



Language  
Technologies  
Institute

Carnegie  
Mellon  
University

# Multimodal Machine Learning

## Lecture 9.1: Multimodal Generation

Paul Liang

*\* Co-lecturer: Louis-Philippe Morency. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

# Administrative Stuff

# Midterm Project Report (Due Sunday 10/29 at 8pm)

---

## Main goals:

1. Experiment with state-of-the-art approaches
  - Run on your own dataset state-of-the-art models
  - Teams of N should have at least N-1 baseline 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 Sunday 10/29 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 10/31 and Thursday 11/2)

---

## 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 10/31 and Thursday 11/2)

---

- 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 10/31 and Thursday 11/2)

---

- 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 10/31 and Thursday 11/2)

---

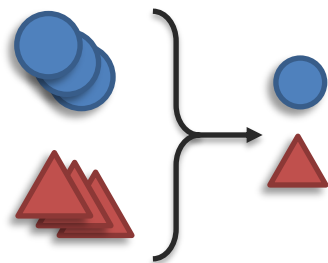
- Grading guidelines for presentations (4 points)
  - Quality of the slides (incl. images, videos and clear explanations)
  - Good motivation and explanation of the problem
  - Future research ideas (describe their future research directions)
  - Presentations skills (incl. explanations, voice and body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



# Challenge 4: Generation

# Generation

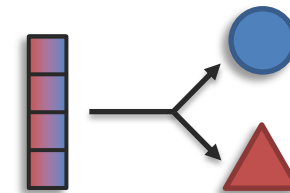
**Definition:** Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



Reduction



Maintenance



Expansion



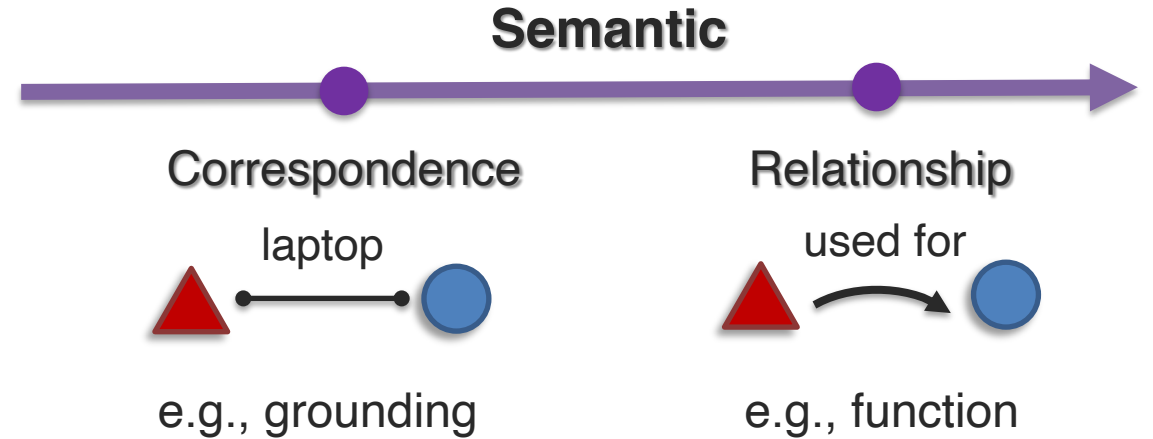
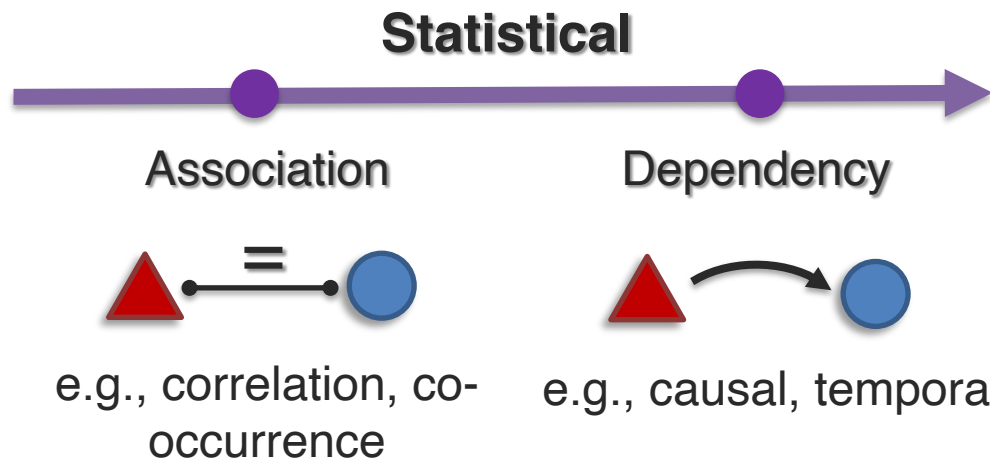
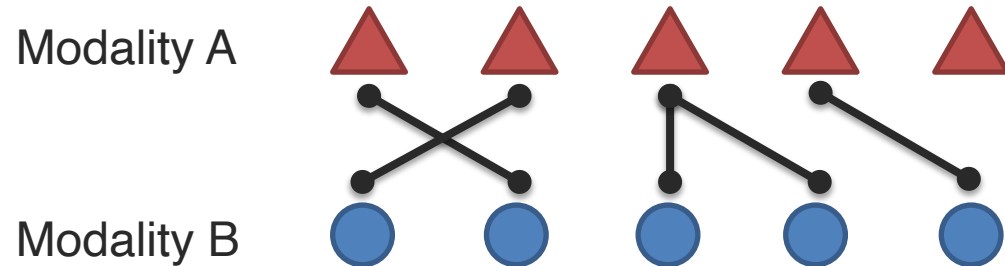
**Information:**  
(content)

# Dimension 1: Information Content

How modality interconnections change across multimodal inputs and generated outputs.

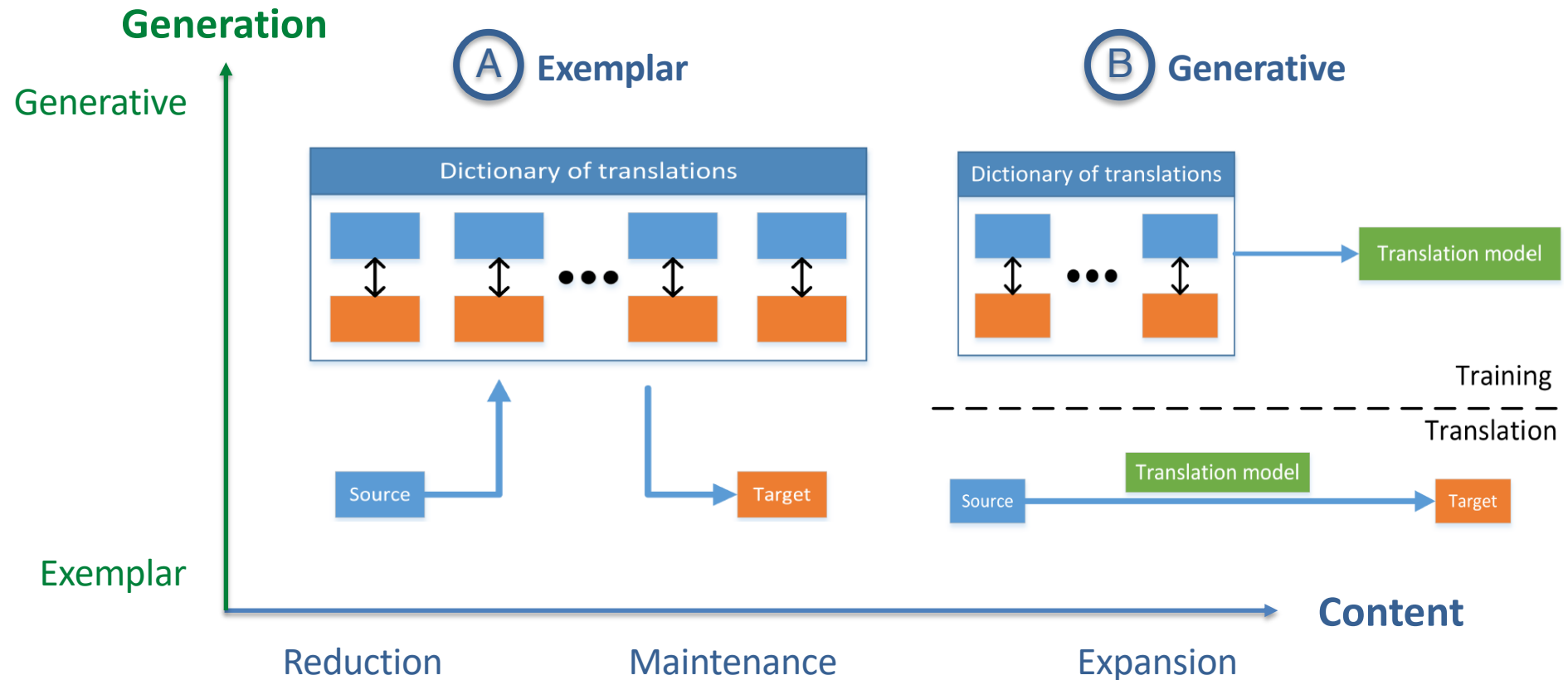
## ① Modality connections

*Modalities are often related and share commonality*



## Dimension 2: Generative Process

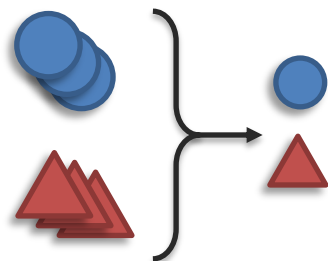
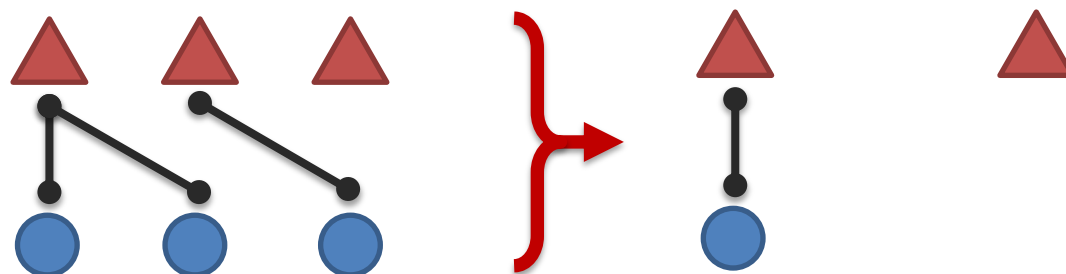
Generative process to respect modality heterogeneity and decode multimodal data.



# Dimension 1: Information Content

---

How modality interconnections change across multimodal inputs and generated outputs.



Reduction



Information:  
(content)

## Sub-challenge 4a: 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 .

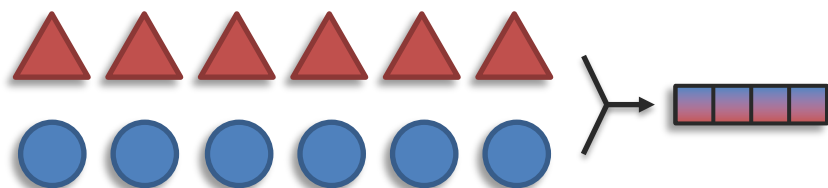
# Sub-challenge 4a: Summarization

## Video summarization

### (A) Content

Fusion via  
**joint representation**

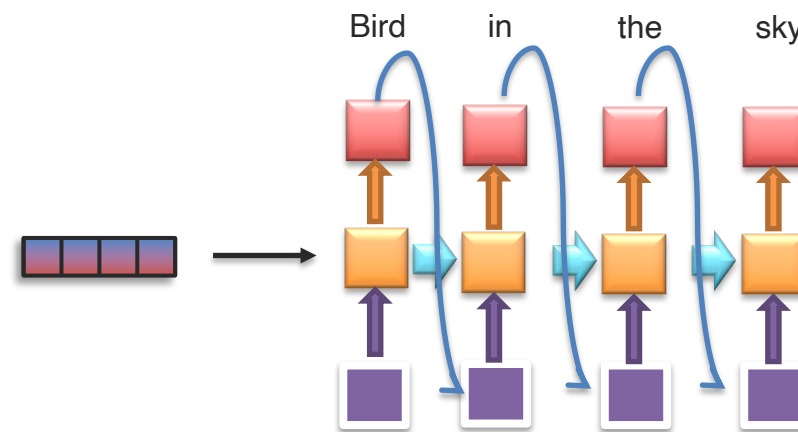
Capture **complementary**  
cross-modal interactions



### (B) Generation

**Generative**  $\approx$  **abstractive summarization**

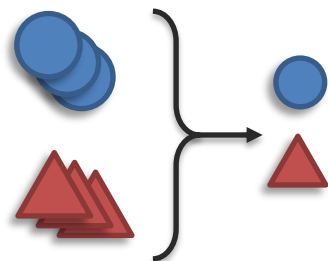
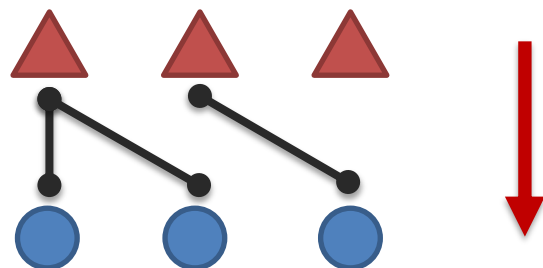
**Exemplar**  $\approx$  **extractive summarization**



# Dimension 1: Information Content

---

How modality interconnections change across multimodal inputs and generated outputs.



