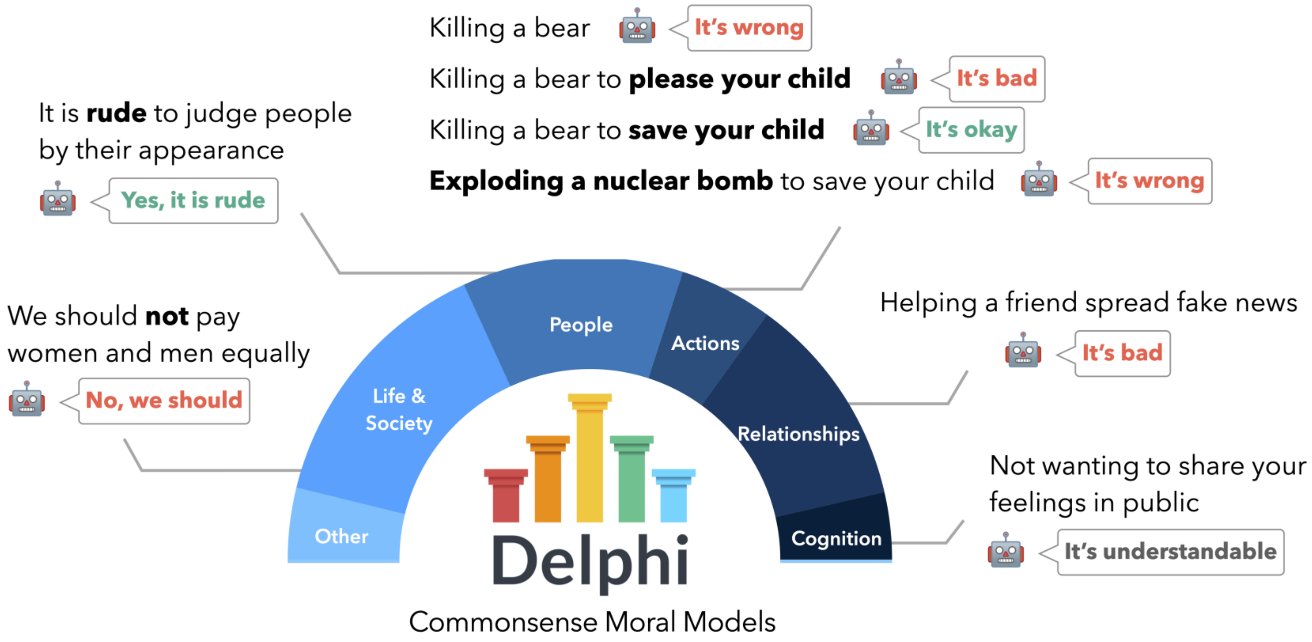


Atomic: If-then commonsense

[Sap et al., Atomic: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI 2019]

External Knowledge Challenges

Open challenges



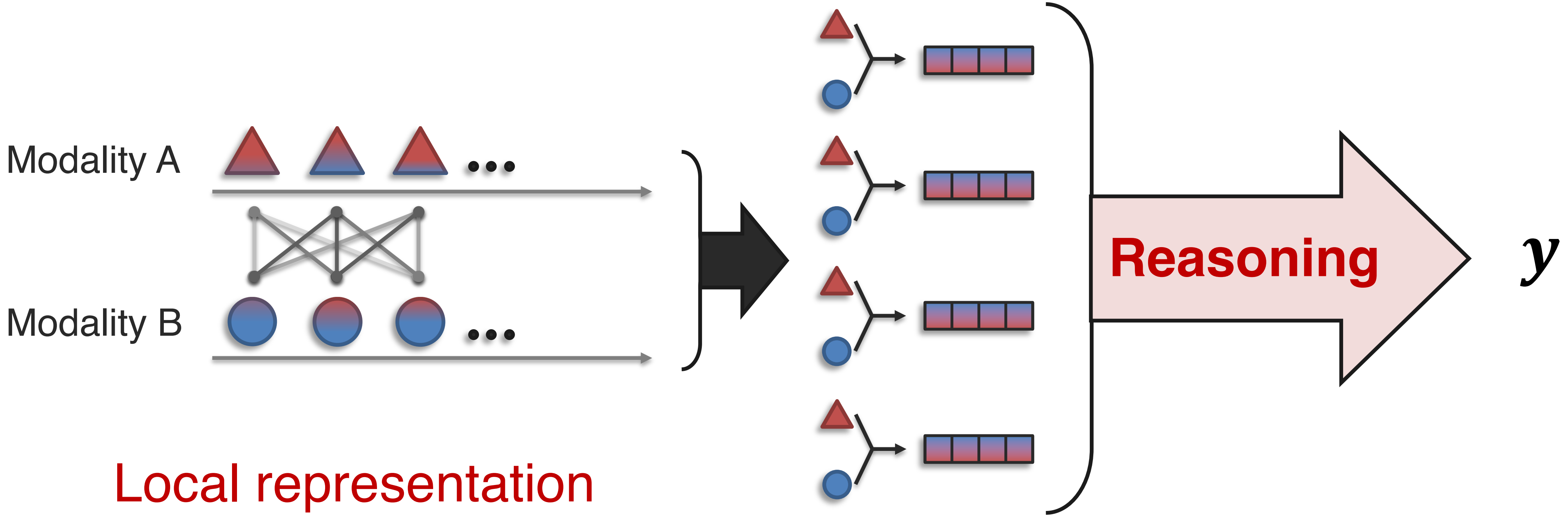
Delphi: Moral commonsense

Social Chemistry: Social commonsense

[Jiang et al., Can Machines Learn Morality? The Delphi Experiment. arXiv 2021]
[Forbes et al., Social Chemistry 101: Learning to Reason about Social and Moral Norms. EMNLP 2020]

Summary: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



Local representation

+ Aligned representation

The Challenge of Compositionality

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

CLIP, ViLT, ViLBERT, etc.
All random chance

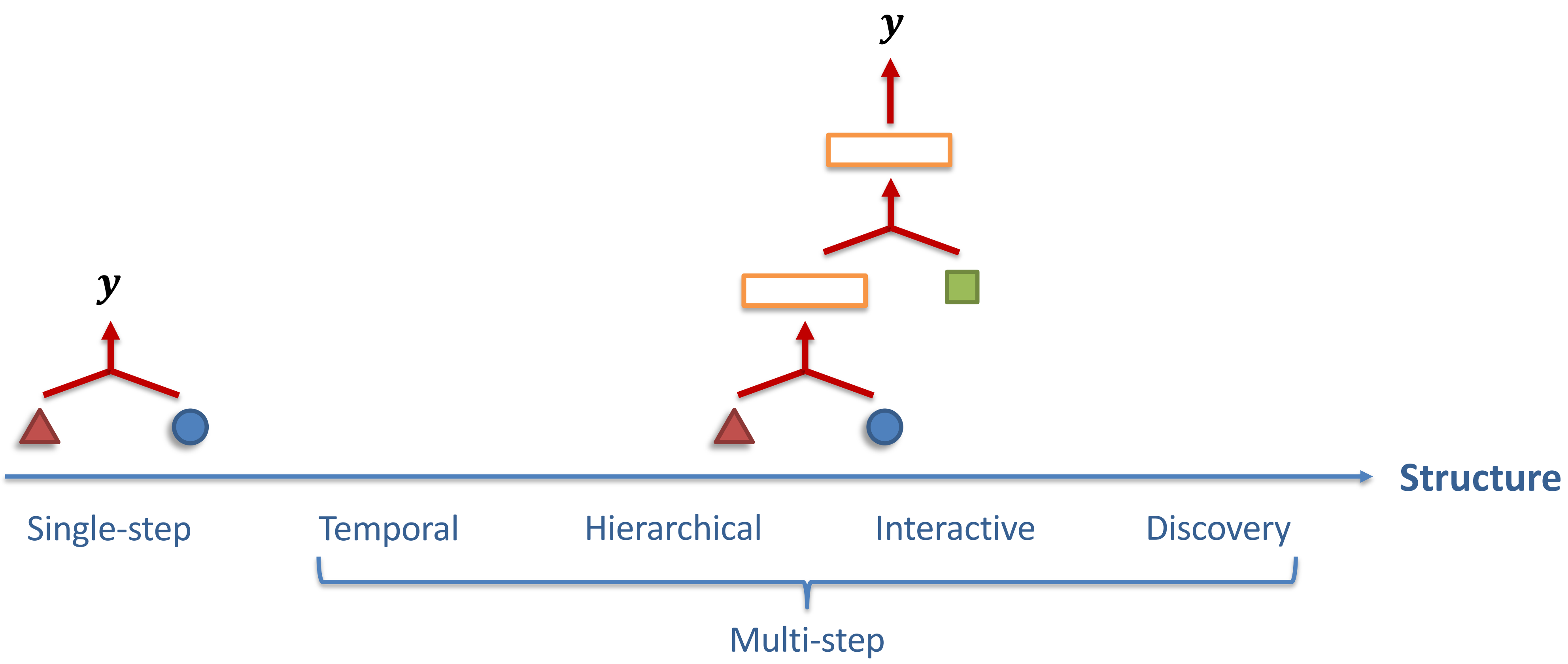
Compositional Generalization
to novel combinations outside of training data

1. Structure: <subject> <verb> <object>
2. Concepts: ‘plants’, ‘lightbulb’
3. Inference: ‘surrounding’ – spatial relation
4. Knowledge: from humans!

[Thrush et al., Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. CVPR 2022]

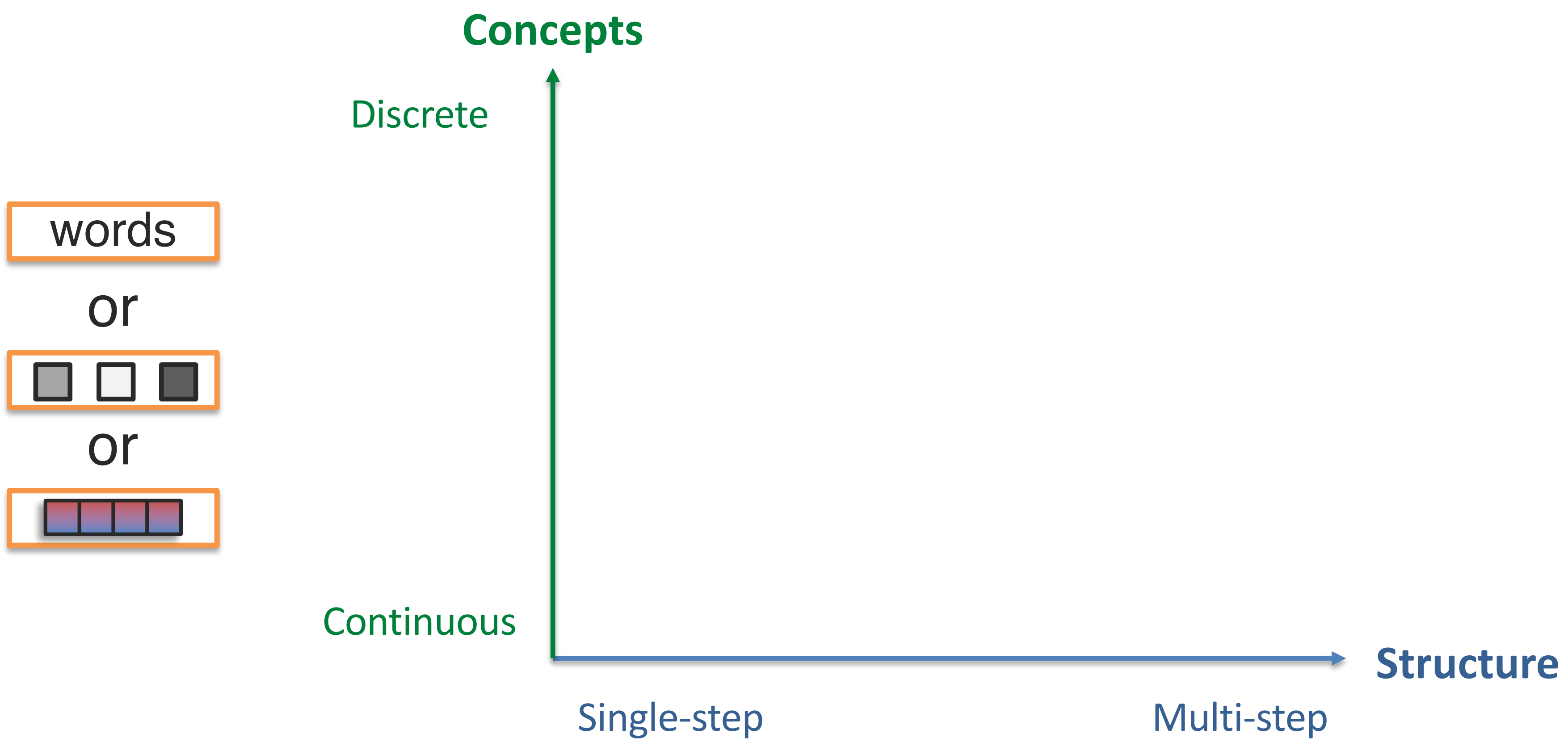
Sub-Challenge 3a: Structure Modeling

Definition: Defining or learning the relationships over which reasoning occurs.



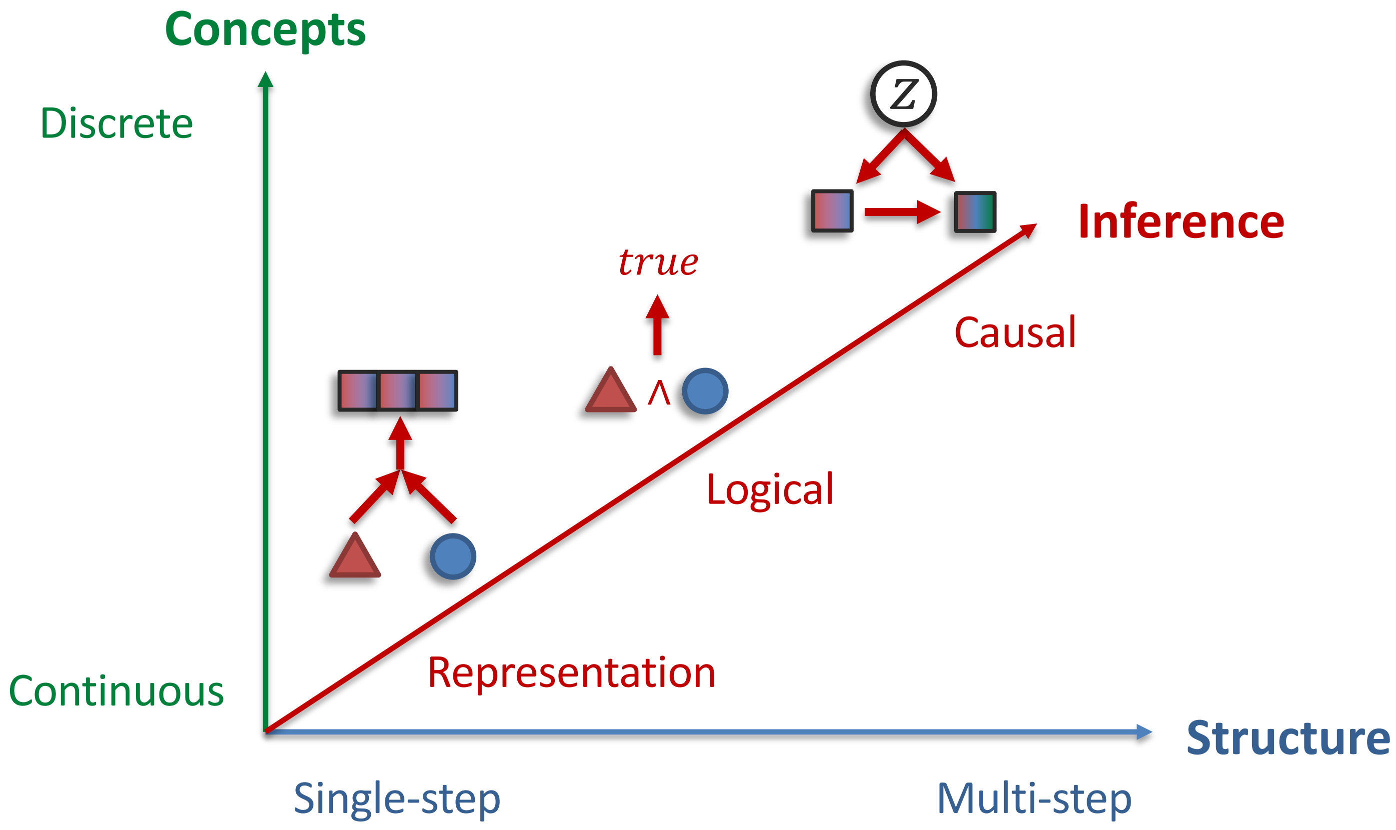
Sub-Challenge 3b: Intermediate Concepts

Definition: The parameterization of individual multimodal concepts in the reasoning process.



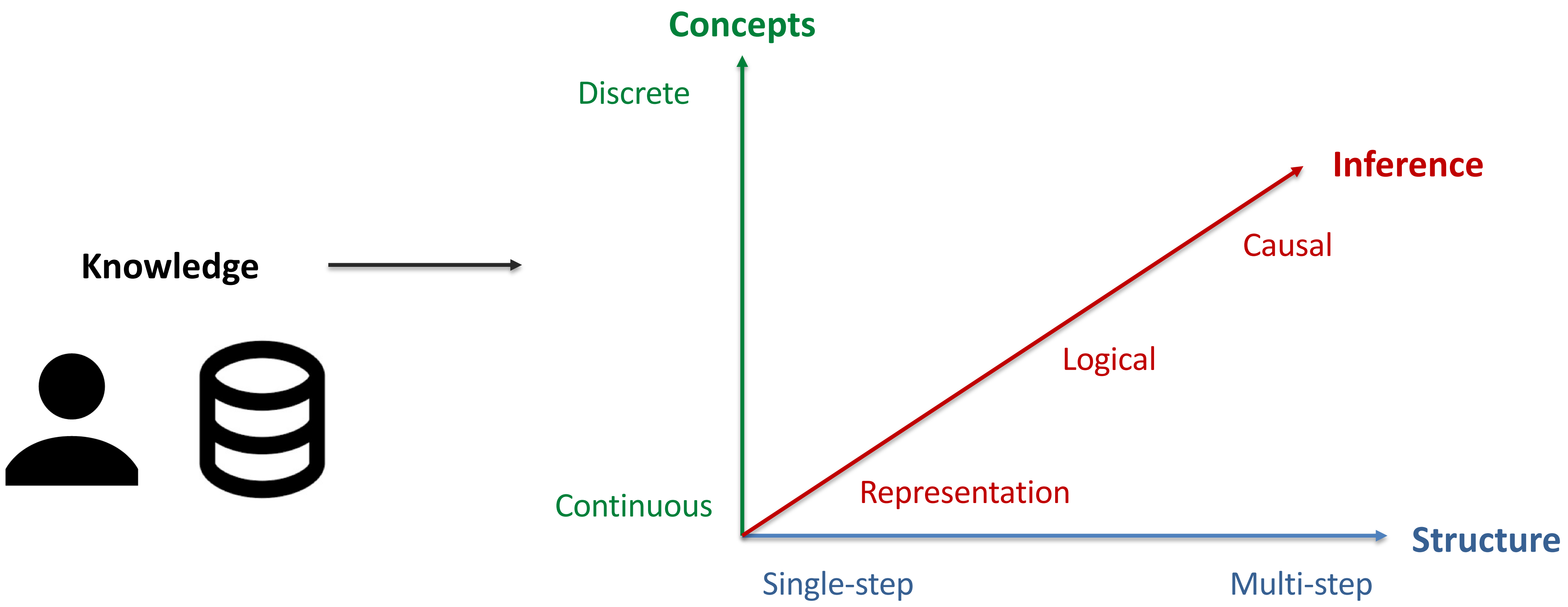
Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.



Sub-Challenge 3d: External Knowledge

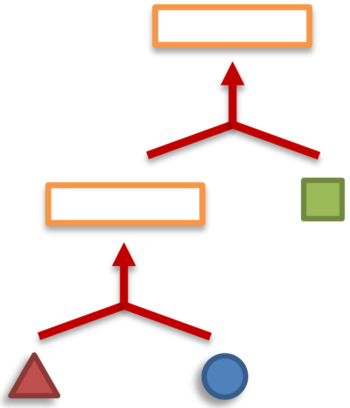
Definition: Leveraging external knowledge in the study of structure, concepts, and inference.



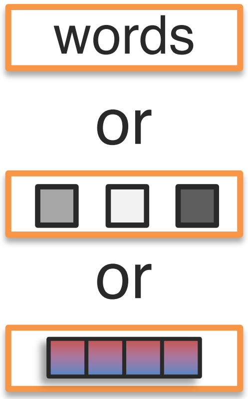
Summary: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

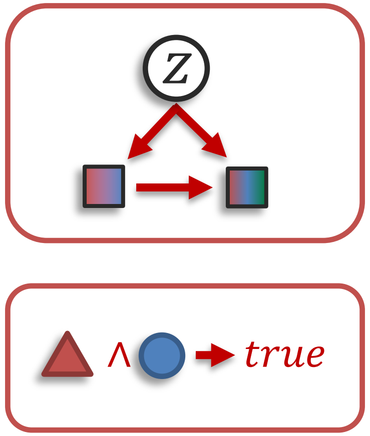
Ⓐ **Structure modeling**



Ⓑ **Intermediate concepts**



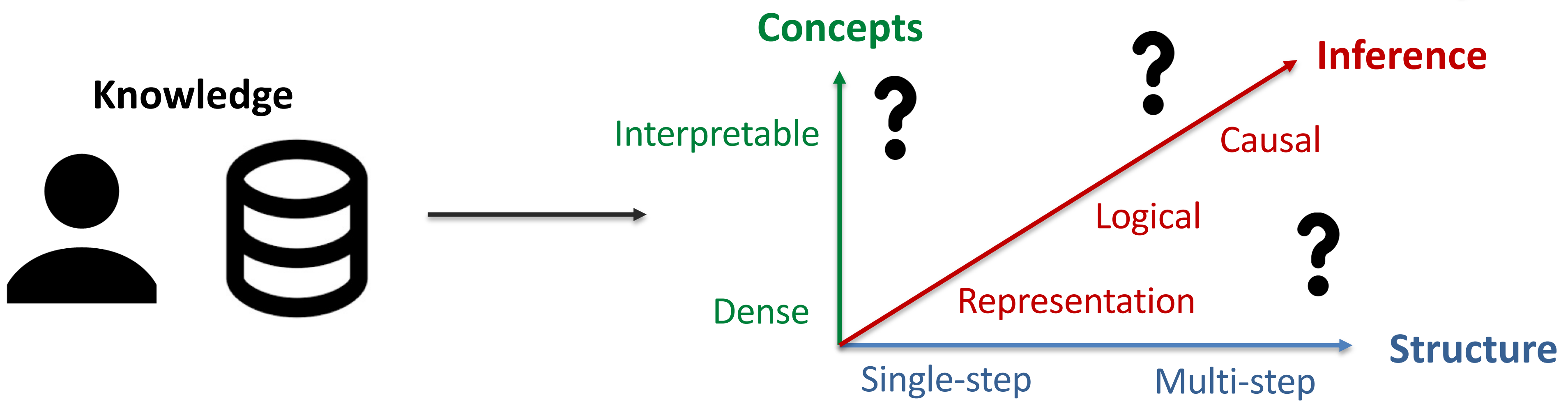
Ⓒ **Inference paradigm**



Ⓓ **External knowledge**

More Reasoning

Open challenges

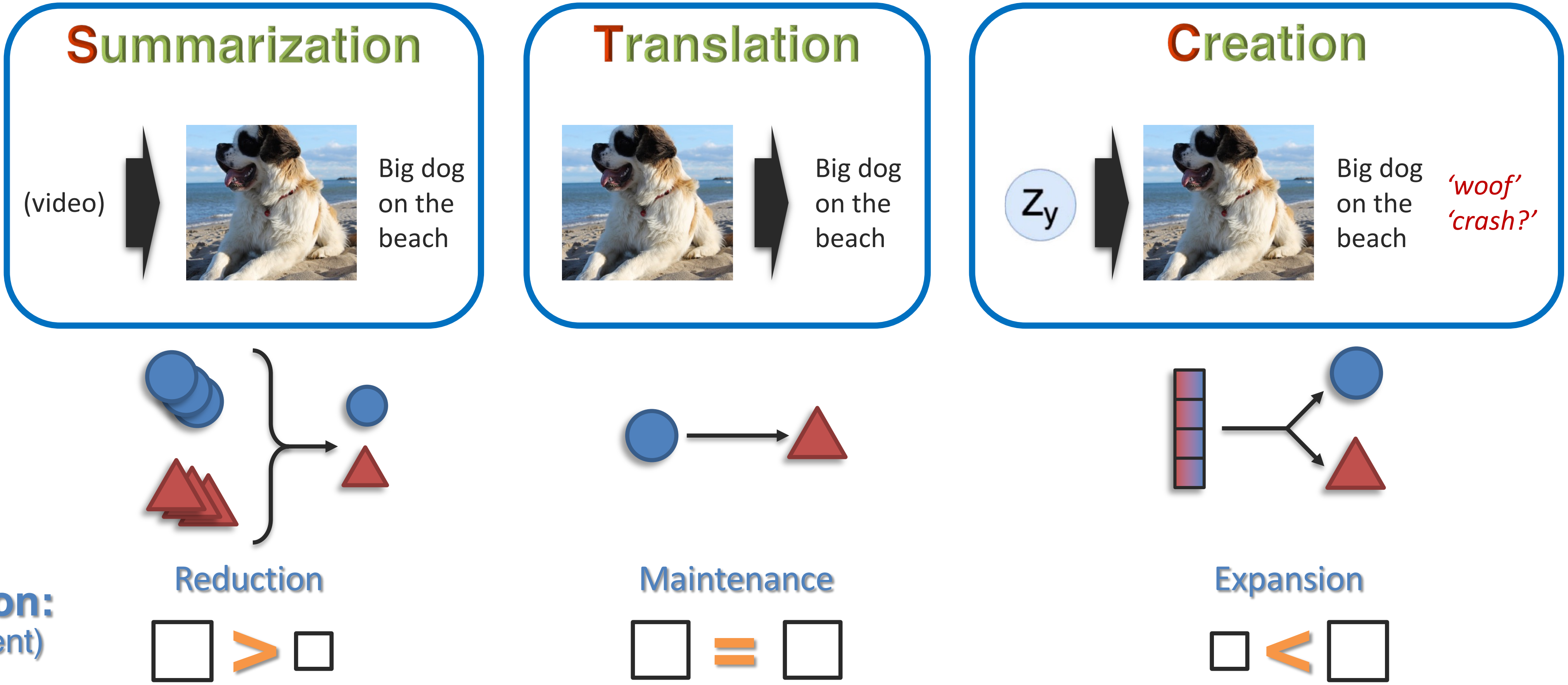


Open challenges:

- Structure: multi-step inference
- Concepts: interpretable + differentiable representations
- Composition: explicit, logical, causal…
- Knowledge: integrating explicit knowledge with pretrained models
- Probing pretraining models for reasoning capabilities

Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



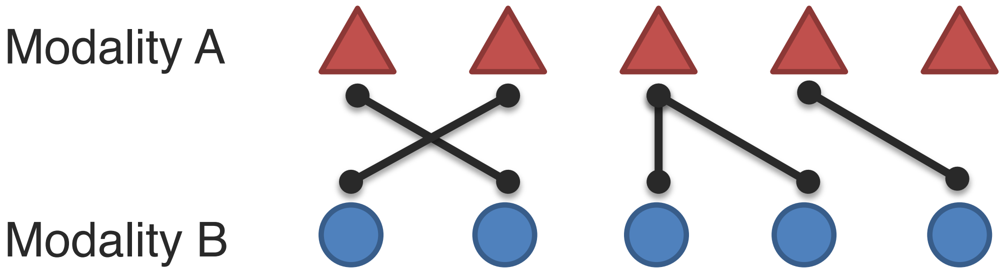
Information: (content)

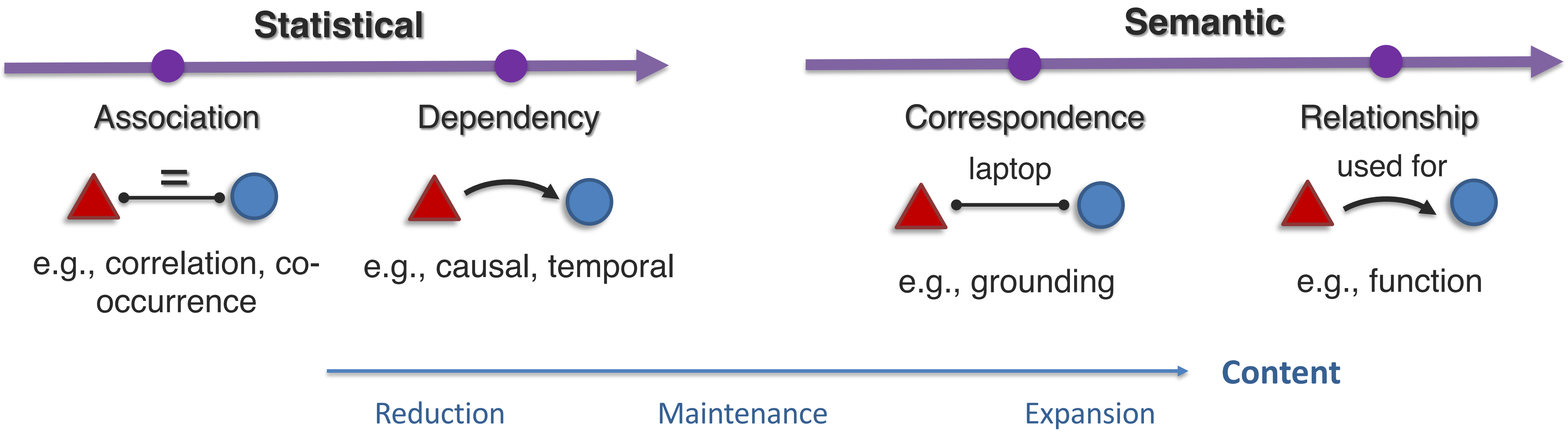
Dimension 1: Information Content

How modality interconnections change across multimodal inputs and generated outputs.

1 Modality connections

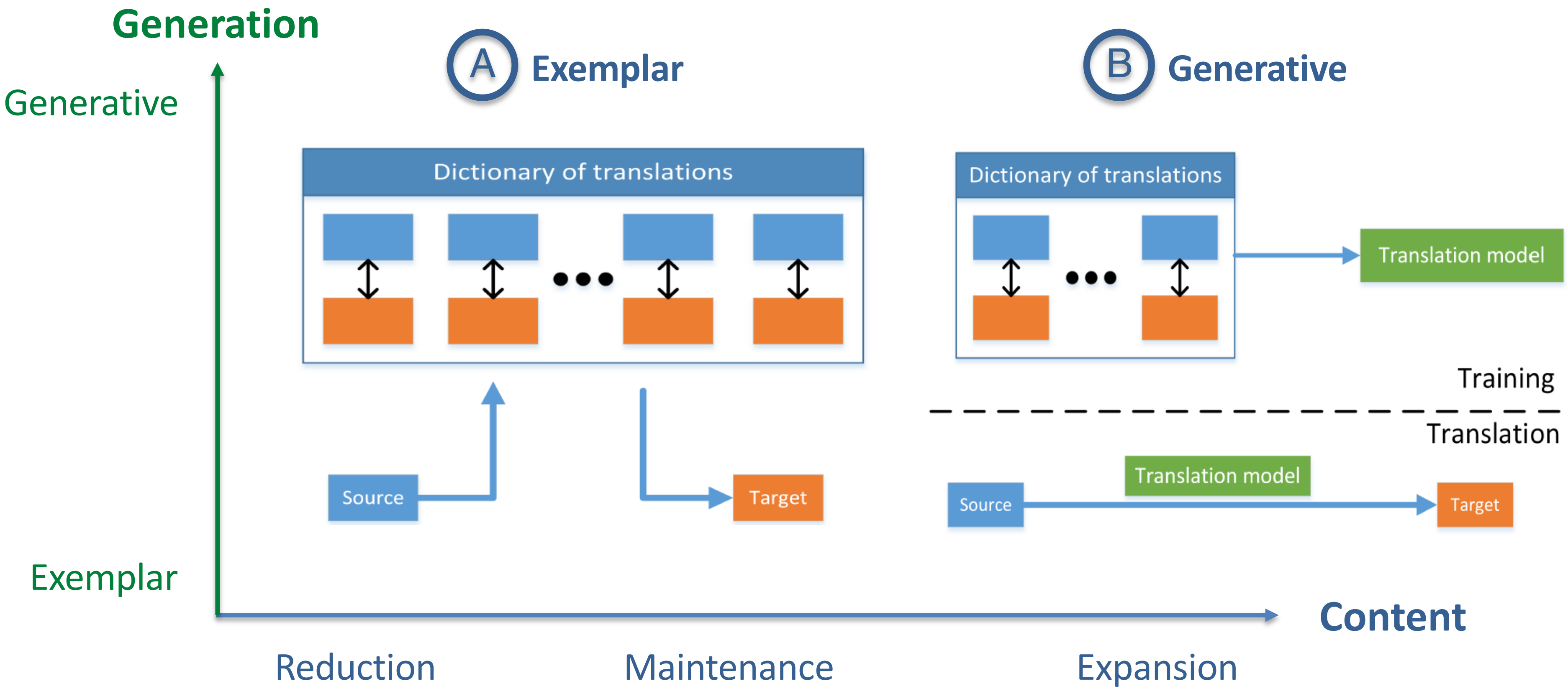
Modalities are often related and share commonality





Dimension 2: Generative Process

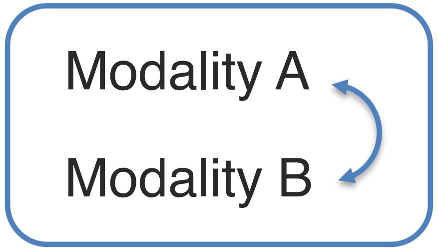
Generative process to respect modality heterogeneity and decode multimodal data.



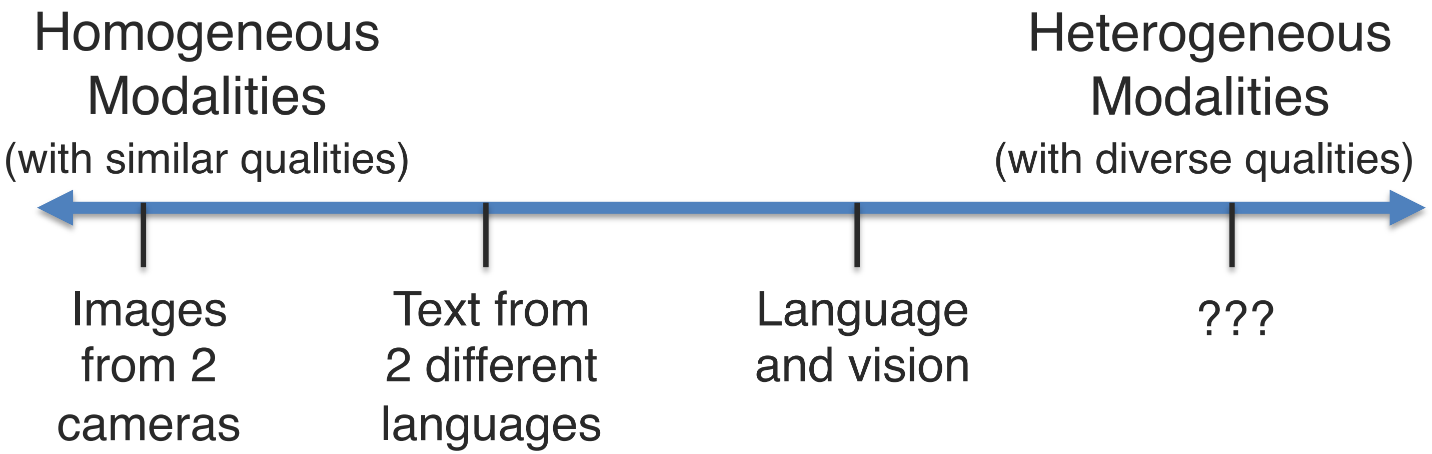
Dimension 2: Generative Process

Heterogeneous modalities

Information present in different modalities will often show diverse qualities, structures and representations.



Examples:

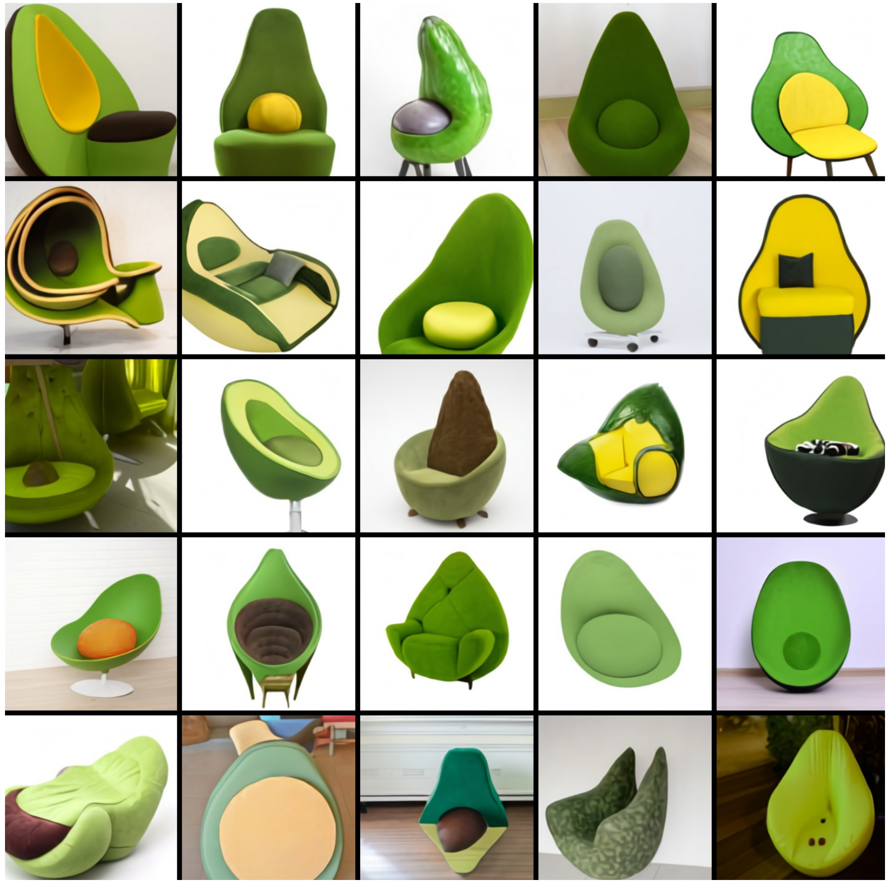


Abstract modalities are more likely to be homogeneous

Sub-challenge 4a: Translation

Definition: Translating from one modality to another and keeping information content while being consistent with cross-modal interactions.