Reduction



Maintenance



Information:  
(content)



## Sub-challenge 4b: 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*

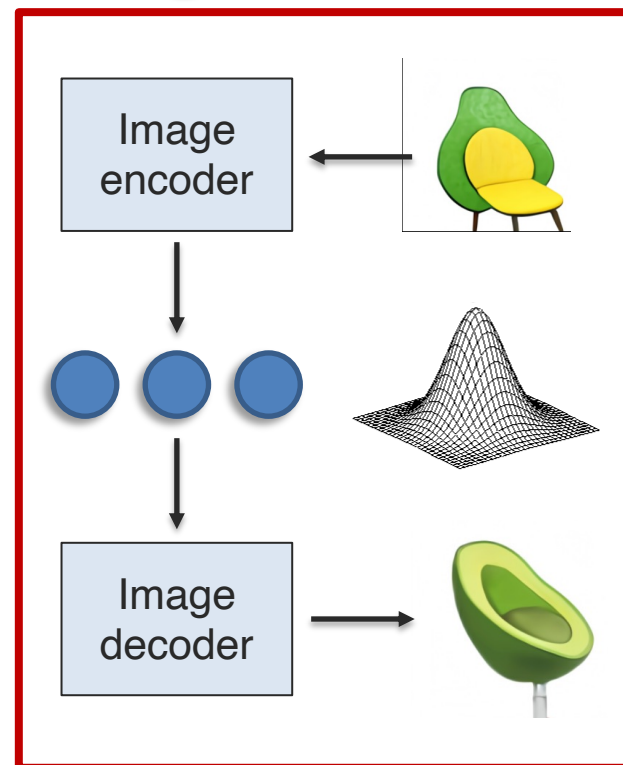


[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

# Sub-challenge 4b: Translation

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

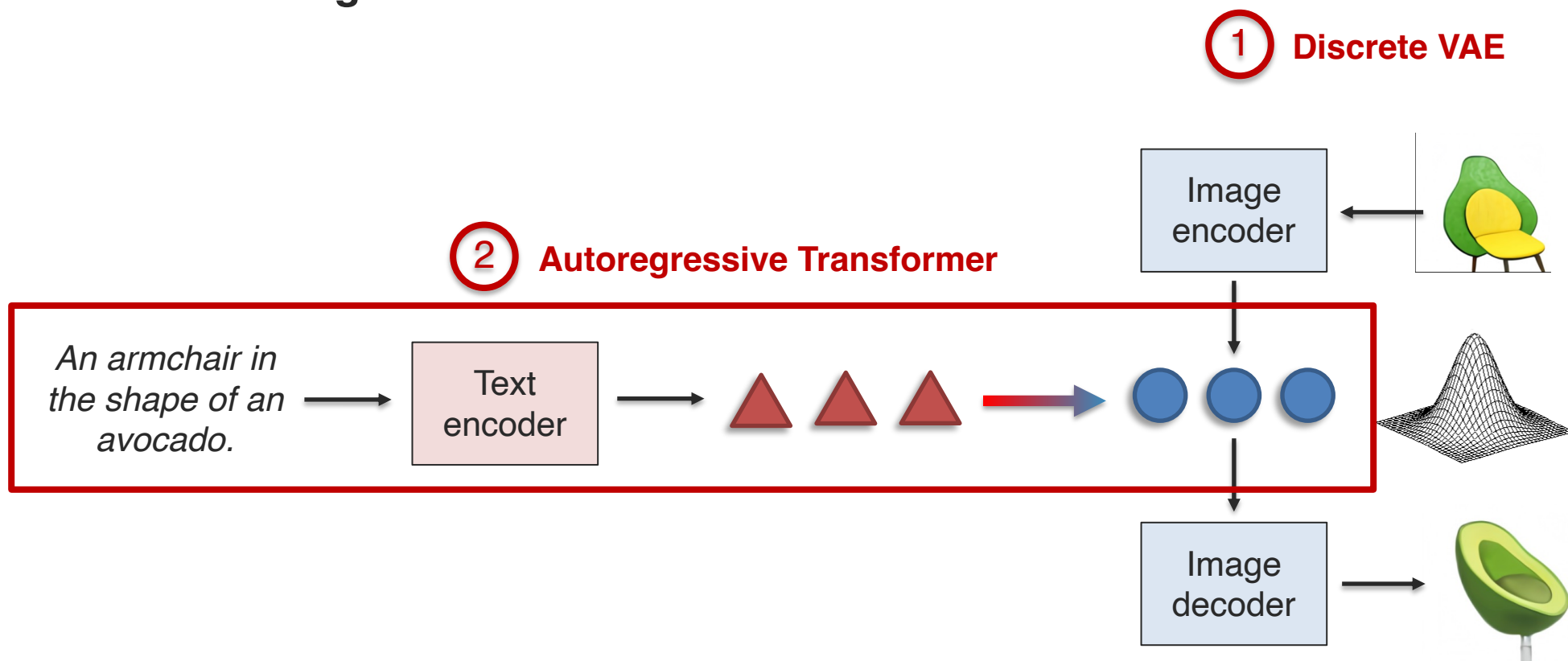
### ① Discrete VAE



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

# Sub-challenge 4a: Translation

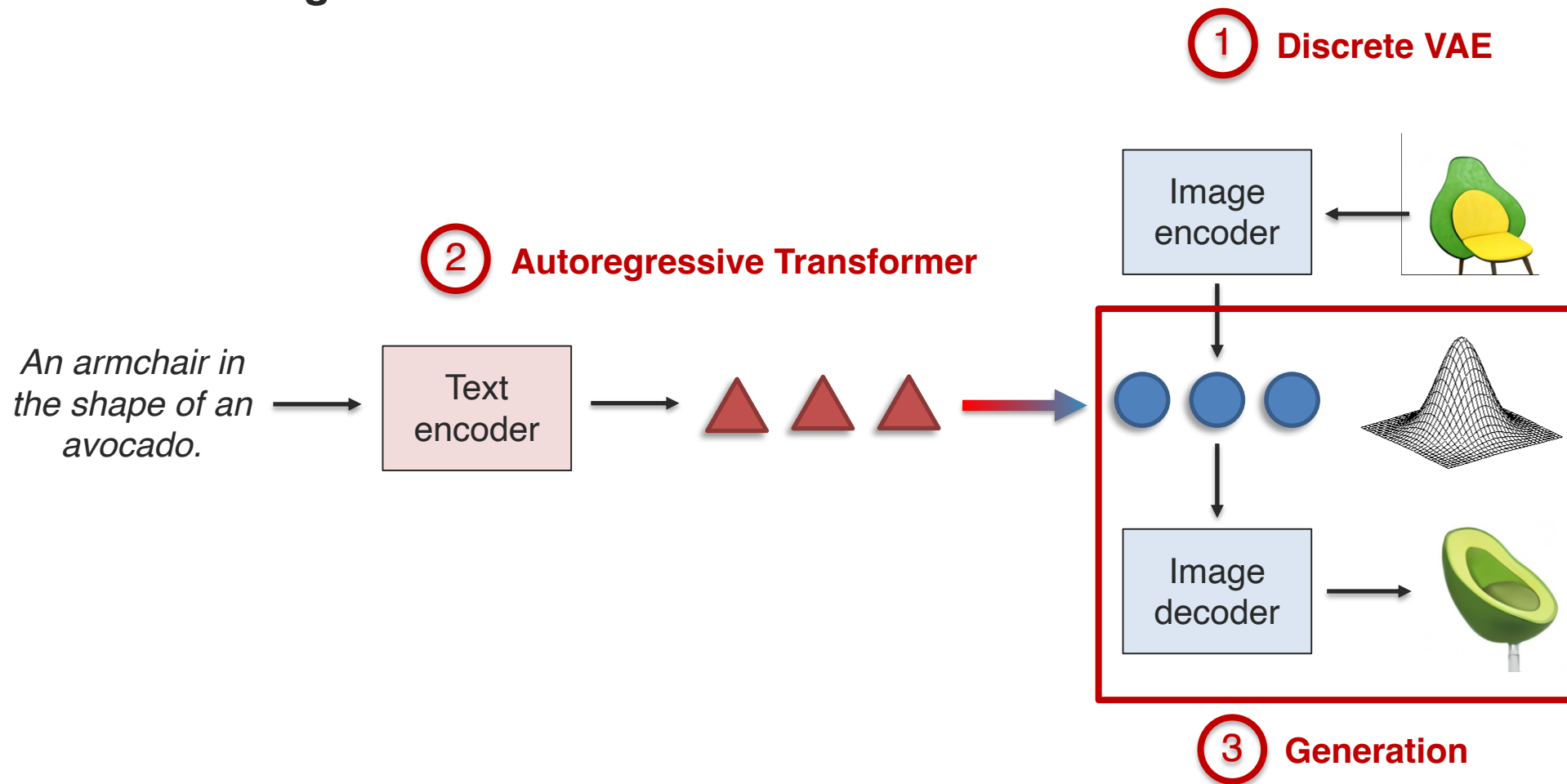
## DALL·E: Text-to-image translation at scale



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

# Sub-challenge 4a: Translation

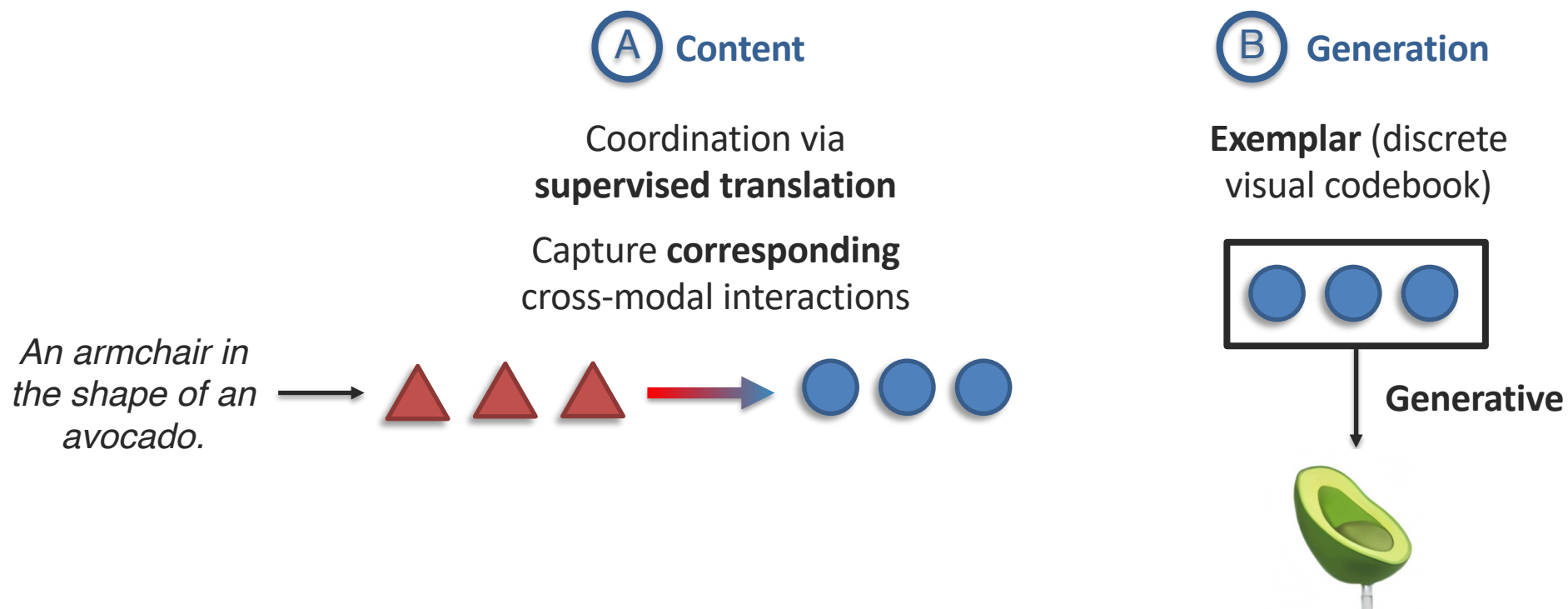
## DALL·E: Text-to-image translation at scale



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

# Sub-challenge 4a: Translation

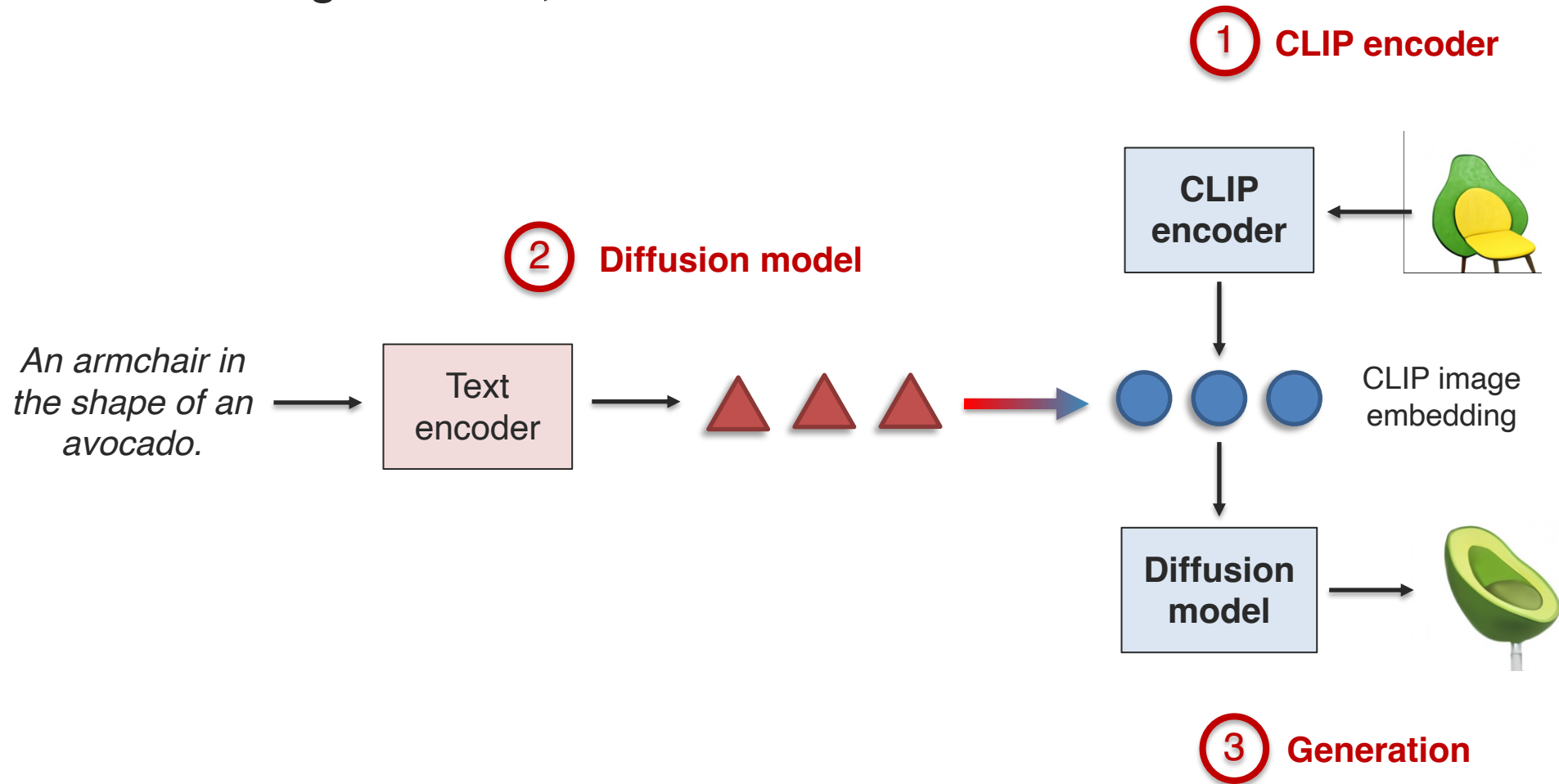
## DALL·E: Text-to-image translation at scale