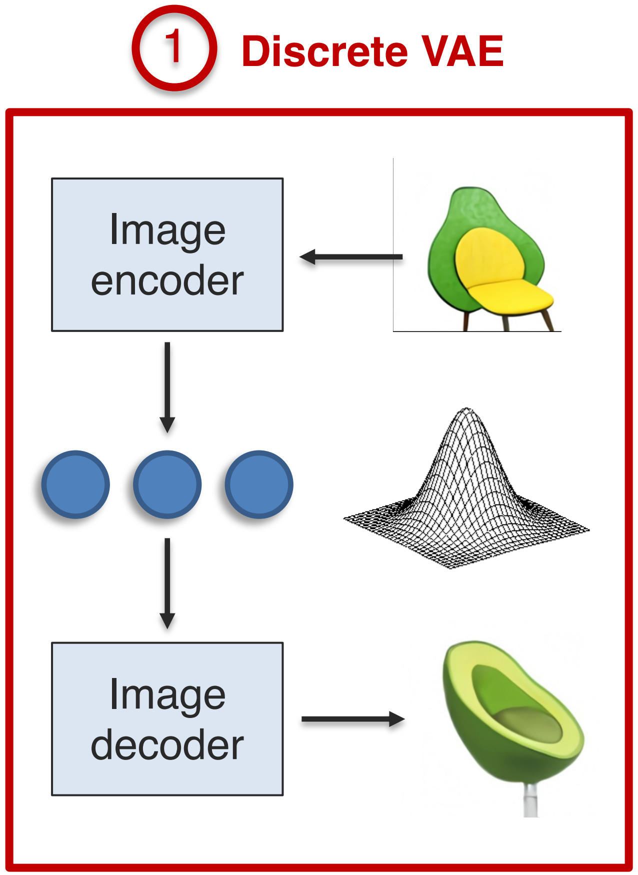
An armchair in the shape of an avocado →



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

Sub-challenge 4a: Translation

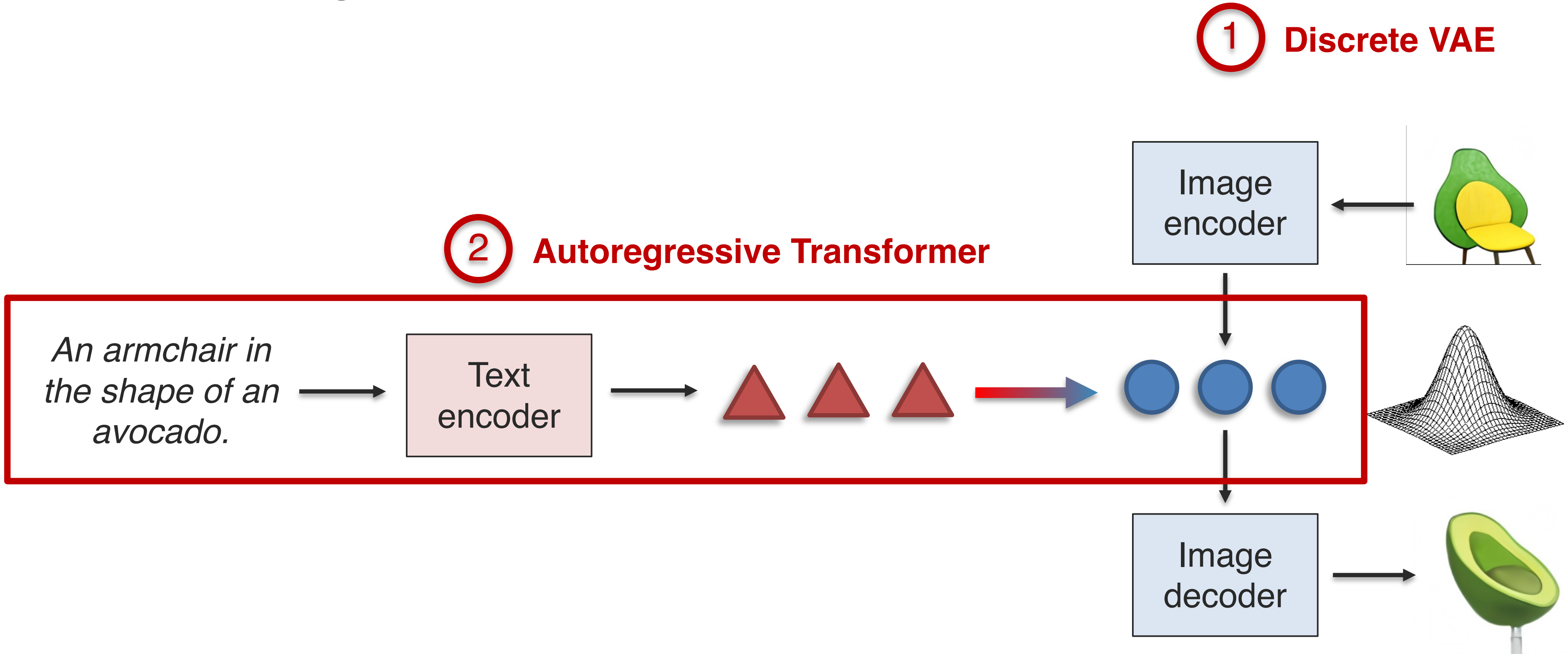
DALL·E: Text-to-image translation at scale



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

Sub-challenge 4a: Translation

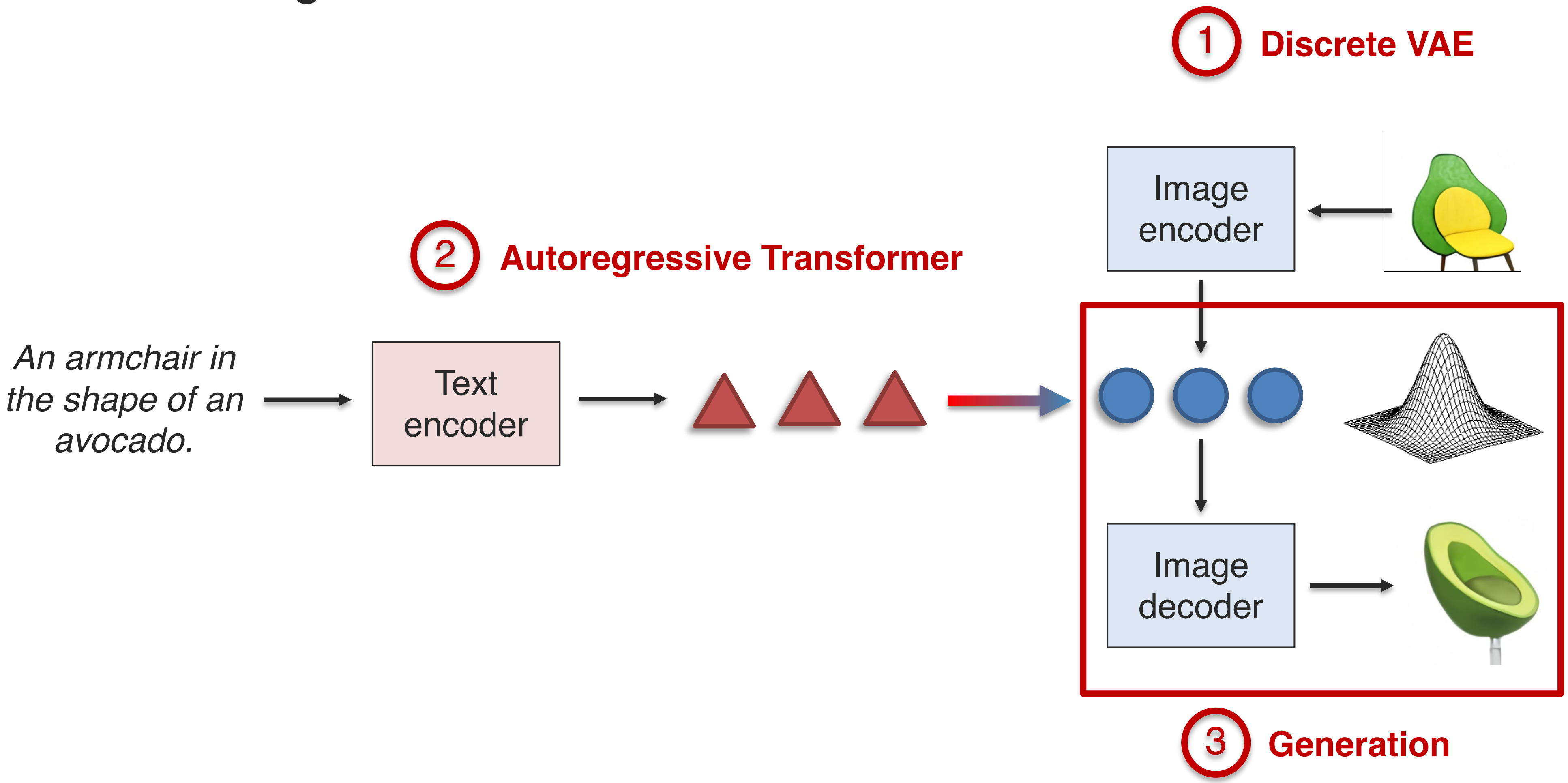
DALL·E: Text-to-image translation at scale



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

Sub-challenge 4a: Translation

DALL·E: Text-to-image translation at scale



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

Sub-challenge 4a: Translation

DALL·E: Text-to-image translation at scale

(A) Content

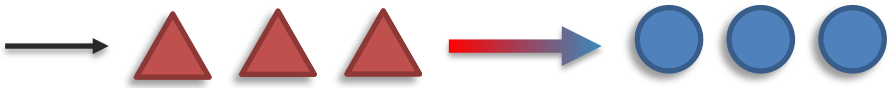
Coordination via
supervised translation

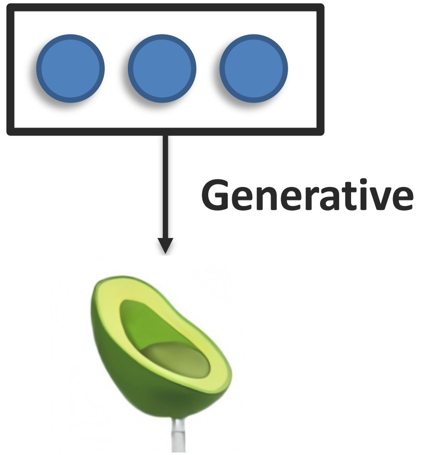
Capture **corresponding**
cross-modal interactions

(B) Generation

Exemplar (discrete
visual codebook)

An armchair in the shape of an avocado.