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

# Sub-challenge 4a: Translation

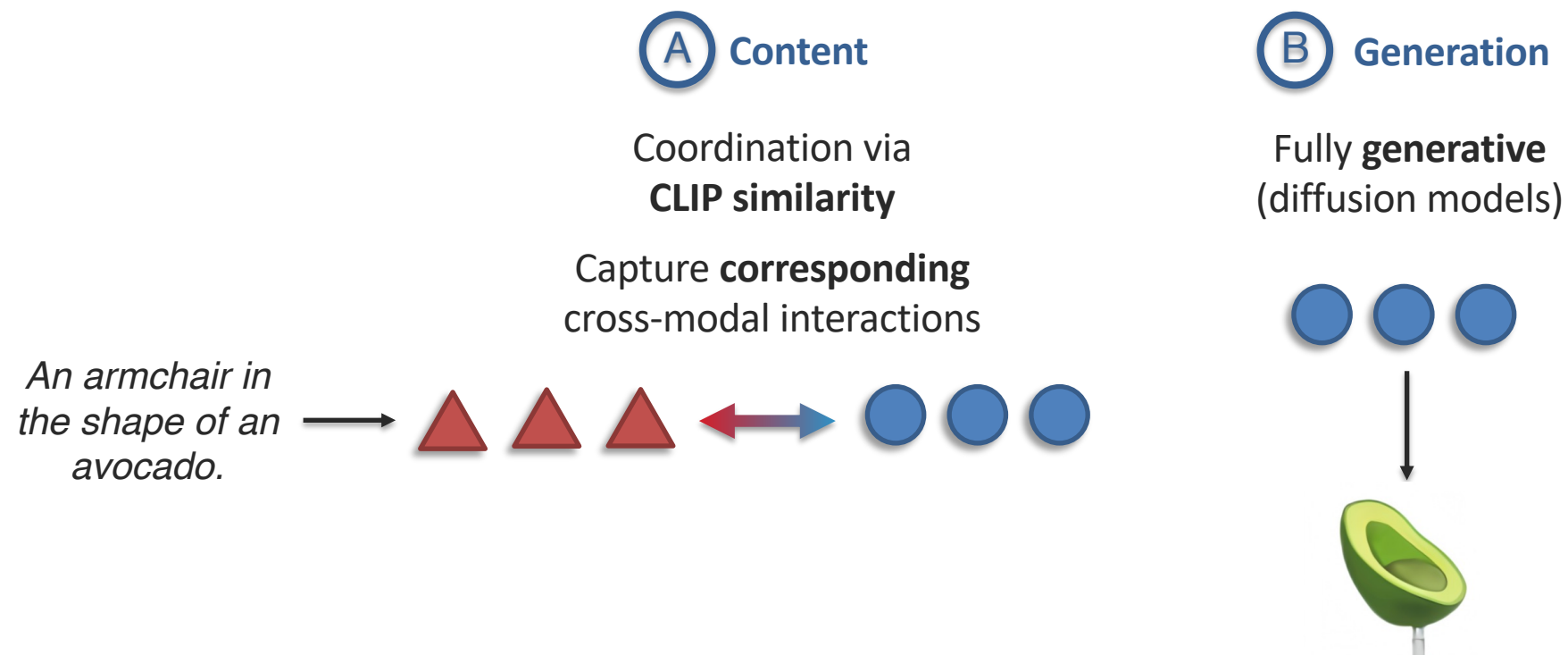
## DALL·E 2: Combining with CLIP, diffusion models



[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

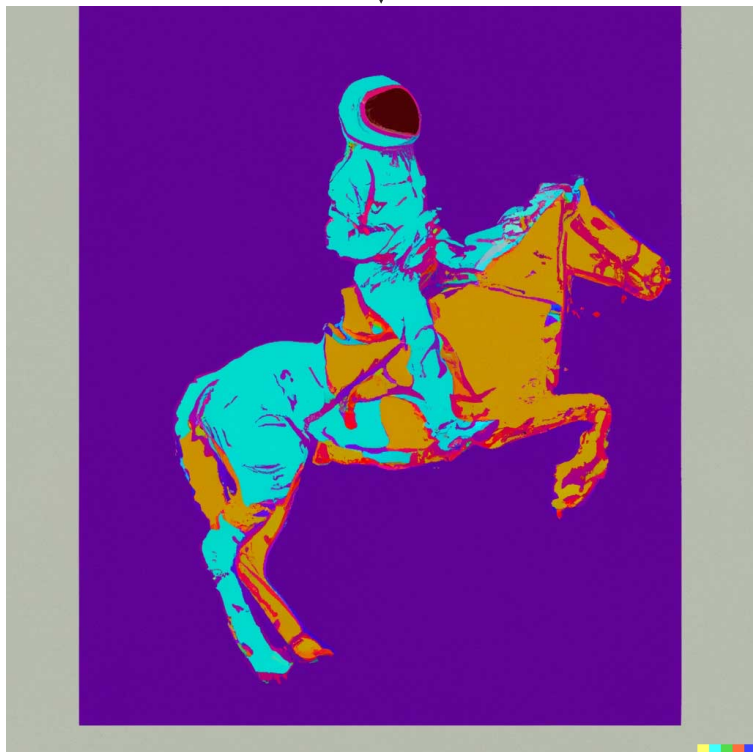


[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

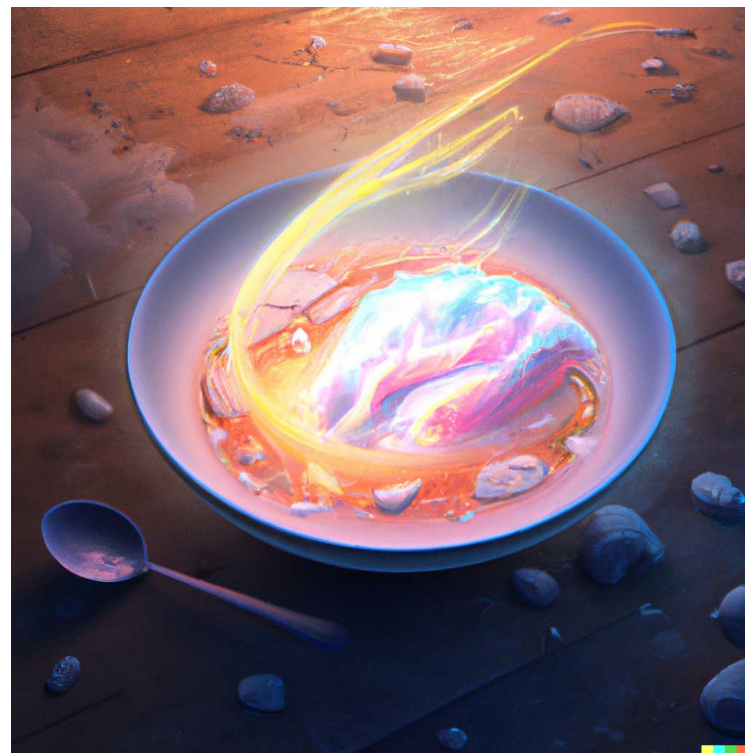
## CLIP + Image Generation

---

*An astronaut riding a horse in the style of Andy Warhol.*



*A bowl of soup that is a portal to another dimension as digital art*



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



# Virtual Humans

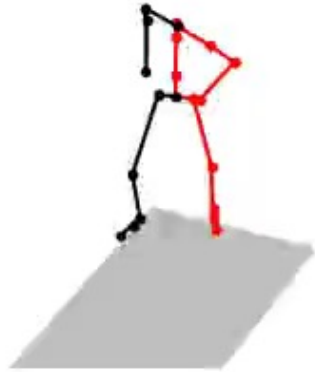
---



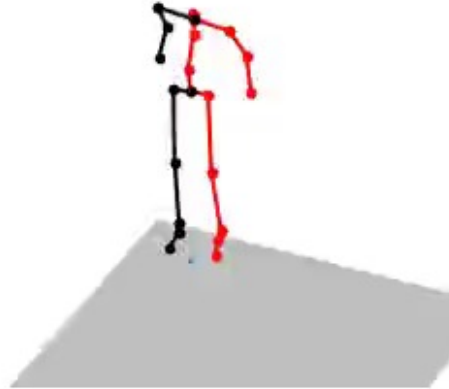
[Marsella et al., Virtual character performance from speech, SIGGRAPH, 2013]

# Language to Pose

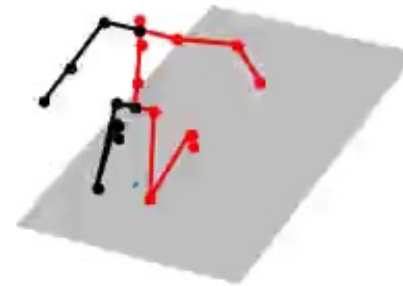
---



**a person jogs a few steps**



**A person steps forward then turns around and steps forwards again.**

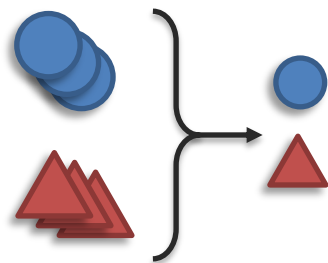
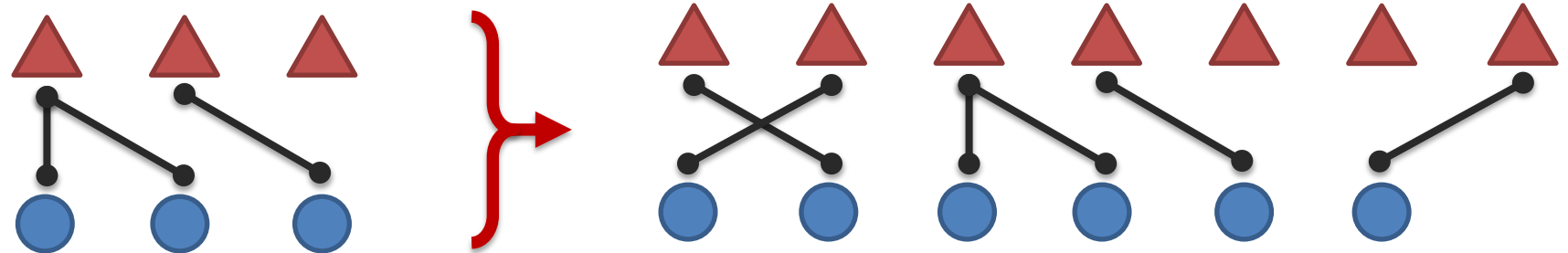


**A kneeling person raises their arms to the sides and stand up.**

[Ahuja & Morency. Language2Pose: Natural Language Grounded Pose Forecasting. Proceedings of 3DV Conference 2019]

# Dimension 1: Information Content

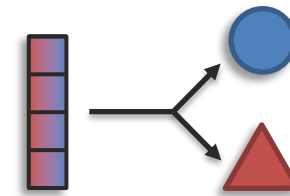
How modality interconnections change across multimodal inputs and generated outputs.



Reduction



Maintenance



Expansion

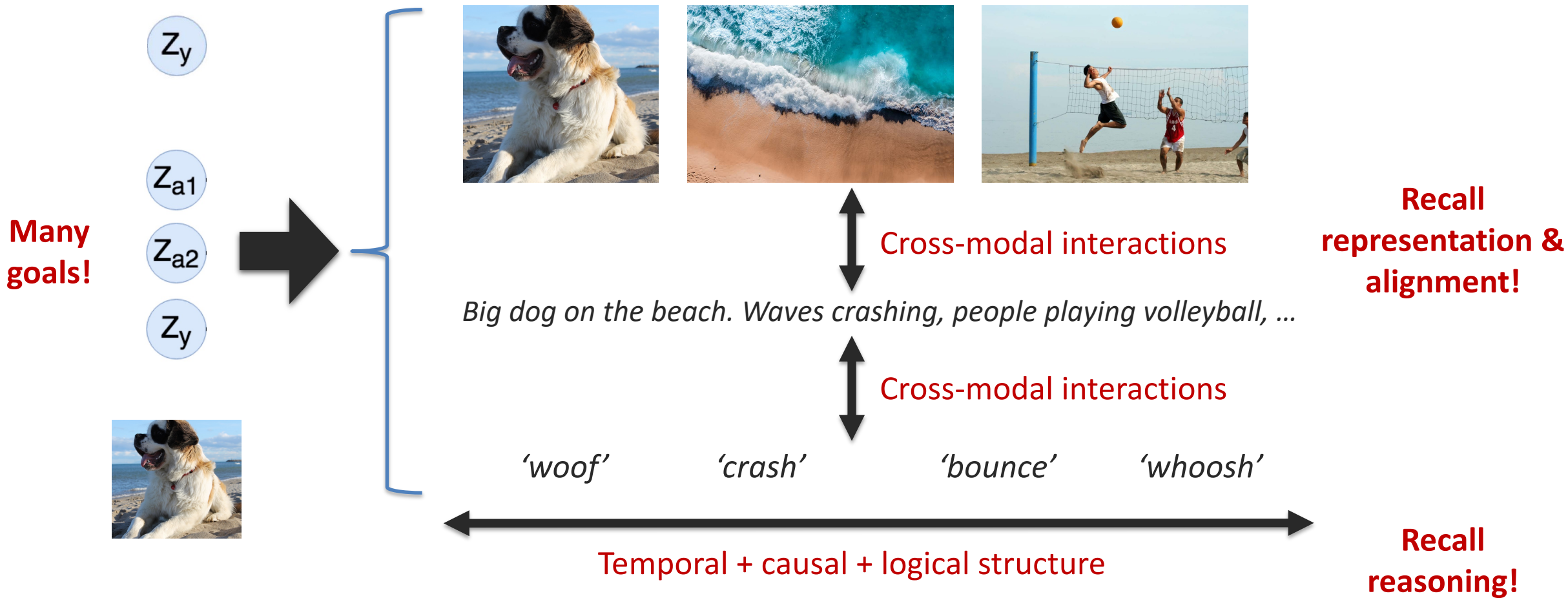


Information:  
(content)

# Sub-challenge 4c: Creation

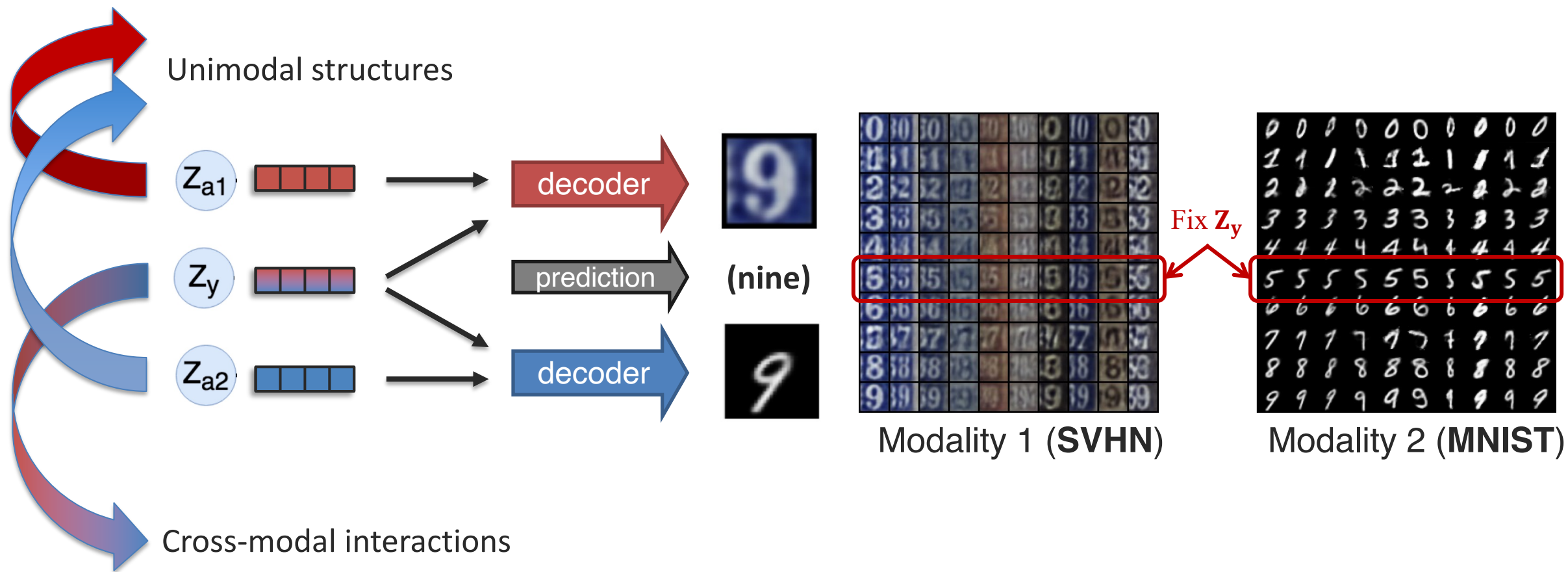


**Definition:** Simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities.



# Sub-challenge 4c: Creation

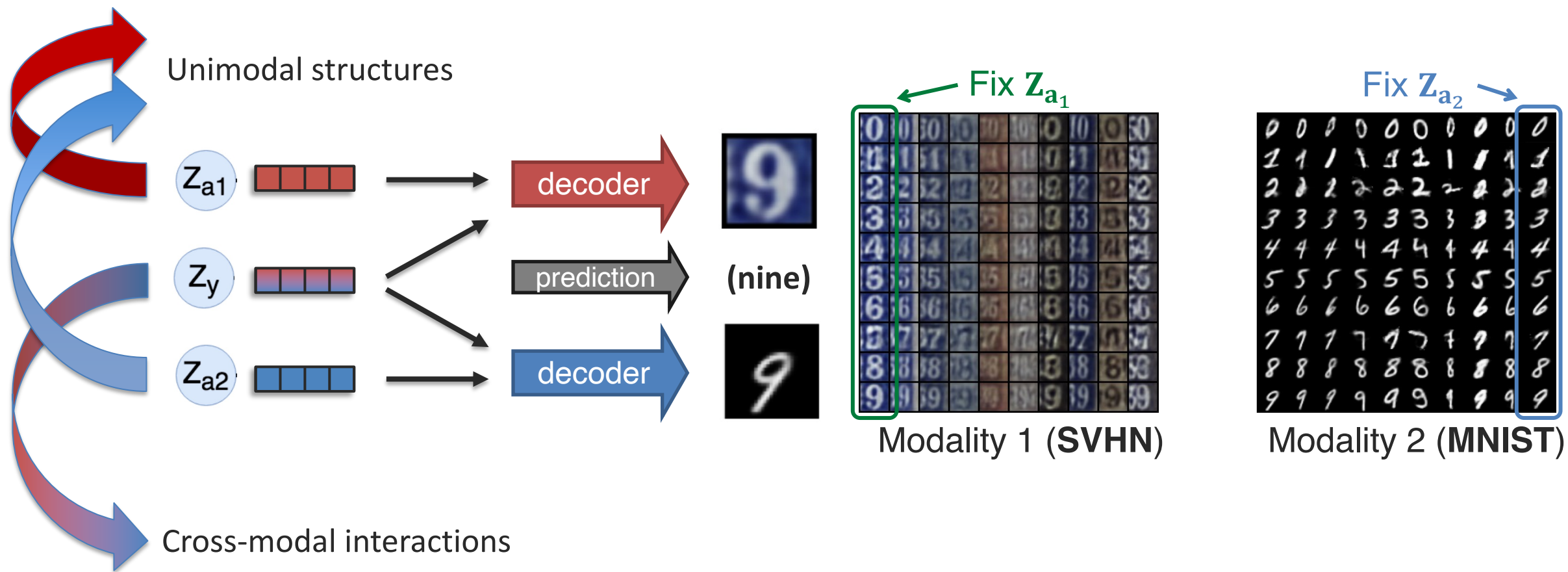
Some initial attempts: factorized generation



[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]

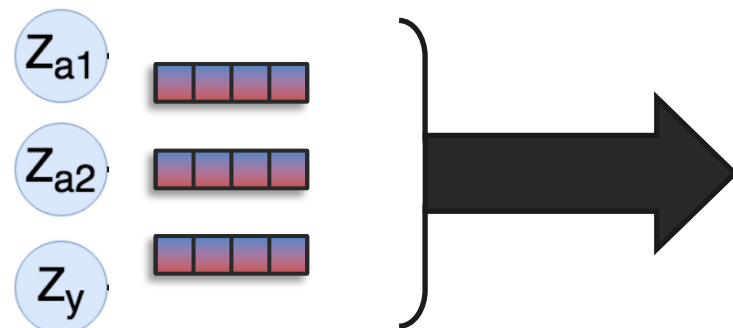
# Sub-challenge 4c: Creation

Some initial attempts: factorized generation

(A) Content

Factorized representation

Expanding **complementary**  
cross-modal interactions



(B) Generation

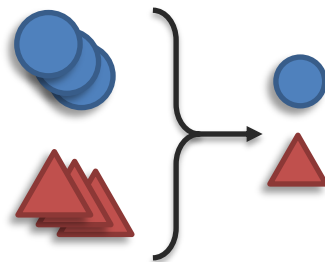
Generative model



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

# Preview: Generation

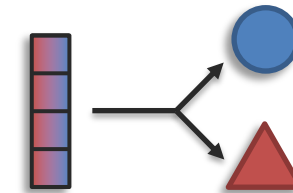
**Definition:** Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



Reduction



Maintenance



Expansion



**Information:**  
(content)

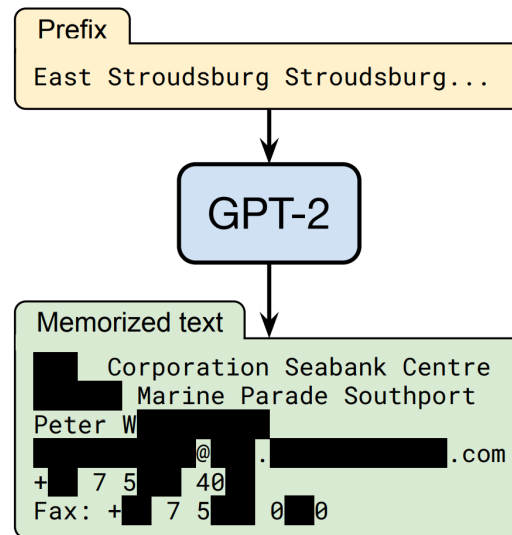
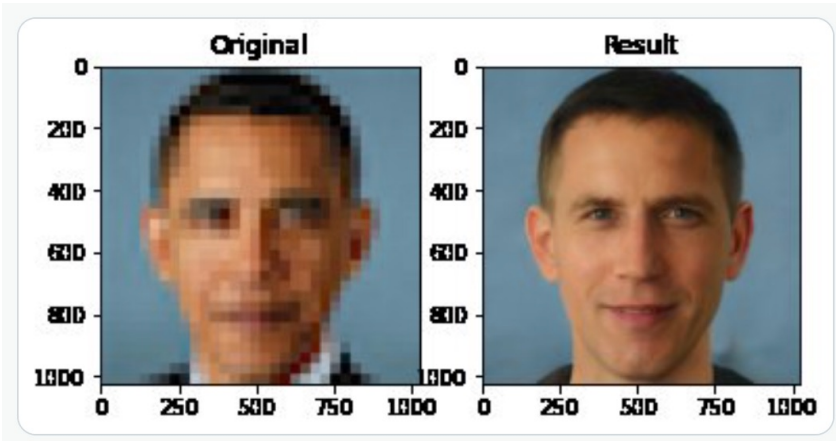


# Model Evaluation & Ethical Concerns

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
- 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.

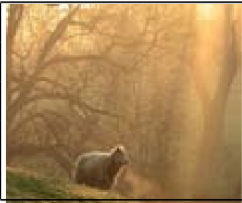


[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]

[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]

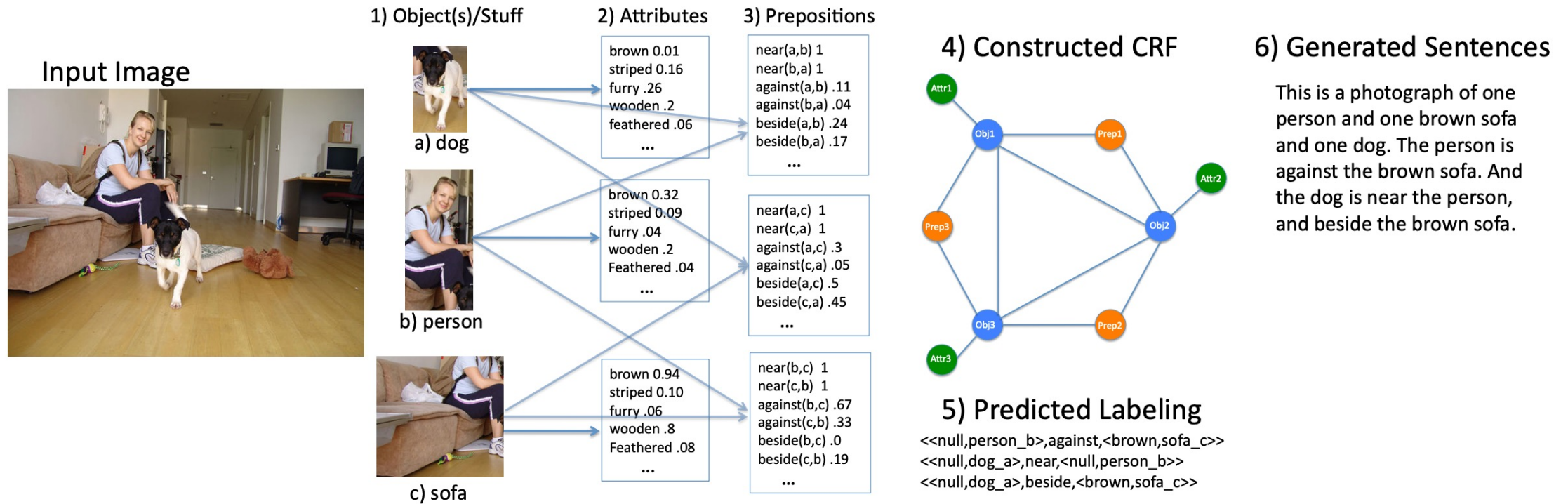
# Captioning as Summarization

- Every Picture Tells a Story [Farhadi 2010]
- Extract <object, action, scene> triplets, and retrieve sentences

	(pet, sleep, ground) (dog, sleep, ground) (animal, sleep, ground) (animal, stand, ground) (goat, stand, ground)	see something unexpected. Cow in the grassfield. Beautiful scenery surrounds a fluffy sheep. Dog herding sheep in open terrain. Cattle feeding at a trough.
	(furniture, place, furniture) (furniture, place, room) (furniture, place, home) (bottle, place, table) (display, place, table)	Refrigerator almost empty. Foods and utensils. Eatables in the refrigerator. The inside of a refrigerator apples, cottage cheese, tupperwares and lunch bags. Squash apenny white store with a hand statue, picnic tables in front of the building.
	(transportation, move, track) (bike, ride, track) (transportation, move, road) (pet, sleep, ground) (bike, ride, road)	A man stands next to a train on a cloudy day A backpacker stands beside a green train This is a picture of a man standing next to a green train There are two men standing on a rocky beach, smiling at the camera. This is a person laying down in the grass next to their bike in front of a strange white building.