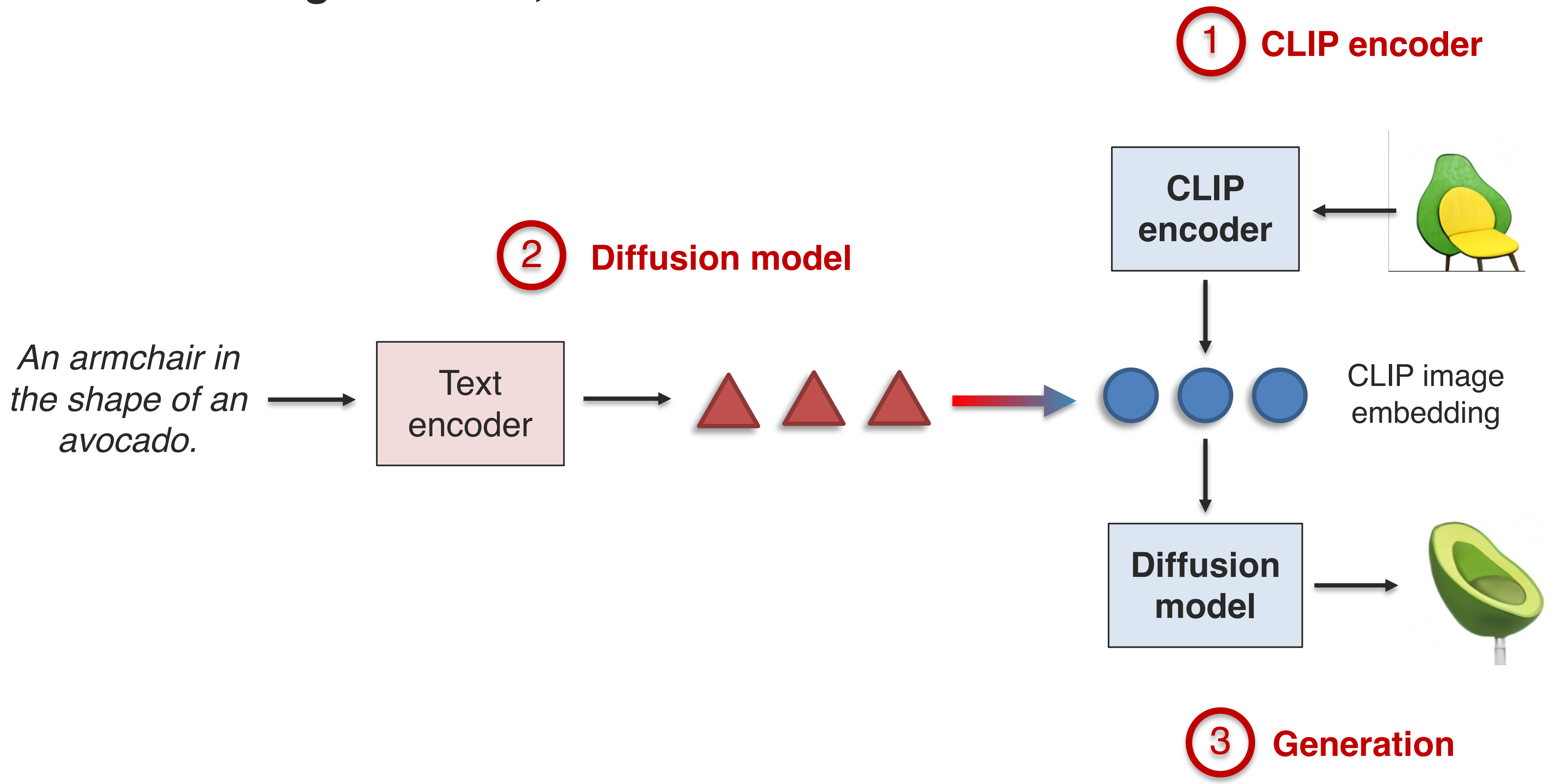




[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

Sub-challenge 4a: Translation

DALL·E 2: Combining with CLIP, diffusion models



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

Sub-challenge 4a: Translation

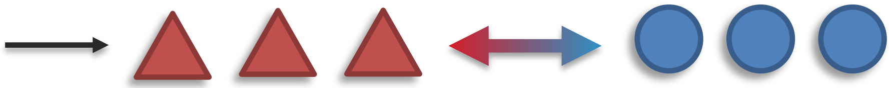
DALL·E 2: Combining with CLIP, diffusion models

A Content

Coordination via
CLIP similarity

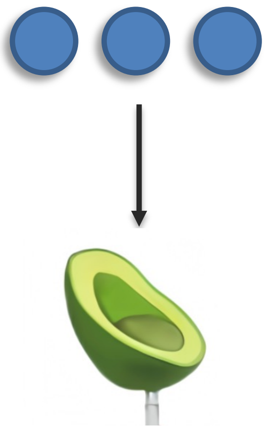
Capture **corresponding**
cross-modal interactions

An armchair in the shape of an avocado.



B Generation

Fully **generative**
(diffusion models)



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

Sub-challenge 4b: Summarization

Definition: Summarizing multimodal data to reduce information content while highlighting the most salient parts of the input.

Transcript

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't .... t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

Video

How2 video dataset

Complementary cross-modal interactions

Cuban breakfast
Free cooking video

(not present in text)

Summary

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

[Palaskar et al., Multimodal Abstractive Summarization for How2 Videos. ACL 2019]

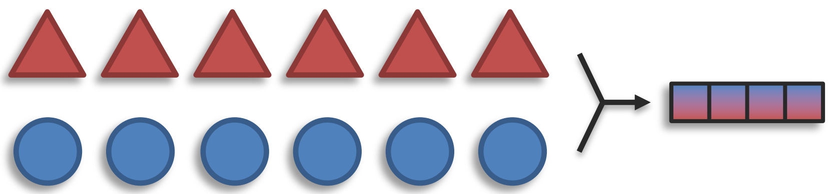
Sub-challenge 4b: Summarization

Video summarization

(A) **Content**

Fusion via
joint representation

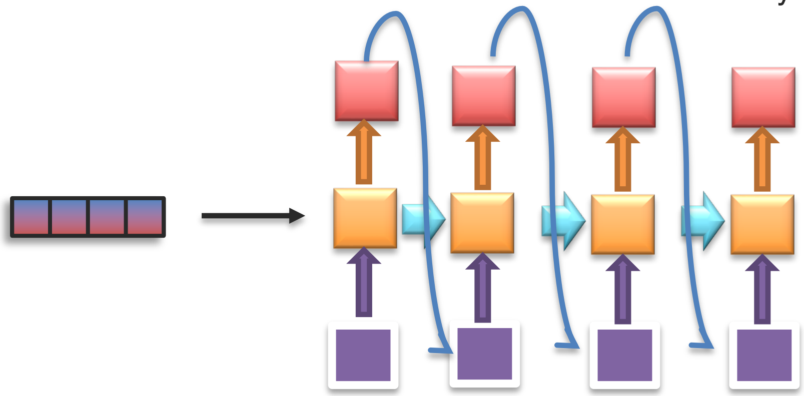
Capture **complementary**
cross-modal interactions



(B) **Generation**

Generative ≈ abstractive summarization

Exemplar ≈ extractive summarization



[Palaskar et al., Multimodal Abstractive Summarization for How2 Videos. ACL 2019]

Sub-challenge 4c: Creation

Open challenges

Definition: Simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities.

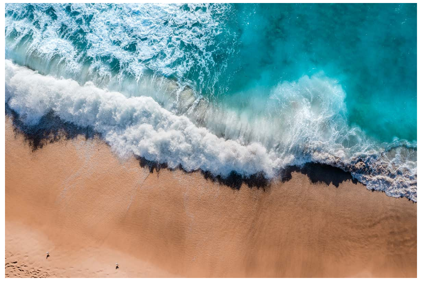
z_y

z_{a1}

z_{a2}

z_y

Many goals!



Cross-modal interactions

Big dog on the beach. Waves crashing, people playing volleyball, …

Cross-modal interactions

'woof' *'crash'* *'bounce'* *'whoosh'*

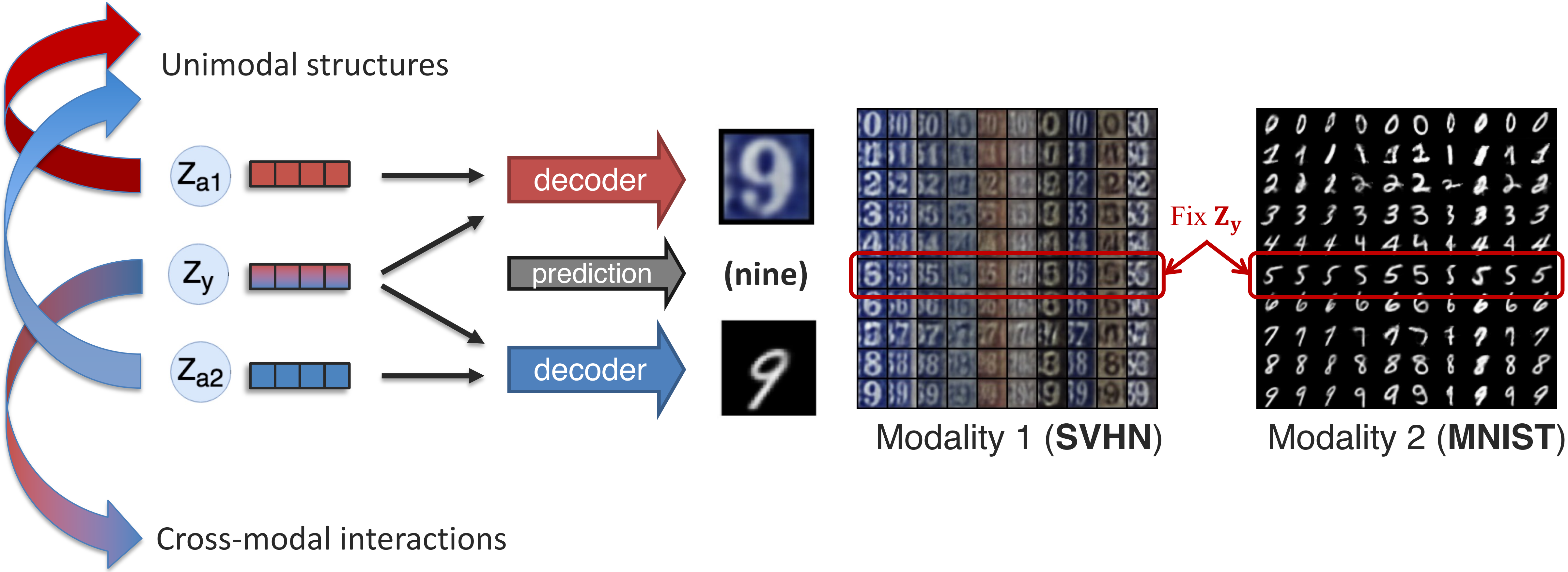
Temporal + causal + logical structure

Recall representation & alignment!

Recall reasoning!

Sub-challenge 4c: Creation

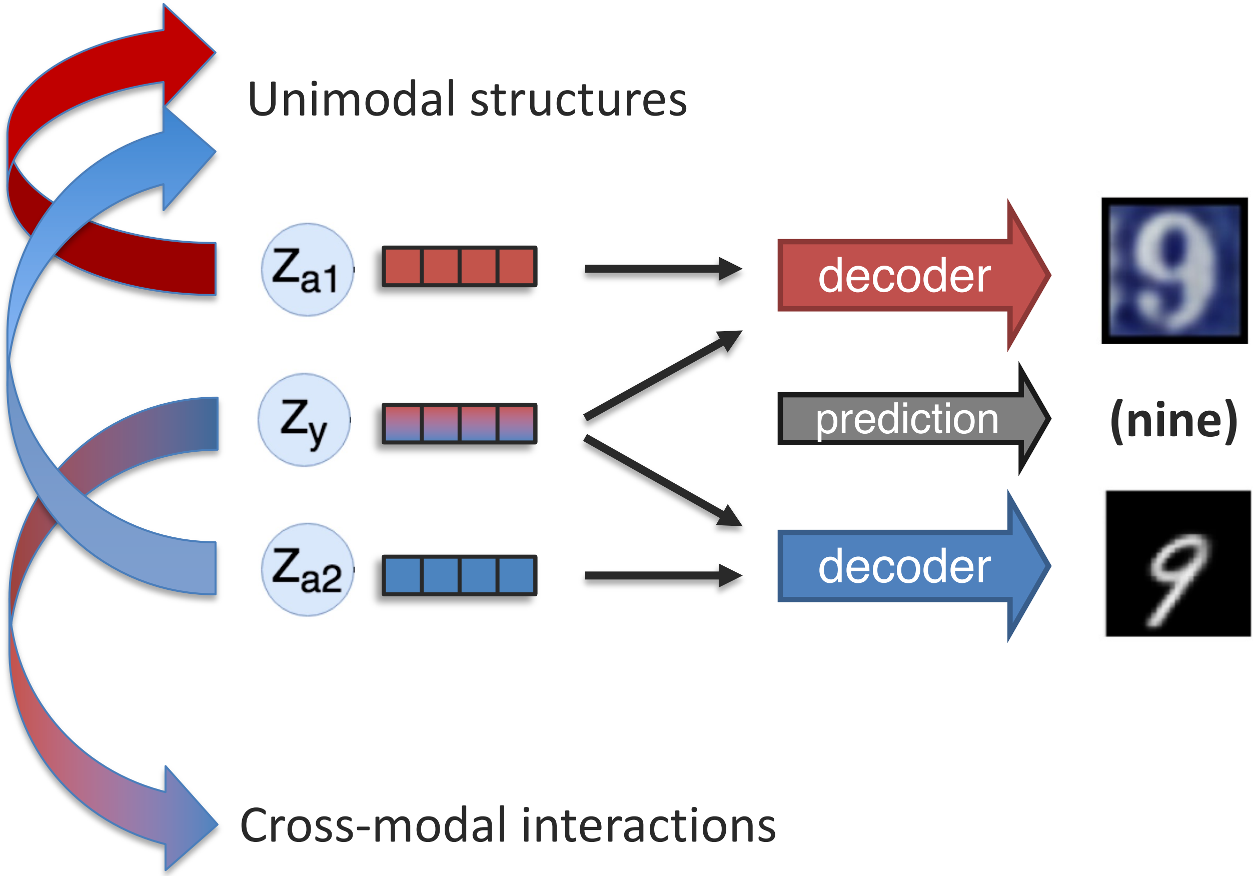
Some initial attempts: factorized generation