# Captioning as Translation

- Baby Talk: Understanding and Generating Simple Image Descriptions
- Templated sentences that describe all detected objects, attributes, and relations



# Captioning as Generation

- Visual Storytelling [Huang et al. 2016]

Isolated Captions	DII	 A black frisbee is sitting on top of a roof.	 A man playing soccer outside of a white house with a red door.	 The boy is throwing a soccer ball by the red door.	 A soccer ball is over a roof by a frisbee in a rain gutter.	 Two balls and a frisbee are on top of a roof.
Sequential Captions	DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
Stories	SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

[https://www.microsoft.com/en-us/research/wp-content/uploads/2016/06/visionToLanguage2015\\_DataRelease-1.pdf](https://www.microsoft.com/en-us/research/wp-content/uploads/2016/06/visionToLanguage2015_DataRelease-1.pdf)

# Generation is Data-dependent

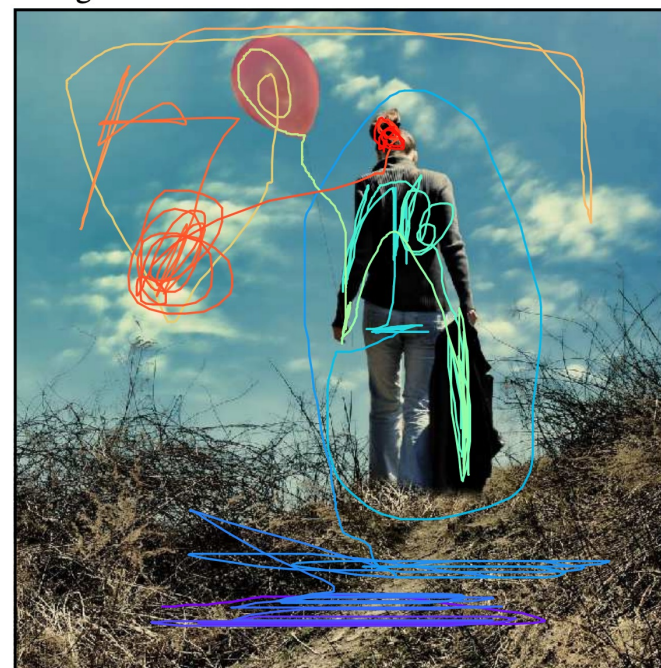
- What were people's goals when they wrote image descriptions?



1. A graying man in a suit is perplexed at a business meeting.
2. A businessman in a yellow tie gives a frustrated look.
3. A man in a yellow tie is rubbing the back of his neck.
4. A man with a yellow tie looks concerned.
5. Gray haired man in black suit and yellow tie working in a financial environment.

Flickr30k, <https://aclanthology.org/Q14-1006/>

Image and Trace:



Caption:

In the front portion of the picture we can see a dried grass area with dried twigs. There is a woman standing wearing light blue jeans and ash colour long sleeve length shirt. This woman is holding a black jacket in her hand. On the other hand she is holding a balloon which is peach in colour. On the top of the picture we see a clear blue sky with clouds. The hair colour of the woman is brownish.

Localized Narratives, <https://arxiv.org/pdf/1912.03098.pdf>

# More Complex Inputs: Video 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**

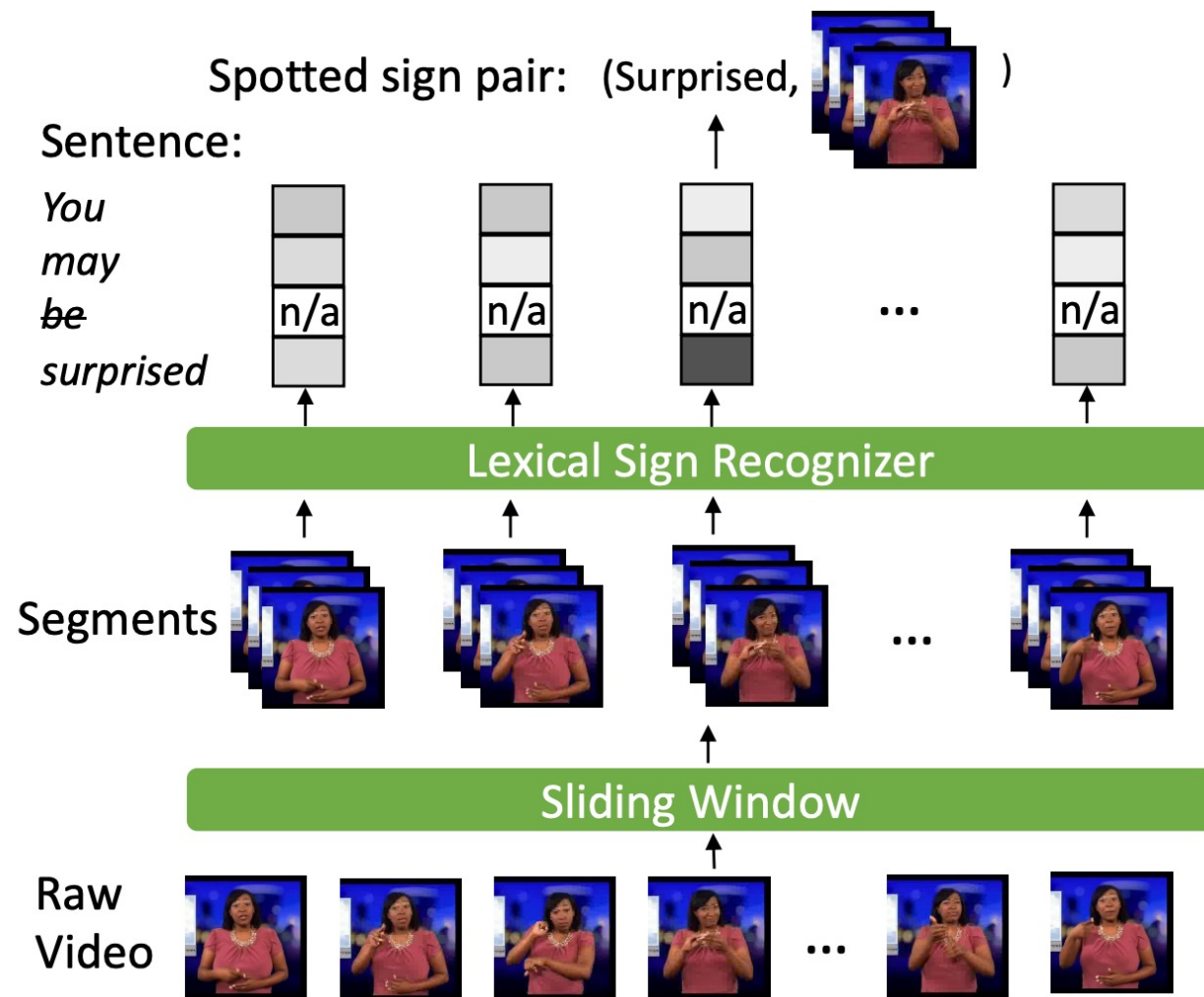
*Cuban breakfast*  
*Free cooking video*

(not present in text)

**Summary**

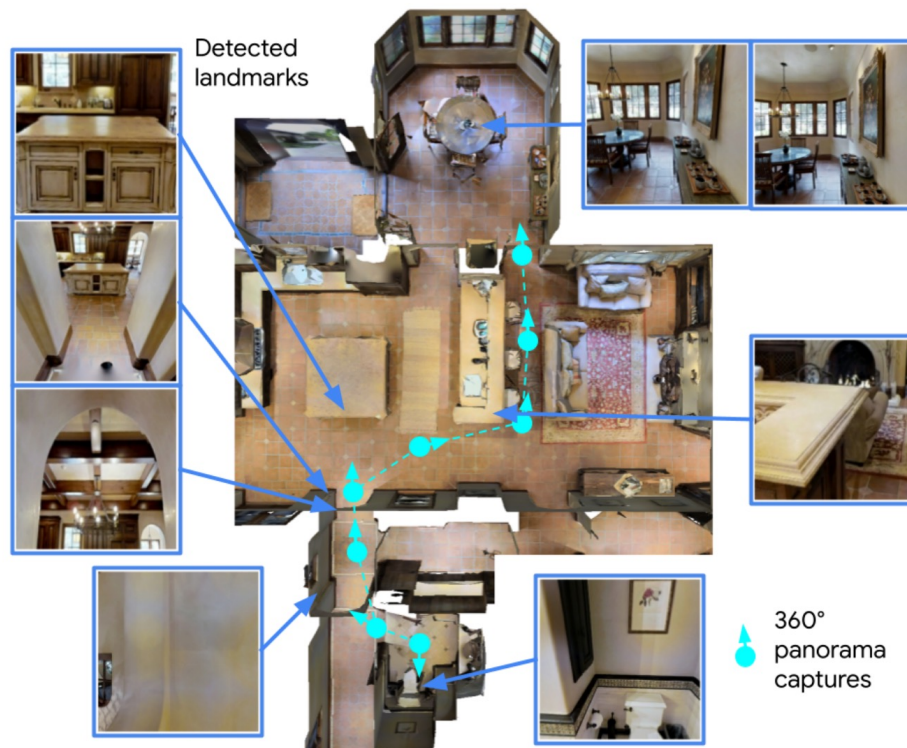
how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

# More Complex Inputs: Sign Language Translation



[Open-Domain Sign Language Translation Learned from Online Video, <https://arxiv.org/pdf/2205.12870.pdf>]

# More Complex Inputs: Instruction Generation



## Generated Instruction:

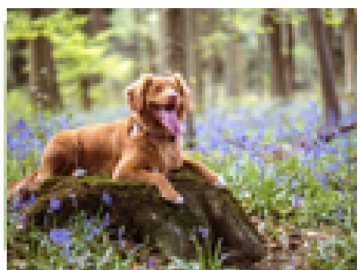
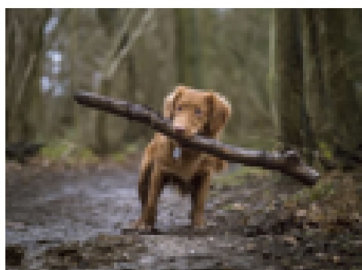
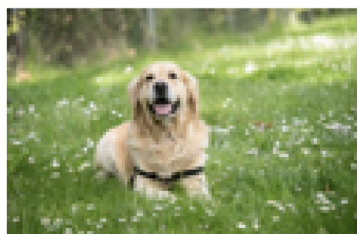
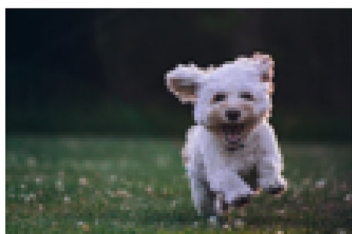
You are facing towards the commode. Turn right and exit the washroom. Turn right and walk straight till you reach the white cabinet in the front. There is an arch in the front. Enter inside the arch. Turn right and walk towards the sofa. Turn left and walk straight till you reach the arch in the front. There is a round table with four chairs towards your left side. You have reached your point.



# Generative Models

Learn to model  $p(\mathbf{x})$  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 ( $\mathbf{x}$ )	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  $p(x_2|x_1,c)$  - **style transfer**

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



---

From negative to positive

---

consistently slow .  
consistently good .  
consistently fast .

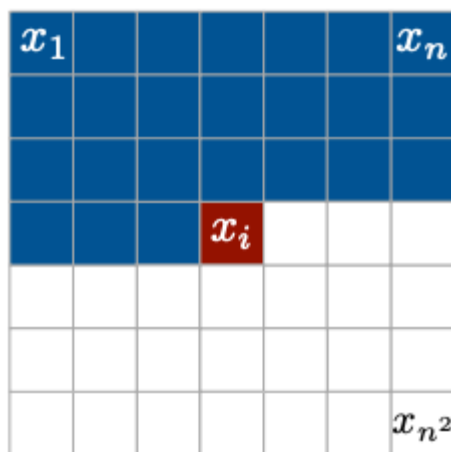
my goodness it was so gross .  
my husband 's steak was phenomenal .  
my goodness was so awesome .

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 .

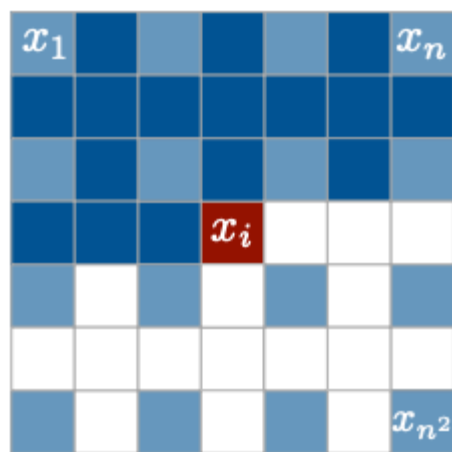
# Autoregressive Models

Autoregressive models

$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, \dots, x_{i-1})$$



Context



Multi-scale context



Figure 1. Image completions sampled from a PixelRNN.

# Autoregressive Models

---

Autoregressive language models

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1})$$

Input Prompt:

Recite the first law of robotics



Output:

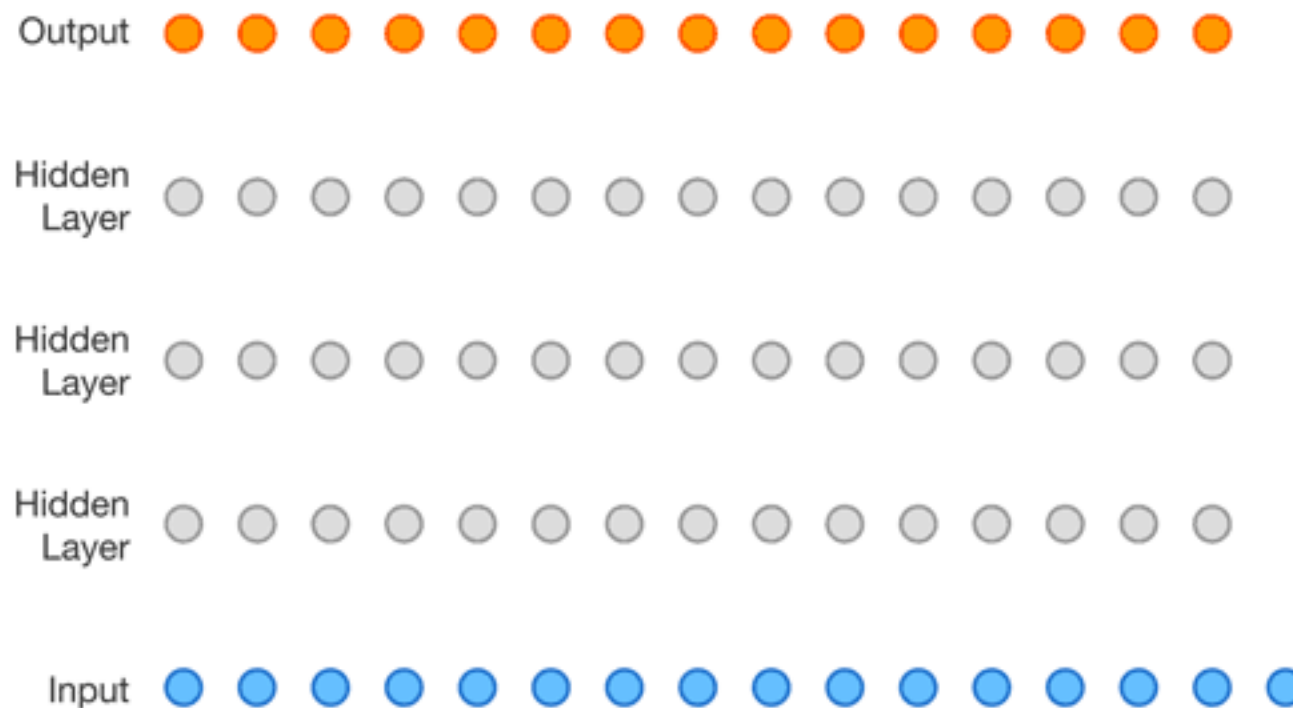
[Brown et al., Language Models are Few-shot Learners. NeurIPS 2020]

# Autoregressive Models

---

Autoregressive audio generation models

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1})$$



[van den Oord et al., WaveNet: A Generative Model for Raw Audio. ICML 2016]

# Conditioning Autoregressive Models

---

We typically want  $p(x|c)$  - **conditional generation**

- $c$  is a category (e.g. faces, outdoor scenes) from which we want to generate images
- $c$  is an image which we want to describe in natural language

We might also care about  $p(x_2|x_1,c)$  - **style transfer**

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



---

From negative to positive

---

consistently slow .  
consistently good .  
consistently fast .

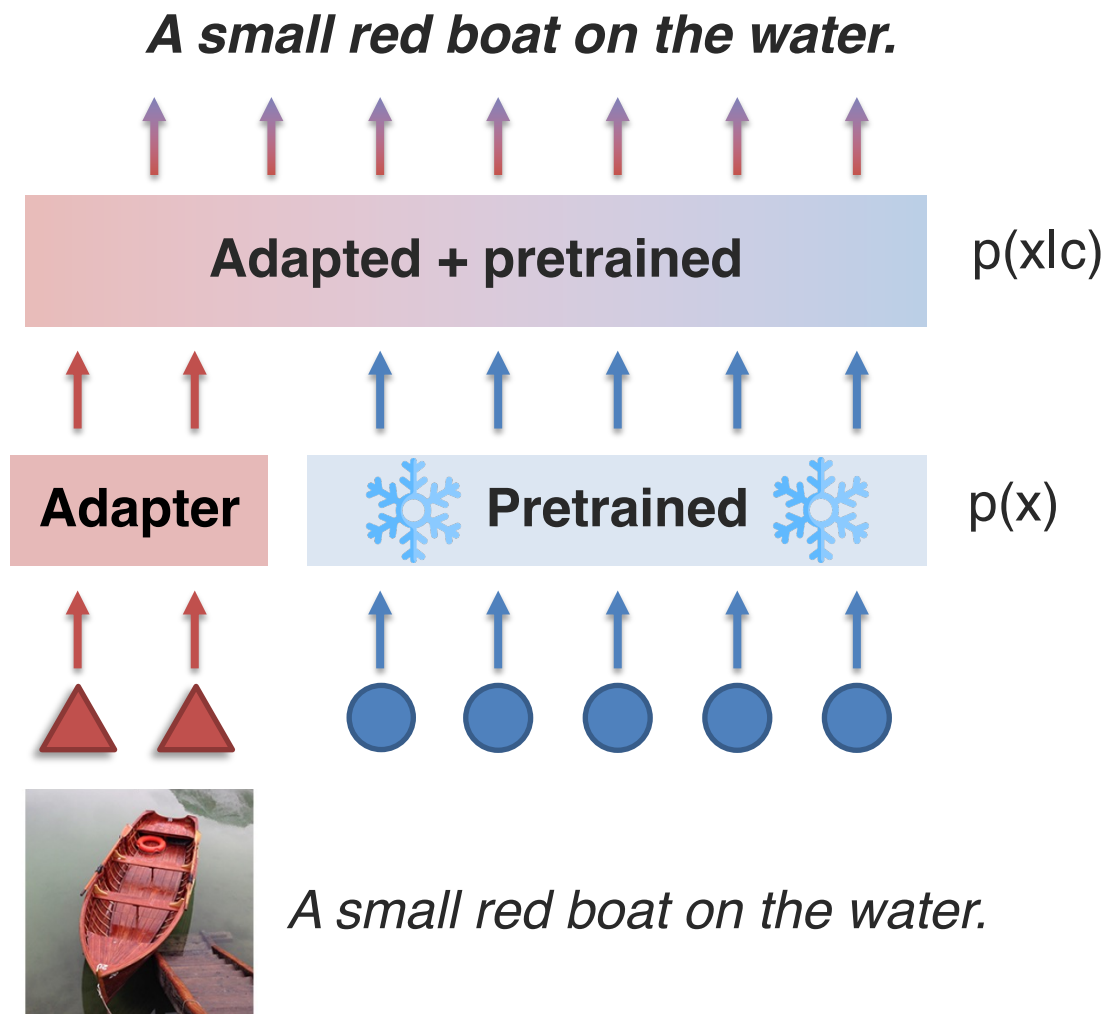
my goodness it was so gross .  
my husband 's steak was phenomenal .  
my goodness was so awesome .

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 .

# Conditioning Autoregressive Models

## Conditioning via prefix tuning

Modeling  $p(x|c)$ :

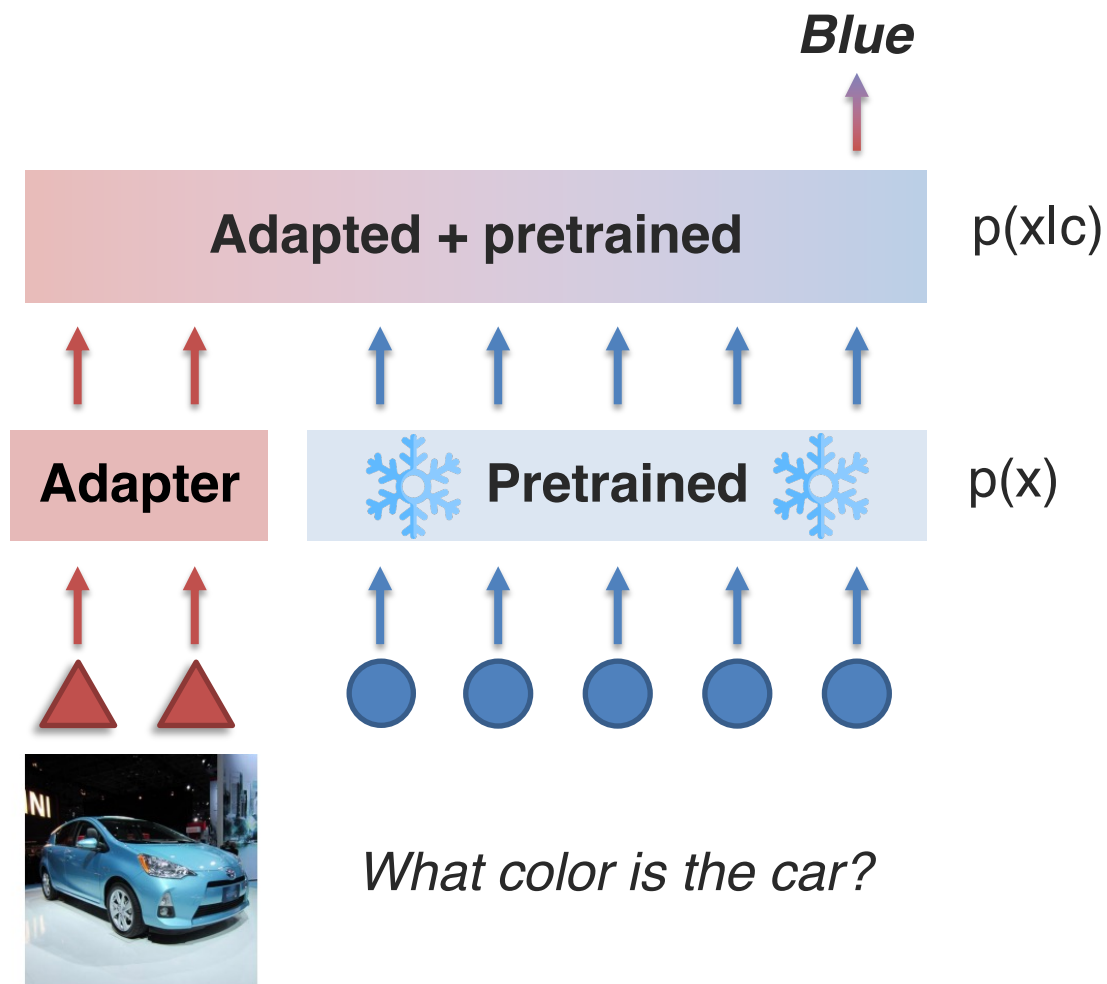


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

# Conditioning Autoregressive Models

## Conditioning via prefix tuning

0-shot VQA:



[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

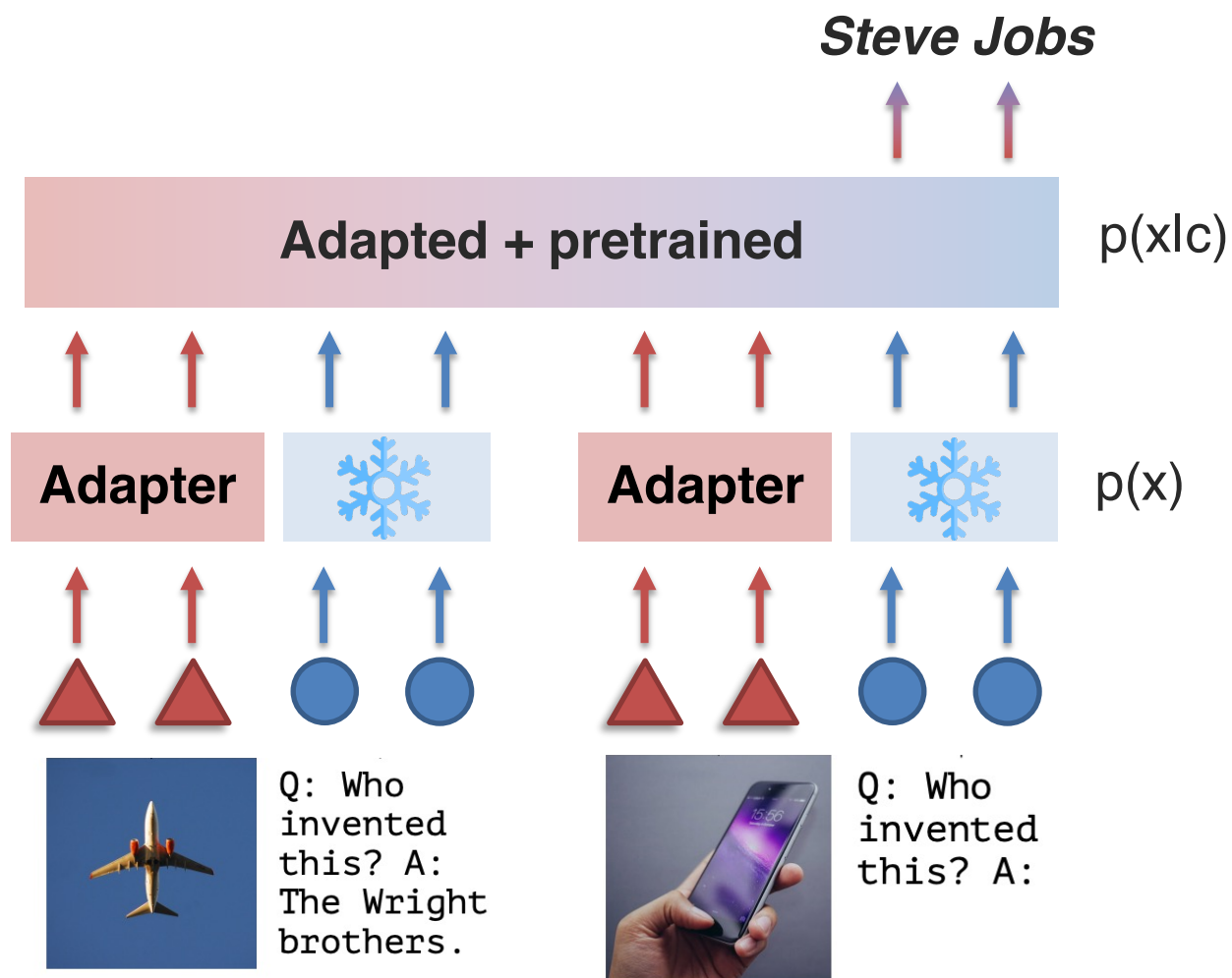


# Conditioning Autoregressive Models

## Conditioning via prefix tuning

1-shot outside knowledge VQA:

Recall reasoning  
– leverage implicit knowledge in LMs

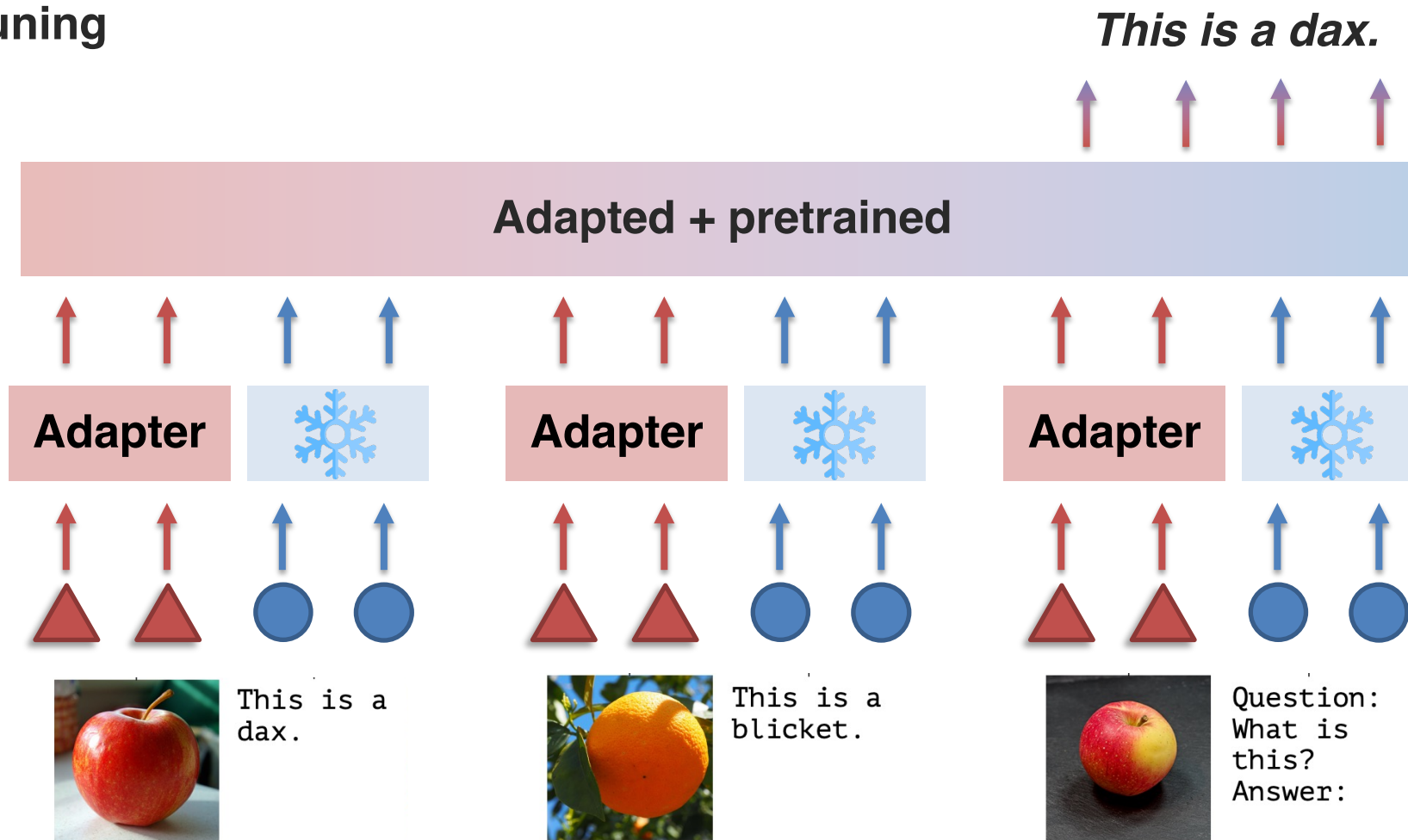


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

# Conditioning Autoregressive Models

## Conditioning via prefix tuning

Few-shot image classification:



[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

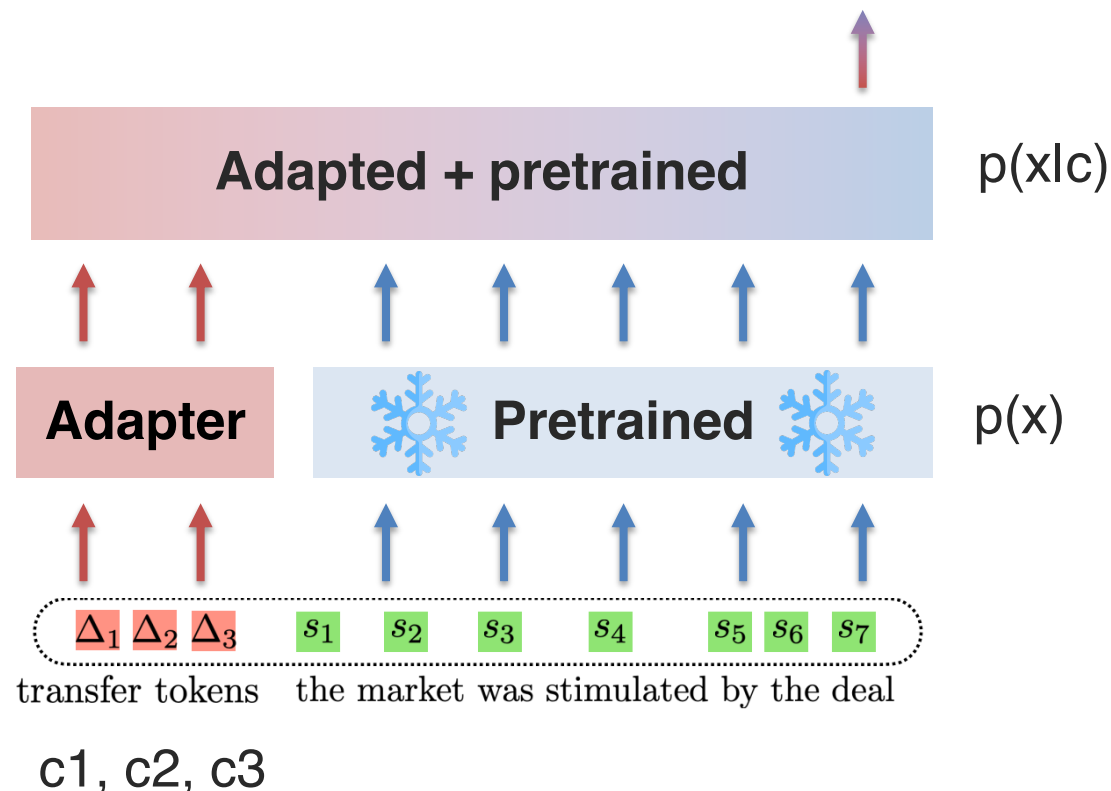
# Conditioning Autoregressive Models

## Conditioning via prefix tuning

*The deal greatly simulated the market.*

Multiclass style transfer:




Also enables compositions of  $c_1, c_2, c_3$ .



[Lyu et al., STYLEPTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer. NAACL 2021]

# Flamingo

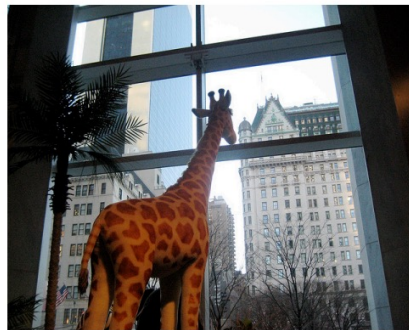
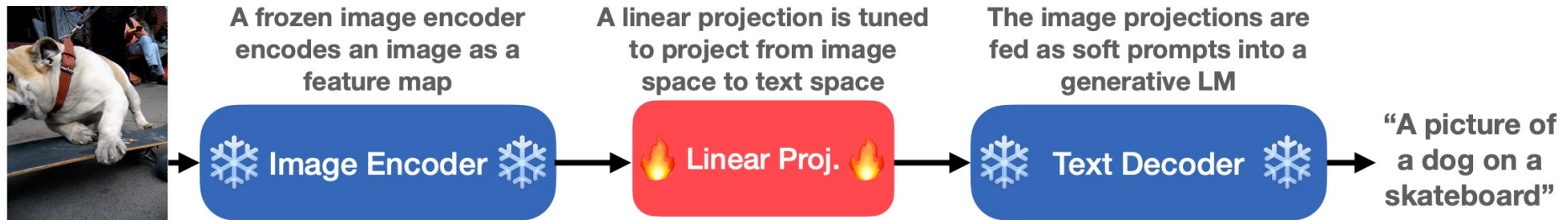
- Scales up Frozen with much bigger models (80B parameter LMs), cross-attention, and richer training data

 <p>This is a picture of two teddy bears on the moon.</p> <p>What are they doing?</p> <p>They are having a conversation.</p> <p>What object are they using?</p> <p>It looks like a computer.</p> <p>Is this surprising?</p> <p>Yes, it is surprising.</p> <p>Why is this picture surprising to you?</p> <p>I think it is surprising because teddy bears are not usually found on the moon.</p>	 <p>What is the common thing about these three images?</p> <p>They are all flamingos.</p> <p>What is the difference between these three images?</p> <p>The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.</p>	 <p>This is an apple with a sticker on it.</p> <p>What does the sticker say?</p> <p>The sticker says "iPod".</p> <p>Where is the photo taken?</p> <p>It looks like it's taken in a backyard.</p> <p>Do you think it is printed or handwritten?</p> <p>It looks like it's handwritten.</p> <p>What color is the sticker?</p> <p>It's white.</p>
---	---	---

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

# LiMBeR

- Maybe we can also use a frozen image encoder?



<b>CLIP</b>	a giraffe in the lobby of the building
<b>NFRN50</b>	the giraffe in the zoo.
<b>BEIT</b>	a peacock in the garden
<b>NFRN50</b>	a man and a woman in a field of flowers
<b>Random</b>	

More Language Supervision\*

Less Language Supervision\*



<b>CLIP</b>	A tennis racket
<b>NFRN50</b>	A tennis racket
<b>BEIT</b>	A baseball bat.
<b>NFRN50</b>	A tree
<b>Random</b>	

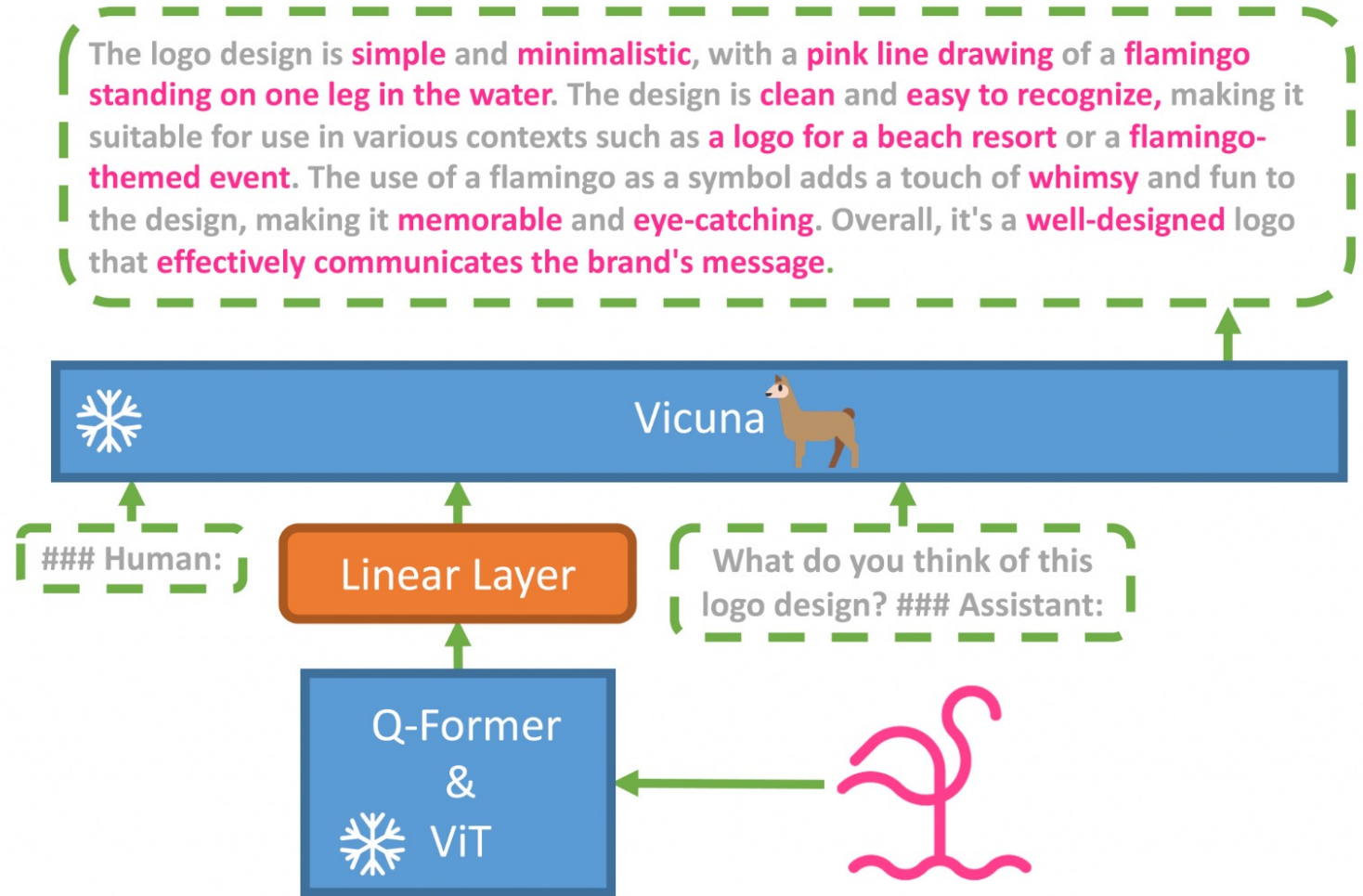
**Q: What is the person holding?**  
**A: tennis racket**

# Conditioning Pretrained Language Models

## Mini-GPT4

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

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



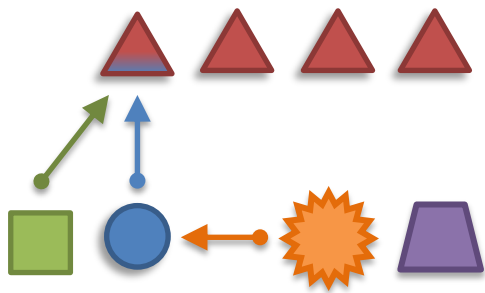
The architecture of MiniGPT-4.

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

# Conditioning Pretrained Language Models

## LLaMA-Adapter

Can be combined with ImageBind – alignment of many modalities to language (i.e., high-modality coordination model)



**LLaMA-Adapter:**  
**Bilingual Multi-modality**  
**Instruction Model**

**Example: 3D Point Cloud to Image (Bilingual)**

Generate an image from the 3D point cloud.