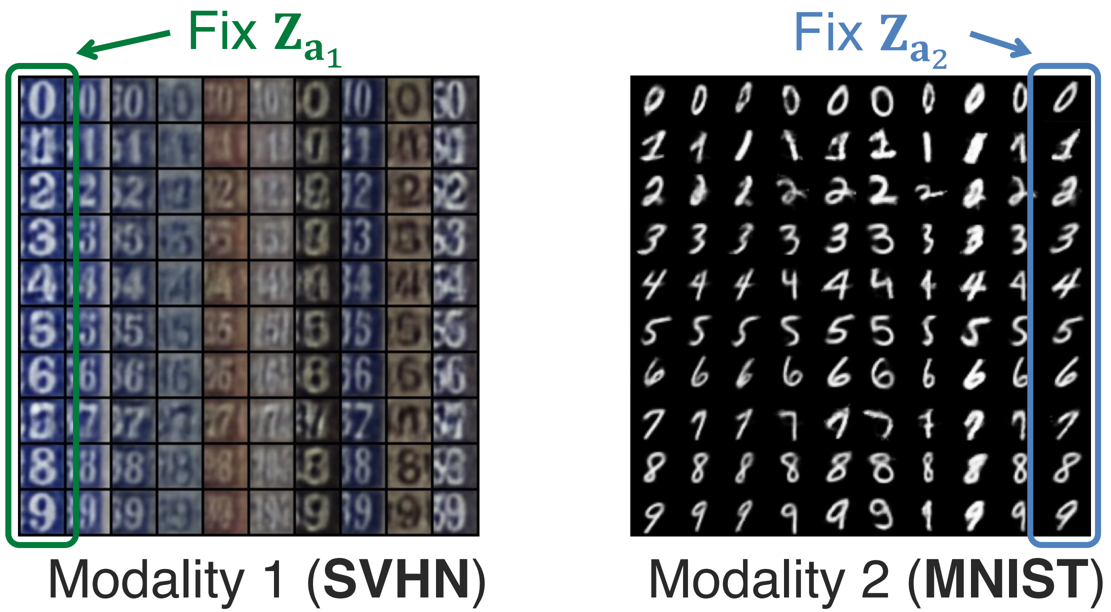


[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Sub-challenge 4c: Creation

Some initial attempts: factorized generation





[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Sub-challenge 4c: Creation

Some initial attempts: factorized generation

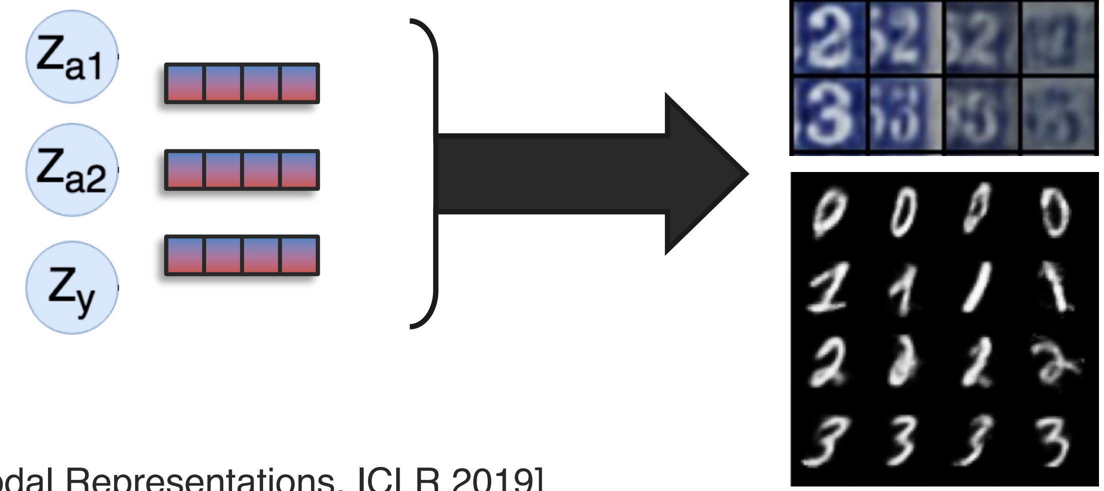
Ⓐ **Content**

Factorized **representation**

Expanding **complementary** cross-modal interactions

Ⓑ **Generation**

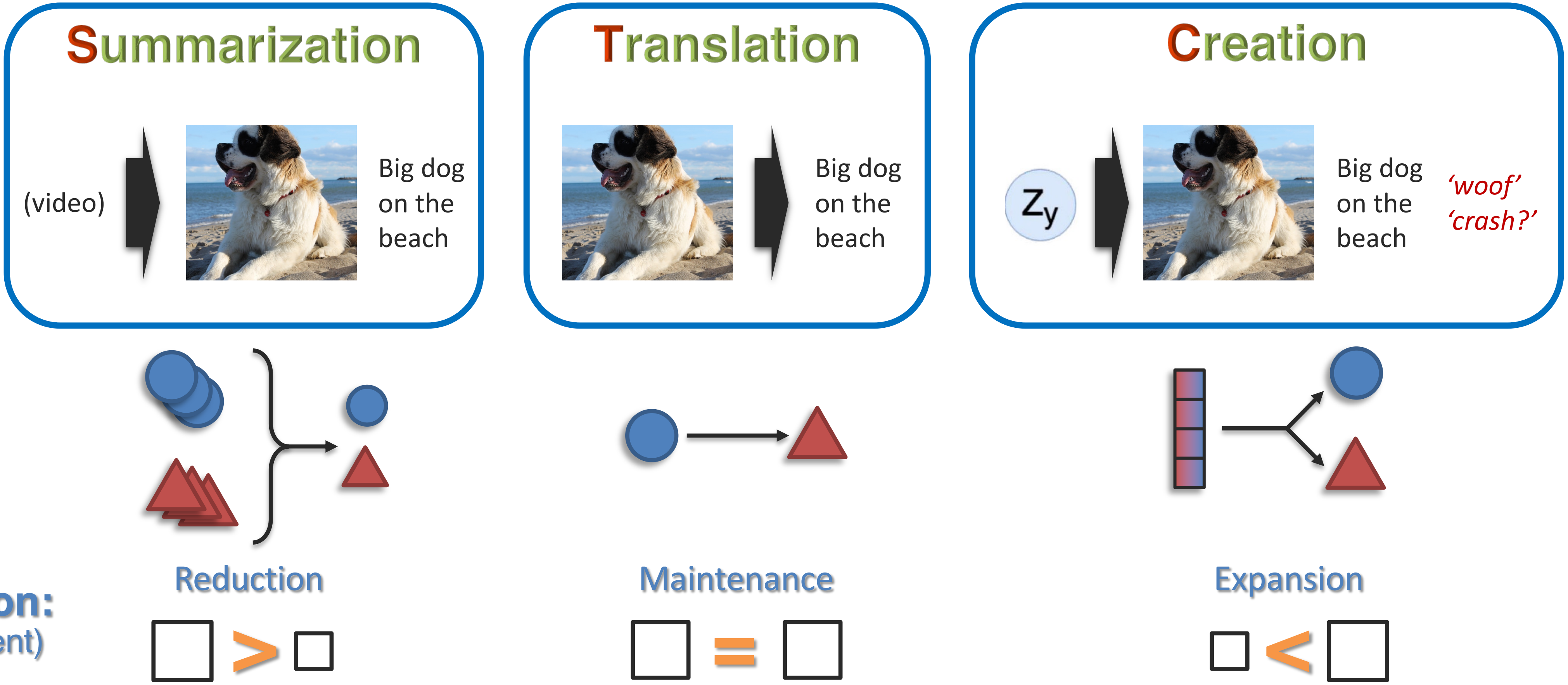
Generative model



[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Preview: Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



Information: (content)

Model Evaluation & Ethical Concerns

Open challenges

Open challenges:

- Modalities beyond text + images or video
- Translation beyond descriptive text and images (beyond corresponding cross-modal interactions)
- Creation: fully multimodal generation, with cross-modal coherence + within modality consistency

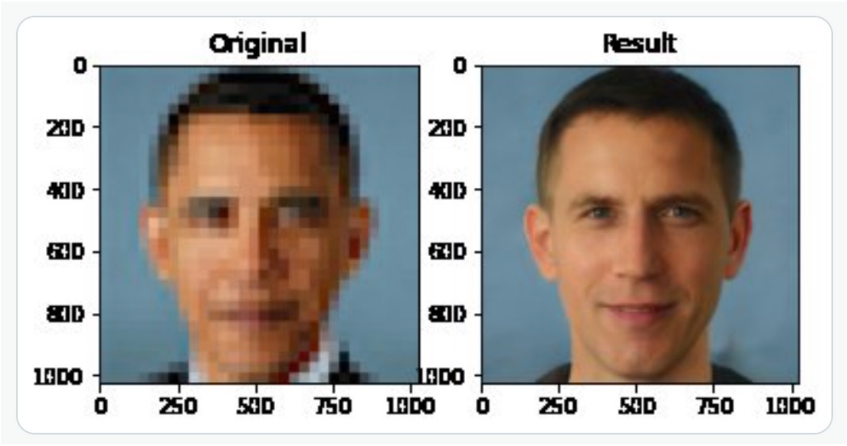
[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]
[Carlini et al., Extracting Training Data from Large Language Models. USENIX 2021]
[Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019]

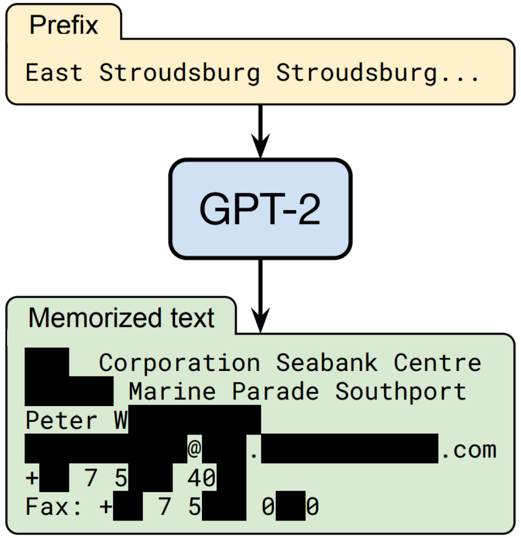
Model Evaluation & Ethical Concerns

Open challenges

Open challenges:

- Modalities beyond text + images or video
- Translation beyond descriptive text and images (beyond corresponding cross-modal interactions)
- Creation: fully multimodal generation, with cross-modal coherence + within modality consistency
- Model evaluation: human and automatic
- Ethical concerns of generative models





Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]
[Carlini et al., Extracting Training Data from Large Language Models. USENIX 2021]
[Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019]

Generative Models

Learn to model **p(x)** where x = text, images, videos, multimodal data

- Given x, **evaluate** p(x) - realistic data should have high p(x) and vice versa
- **Sample** new x according to p(x) - sample realistic looking images
- Unsupervised **representation** learning - we should be able to learn what these images have in common, e.g., ears, tail, etc. (features)

INPUT (x)	RECONSTRUCTION (AUTR)	RECONSTRUCTION (Gen-RNN)
unable to stop herself, she briefly, gently, touched his hand.	unable to stop herself, she leaned forward, and touched his eyes.	unable to help her , and her back and her into my way.
why didn't you tell me?	why didn't you tell me?	why didn't you tell me?"
a strange glow of sunlight shines down from above, paper white and blinding, with no heat.	the light of the sun was shining through the window, illuminating the room.	a tiny light on the door, and a few inches from behind him out of the door.
he handed her the slip of paper.	he handed her a piece of paper.	he took a sip of his drink.

Generative Models

Sometimes we also care about p(x|c) - **conditional generation**

- c is a category (e.g. faces, outdoor scenes) from which we want to generate images

We might also care about p(x2|x1,c) - **style transfer**

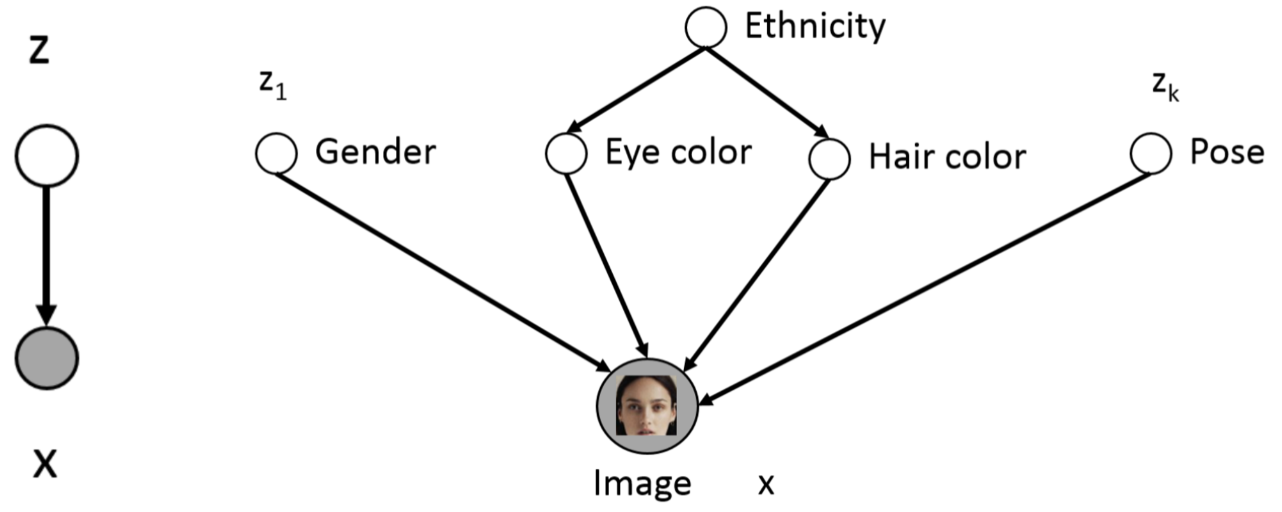
- c is a stylistic change e.g. negative to positive

From negative to positive
consistently slow . consistently good . consistently fast .
my goodness it was so gross . my husband 's steak was phenomenal . my goodness was so awesome .
it was super dry and had a weird taste to the entire slice . it was a great meal and the tacos were very kind of good . it was super flavorful and had a nice texture of the whole side .

Latent Variable Models

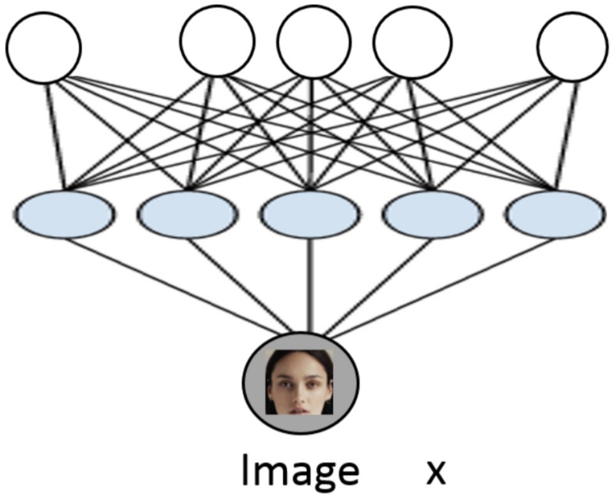
- Lots of variability in images **x** due to gender, eye color, hair color, pose, etc.
- However, unless images are annotated, these factors of variation are not explicitly available (latent).
- Idea: explicitly model these factors using latent variables **z**

Latent Variable Models



- Only shaded variables **x** are observed in the data
- Latent variables **z** are unobserved - correspond to high-level features
  - We want z to represent useful features e.g. hair color, pose, etc.
  - But very difficult to specify these conditionals by hand and they're unobserved
  - Let's **learn** them instead

Latent Variable Models



- Put a prior on z $\mathbf{z} \sim \mathcal{N}(0, I)$

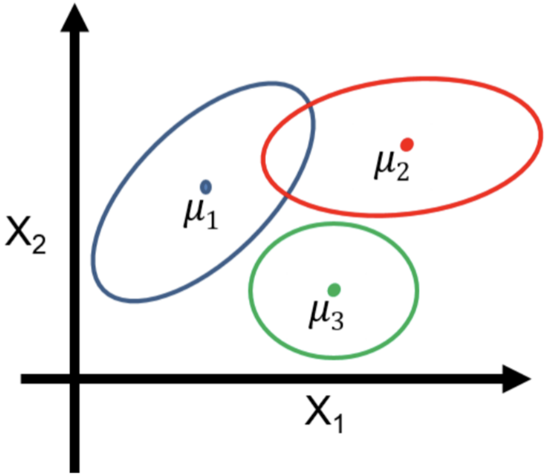
  $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$ where μ_θ,Σ_θ are neural networks
- Hope that after training, z will correspond to meaningful latent factors of variation - useful features for unsupervised representation learning
- Given a new image x, features can be extracted via p(zlx)

Mixture of Gaussians

Mixture of Gaussians (Bayes network z -> x)

$$\mathbf{z} \sim \text{Categorical}(1, \cdots, K)$$

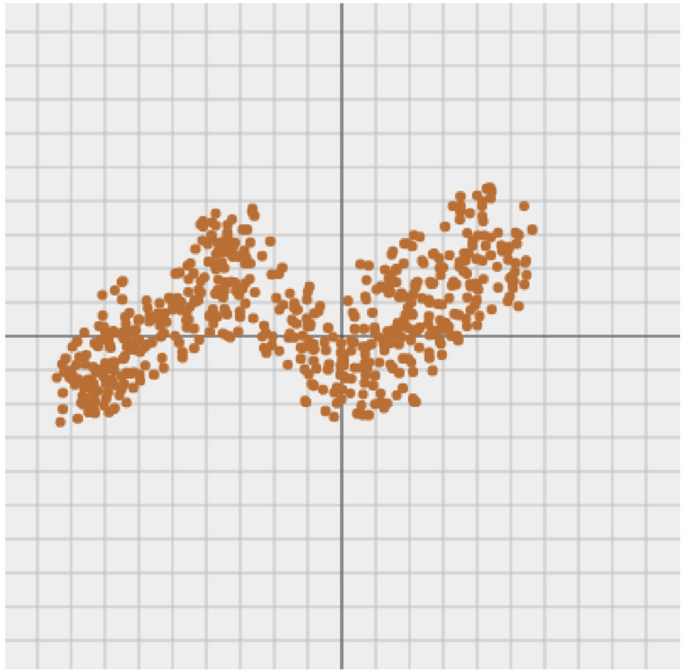
$$p(\mathbf{x} \mid \mathbf{z} = k) = \mathcal{N}(\mu_k, \Sigma_k)$$

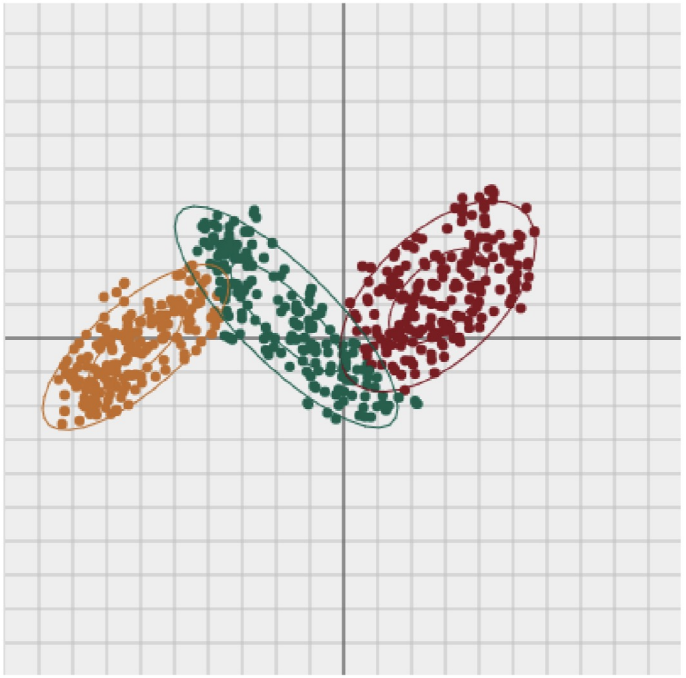


Generative process

1. Pick a mixture component by sampling z
2. Generate a data point by sampling from that Gaussian

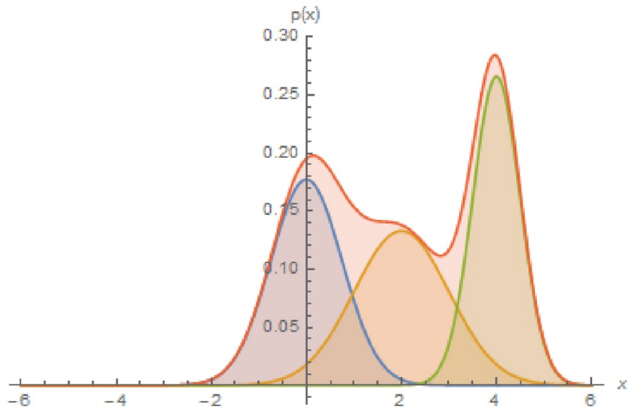
Mixture of Gaussians





Mixture of Gaussians

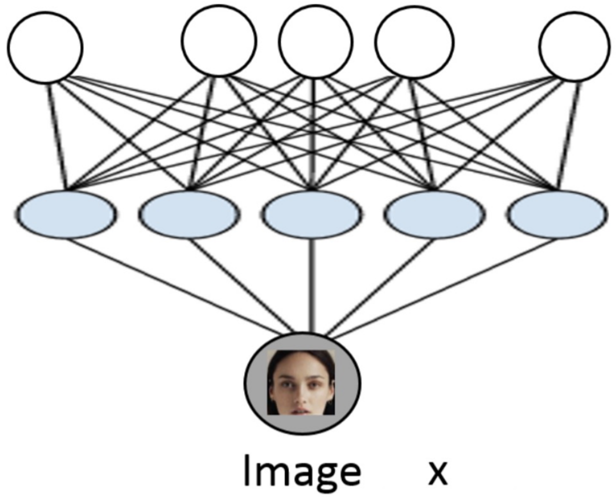
Combining simple models into more expressive ones



$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{z} = k) \underbrace{\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)}_{\text{component}}$$

can solve using expectation maximization

From GMMs to VAEs



- Put a prior on z $\mathbf{z} \sim \mathcal{N}(0, I)$

  $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$ where μ_θ,Σ_θ are neural networks
- Hope that after training, z will correspond to meaningful latent factors of variation - useful features for unsupervised representation learning
- Even though p(xlz) is simple, marginal p(x) is much richer/complex/flexible
- Given a new image x, features can be extracted via p(zlx): natural for unsupervised learning tasks (clustering, representation learning, etc.)

Learning parameters of VAEs

- Learning parameters of VAE: we have a joint distribution $p(\mathbf{X}, \mathbf{Z}; \theta)$
- We have a dataset **D** where for each datapoint the **x** variables are observed (e.g. images, text) and the variables **z** are not observed (latent variables)
- We can try maximum likelihood estimation:

$$\log \prod_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$$

Need cheaper approximations to optimize for VAE parameters

intractable :-(
- if z binary with 30 dimensions, need sum 2^30 terms
- if z continuous, integral is hard

Evidence Lower Bound

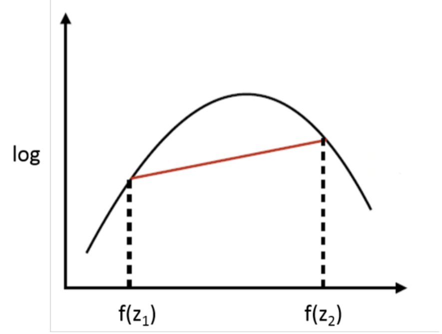
- Log-likelihood function with partially observed latent variables is hard to compute:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

q(**z**) should be a simple distribution

- Use Jensen's inequality for concave functions: $\log(px + (1-p)x') \geq p\log(x) + (1-p)\log(x')$.

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[f(\mathbf{z})]\right) = \log\left(\sum_{\mathbf{z}} q(\mathbf{z}) f(\mathbf{z})\right) \geq \sum_{\mathbf{z}} q(\mathbf{z}) \log f(\mathbf{z})$$



Evidence Lower Bound

- Log-likelihood function with partially observed latent variables is hard to compute:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

q(**z**) should be a simple distribution

- Use Jensen's inequality for concave functions: $\log(px + (1-p)x') \geq p\log(x) + (1-p)\log(x')$.

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[f(\mathbf{z})]\right) = \log\left(\sum_{\mathbf{z}} q(\mathbf{z})f(\mathbf{z})\right) \geq \sum_{\mathbf{z}} q(\mathbf{z})\log f(\mathbf{z})$$

Choosing $f(\mathbf{z}) = \frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}$

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right) \geq \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right]$$

Evidence Lower Bound (ELBO)

Evidence Lower Bound

- ELBO holds for any probability distribution q(**z**) over latent variables:

$$\begin{aligned}\log p(\mathbf{x};\theta) &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right) \\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{\text{Entropy } H(q) \text{ of } q} \\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)\end{aligned}$$

- Equality holds if q(z) = p(z|x):

$$\log p(\mathbf{x};\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z},\mathbf{x};\theta) + H(q)$$

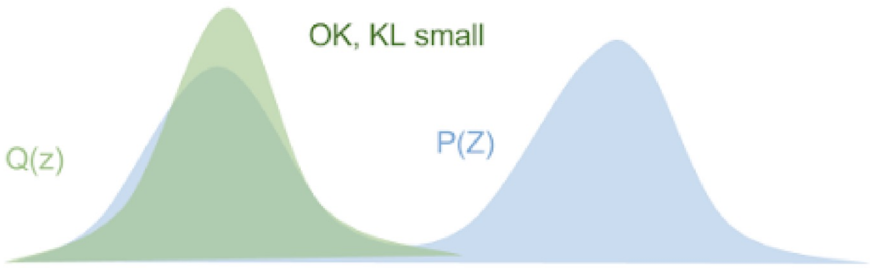
- We want to choose q(z) to be as close to p(z|x) as possible, while being easy to compute

KL Divergence

- The KL divergence for variational inference is:

$$\mathbf{D}_{KL}(q(z)||p(z|x)) = \int q(z) \log \frac{q(z)}{p(z|x)} dz$$

- Intuitively, there are three cases
  a. If **q** is low then we don't care (because of the expectation).
  b. If **q** is high and **p** is high then we are happy.
  c. If **q** is high and **p** is low then we pay a price.
- Note that p must be > 0 wherever q > 0



Evidence Lower Bound

- Starting from the KL divergence:

$$D_{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x};\theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z},\mathbf{x};\theta) + \log p(\mathbf{x};\theta) - H(q) \geq 0$$

- Re-derive ELBO from KL divergence:

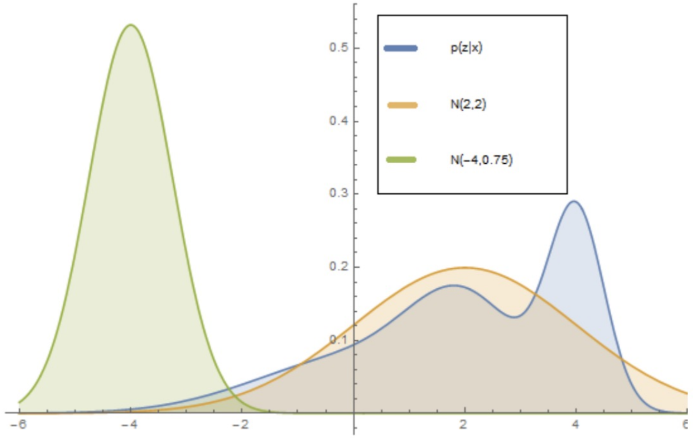
$$\log p(\mathbf{x};\theta) \geq \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z},\mathbf{x};\theta) + H(q)$$

- Equality holds if q = p(z|x) because KL(q||p) = 0:

$$\log p(\mathbf{x};\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z},\mathbf{x};\theta) + H(q)$$

- In general, $\log p(\mathbf{x};\theta) = \mathrm{ELBO} + D_{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x};\theta))$
- The closer the chosen q is to p(z|x), the closer the ELBO is to the true likelihood.

Variational Inference



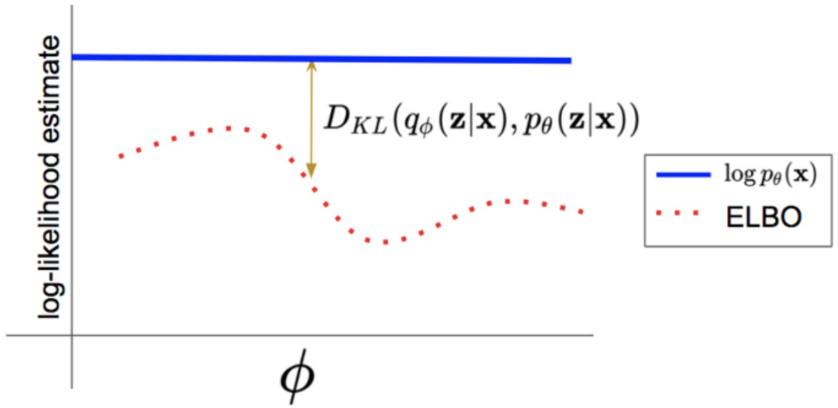
Suppose $q(\mathbf{z};\phi)$ is a (tractable) probability distribution over the hidden variables parameterized by ϕ (variational parameters)

- For example, a Gaussian with mean and covariance specified by ϕ

$$q(\mathbf{z};\phi) = \mathcal{N}(\phi_1, \phi_2)$$

- Variational inference: optimize variational parameters so that $q(\mathbf{z};\phi)$ is as close as possible to $p(\mathbf{z}|\mathbf{x};\theta)$ while being simple to compute
- E.g. in figure, posterior (in blue) is better approximated by orange Gaussian than green

Variational Inference



$$
\begin{aligned}
\log p(\mathbf{x};\theta) &\geq \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi)) = \underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} \\
&= \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}|\mathbf{x};\theta))
\end{aligned}
$$

- In practice how can we learn encoder parameters $p(\mathbf{z}|\mathbf{x};\theta)$ and variational (decoder) parameters jointly? $q(\mathbf{z};\phi)$

Learning parameters of VAEs

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

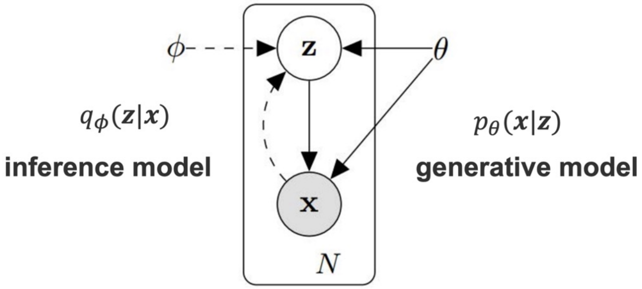
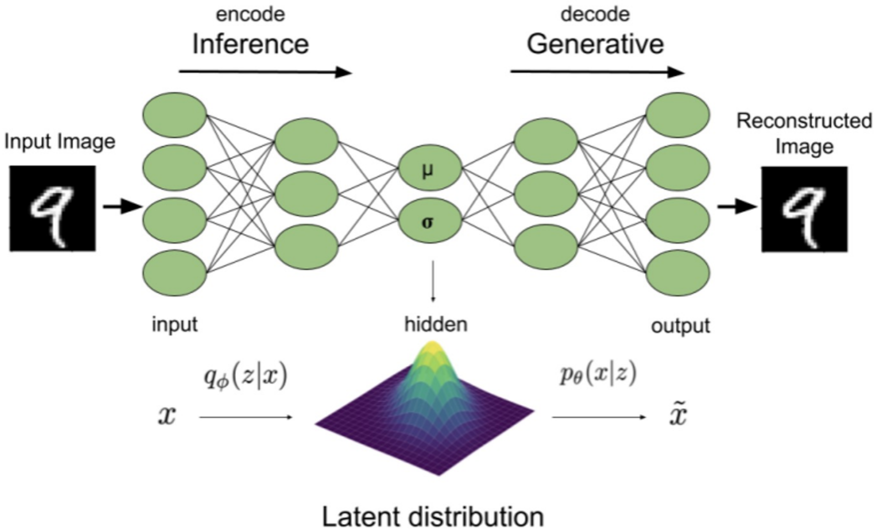


Figure courtesy: Kingma & Welling, 2014

What does the training objective $\mathcal{L}(\mathbf{x};\theta,\phi)$ do?

- First term encourages $\hat{\mathbf{x}} \approx \mathbf{x}^i$ ($\mathbf{x}^i$ likely under $p(\mathbf{x}|\hat{\mathbf{z}};\theta)$)
- Second term encourages $\hat{\mathbf{z}}$ to be likely under the prior $p(\mathbf{z})$

1. Take a data point $\mathbf{x}^i$
2. Map it to $\hat{\mathbf{z}}$ by sampling from $q_\phi(\mathbf{z}|\mathbf{x}^i)$ (*encoder*)
3. Reconstruct $\hat{\mathbf{x}}$ by sampling from $p(\mathbf{x}|\hat{\mathbf{z}};\theta)$ (*decoder*)



[Slides from Ermon and Grover]

Learning parameters of VAEs

$$\begin{aligned}\mathcal{L}(\mathbf{x};\theta,\phi) &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z},\mathbf{x};\theta) - \log q_\phi(\mathbf{z}|\mathbf{x}))] \\ &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z},\mathbf{x};\theta) - \log p(\mathbf{z}) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}))] \\ &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z};\theta)] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\end{aligned}$$

- We need to compute the gradients $\underbrace{\nabla_\theta \mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{easy}}$ and $\nabla_\phi \mathcal{L}(\mathbf{x};\theta,\phi)$

$$\begin{aligned}\nabla_\theta \mathcal{L}(x;\theta,\phi) &= \nabla_\theta E_{q_\phi(z|x)}[\log p(x|z;\theta)] - D_{KL}(q_\phi(z|x)||p(z)) \\ &= \nabla_\theta E_{q_\phi(z|x)}[\log p(x|z;\theta)] \\ &= E_{q_\phi(z|x)}[\nabla_\theta \log p(x|z;\theta)] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log p(x|z_i;\theta)\end{aligned}$$

Learning parameters of VAEs

$$\begin{aligned}\mathcal{L}(\mathbf{x};\theta,\phi) &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z},\mathbf{x};\theta) - \log q_\phi(\mathbf{z}|\mathbf{x}))] \\ &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z},\mathbf{x};\theta) - \log p(\mathbf{z}) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}))] \\ &= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z};\theta)] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\end{aligned}$$

- We need to compute the gradients $\underbrace{\nabla_\theta\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{easy}}$ and $\underbrace{\nabla_\phi\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{tricky}}$

- Expectations also depend on

$$\nabla_\phi\mathcal{L}(x;\theta,\phi) = \nabla_\phi E_{q_\phi(z|x)}[\log p(x|z;\theta)] - D_{KL}(q_\phi(z|x)||p(z))$$

Reparameterization Trick

- Want to compute a gradient with respect to ϕ of

$$E_{q(\mathbf{z};\phi)}[r(\mathbf{z})] = \int q(\mathbf{z};\phi) r(\mathbf{z}) d\mathbf{z}$$

where $\mathbf{z}$ is now **continuous**

- Suppose $q(\mathbf{z};\phi) = \mathcal{N}(\mu, \sigma^2 I)$ is Gaussian with parameters $\phi = (\mu, \sigma)$. These are equivalent ways of sampling:
  - Sample $\mathbf{z} \sim q_\phi(\mathbf{z})$
  - Sample $\epsilon \sim \mathcal{N}(0, I)$, $\mathbf{z} = \mu + \sigma\epsilon = g(\epsilon;\phi)$
- Using this equivalence we compute the expectation in two ways:

$$E_{\mathbf{z}\sim q(\mathbf{z};\phi)}[r(\mathbf{z})] = E_{\epsilon\sim\mathcal{N}(0,I)}[r(g(\epsilon;\phi))] = \int p(\epsilon) r(\mu + \sigma\epsilon) d\epsilon$$

$$\nabla_\phi E_{q(\mathbf{z};\phi)}[r(\mathbf{z})] = \nabla_\phi E_\epsilon[r(g(\epsilon;\phi))] = E_\epsilon[\nabla_\phi r(g(\epsilon;\phi))]$$

- Easy to estimate via Monte Carlo if r and g are differentiable w.r.t. ϕ and ϵ is easy to sample from (backpropagation)
- $E_\epsilon[\nabla_\phi r(g(\epsilon;\phi))] \approx \frac{1}{k}\sum_k \nabla_\phi r(g(\epsilon^k;\phi))$ where $\epsilon^1, \cdots, \epsilon^k \sim \mathcal{N}(0, I)$.

[Slides from Ermon and Grover]

Reparameterization Trick

$$\nabla_\phi \mathcal{L}(x;\theta,\phi) = \nabla_\phi E_{q_\phi(z|x)}[\log p(x|z;\theta)] - D_{KL}(q_\phi(z|x)||p(z))$$

$$\begin{aligned}\nabla_\phi E_{q_\phi(z|x)}[\log p(x|z;\theta)] &= \nabla_\phi E_\epsilon[\log p(x|\mu+\sigma\epsilon;\theta)] \quad \textbf{reparameterize} \\ &= E_\epsilon[\nabla_\phi \log p(x|\mu+\sigma\epsilon;\theta)] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}[\nabla_\phi \log p(x|\mu+\sigma\epsilon_i;\theta)]\end{aligned}$$