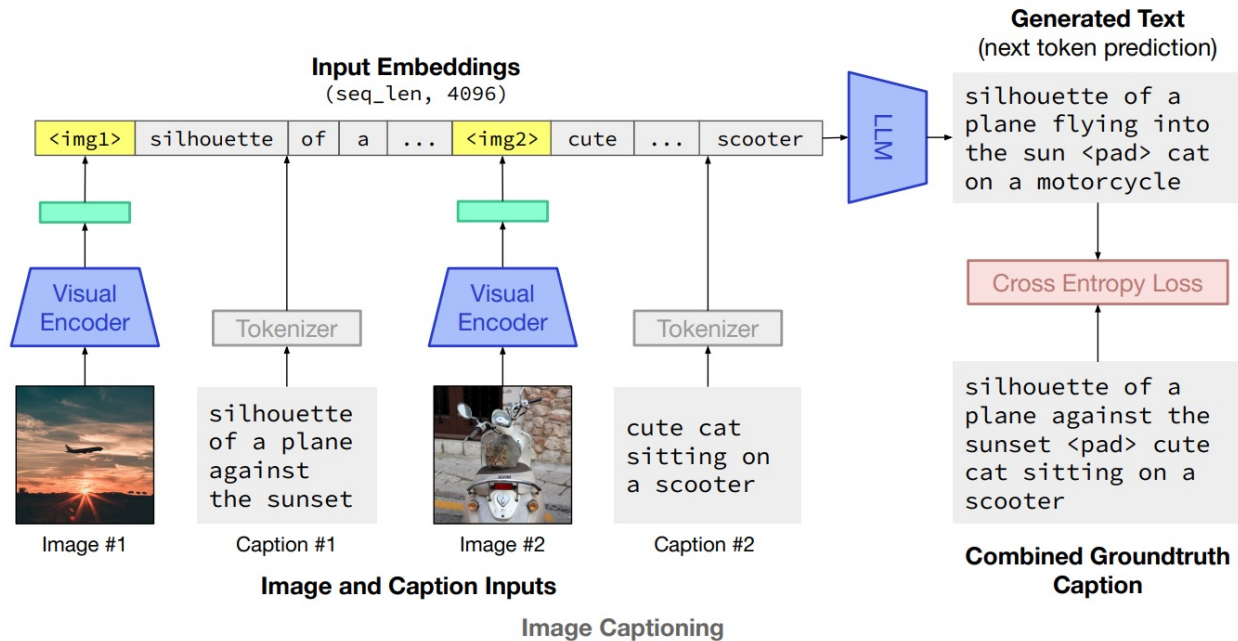
根据这个3D点云生成一张图片。

[Gridhar et al., ImageBind: One Embedding Space To Bind Them All. CVPR 2023]  
[Gao et al., LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arxiv 2023]

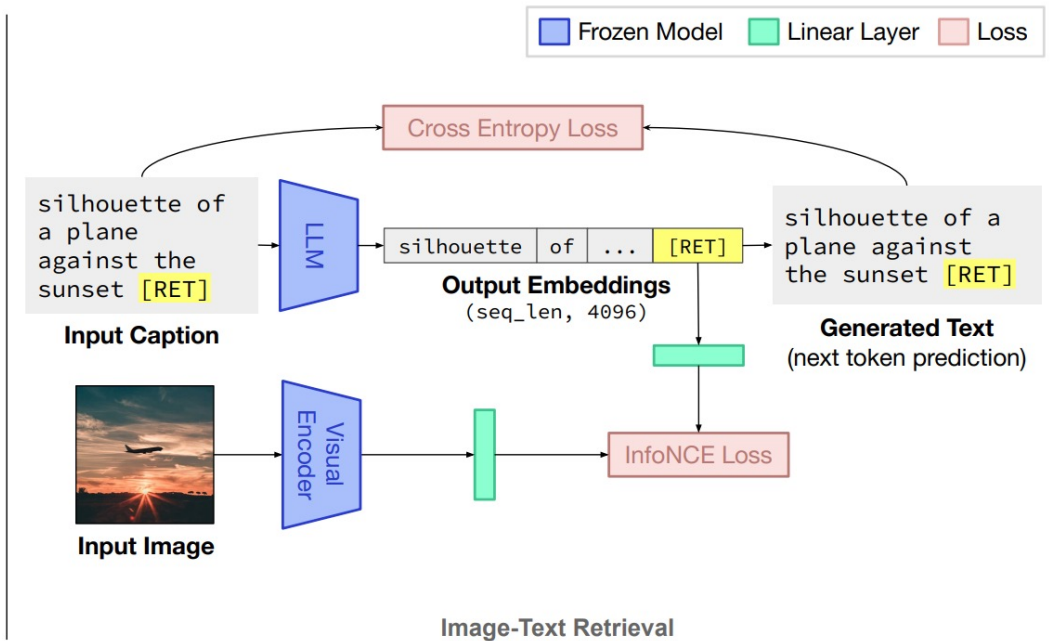
# FROMAGe: Grounding Frozen LMs to Images

- LIMBeR + CLIP. Trainable in 1 day on 1 GPU

## LIMBeR, basically



## CLIP, with a frozen LLM



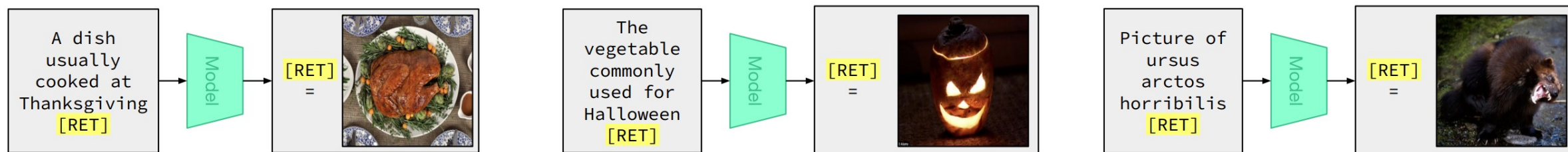


# FROMAGe: Grounding Frozen LMs to Images

- Freezing the LM allows generalizing to novel tasks and domains



Concept composition



Knowledge-aware retrieval

[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

# FROMAGe: Grounding Frozen LMs to Images

I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.


A beaver.

That's right, what do beavers look like?


They have a tail.

Show me an illustration of a beaver



A beaver.



Yes, what about a pencil drawing of one?



What about a photo of one?



What is this?


This is a house with a unique design.

What is its shape?

The house is shaped like a dome.


Can you show me more houses like this?

The house is designed by architect.



That's nice, any others?

Yes, there are many houses like this in the world.



Provide instructions on how to cook.


How do I make macarons?

Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar.

What does the stiff egg white look like?

The egg white should be stiff and glossy.

Show me a picture of what that looks like

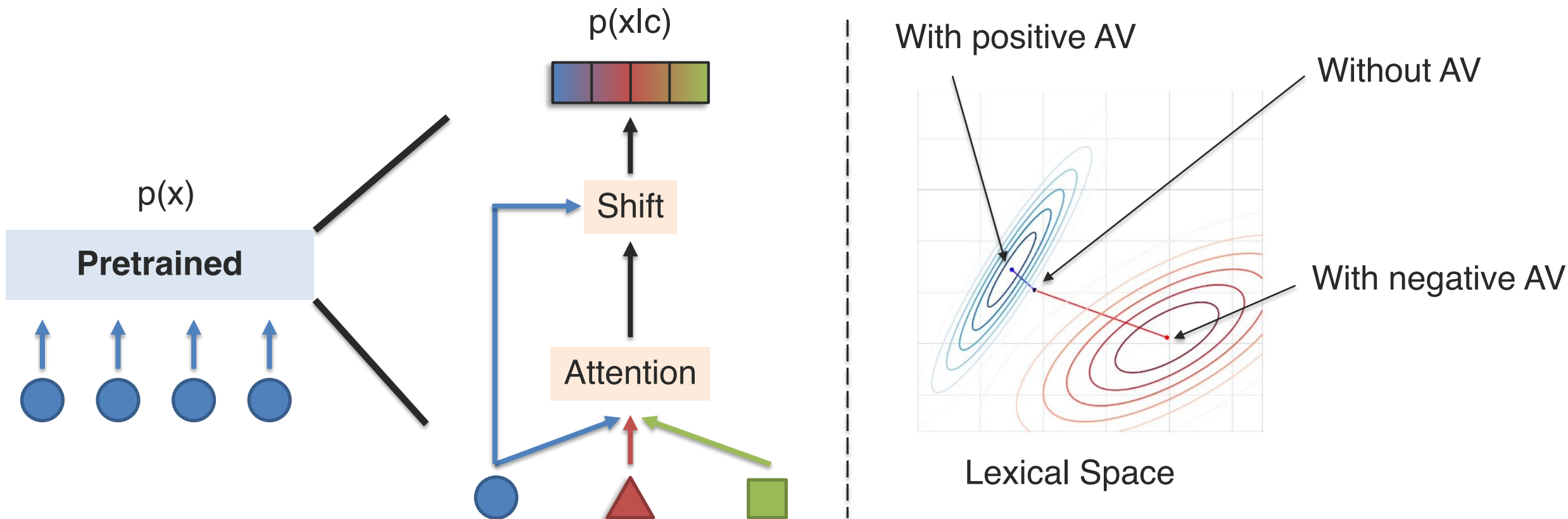


After this, what do I do?

The macaron is then baked in a hot oven.

# Conditioning Autoregressive Models

## Conditioning via representation tuning

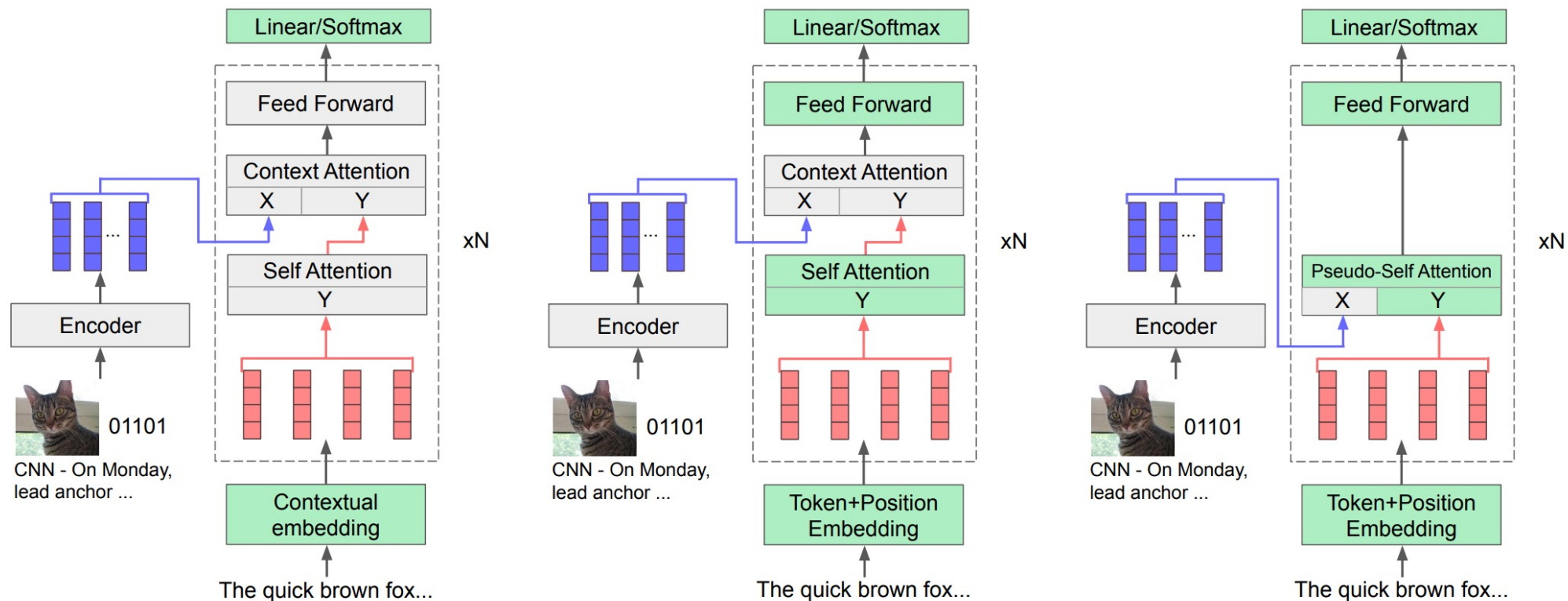


[Ziegler et al., Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv 2019]

[Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers. ACL 2020]

# Conditioning Autoregressive Models

## Conditioning via pseudo-attention



(a) Repr-Transformer

(b) Context-Attn

(c) Pseudo-Self

[Ziegler et al., Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv 2019]

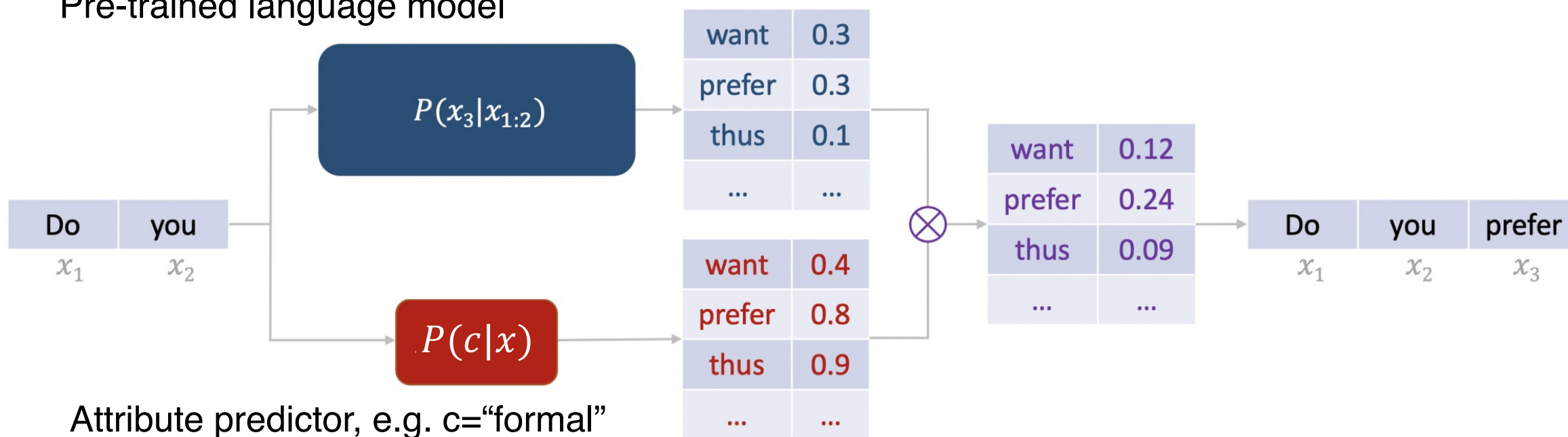
[Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers. ACL 2020]

# Conditioning Autoregressive Models

## Conditioning via Bayes' rule

$$p(x|c) \propto p(c|x)p(x)$$

Pre-trained language model

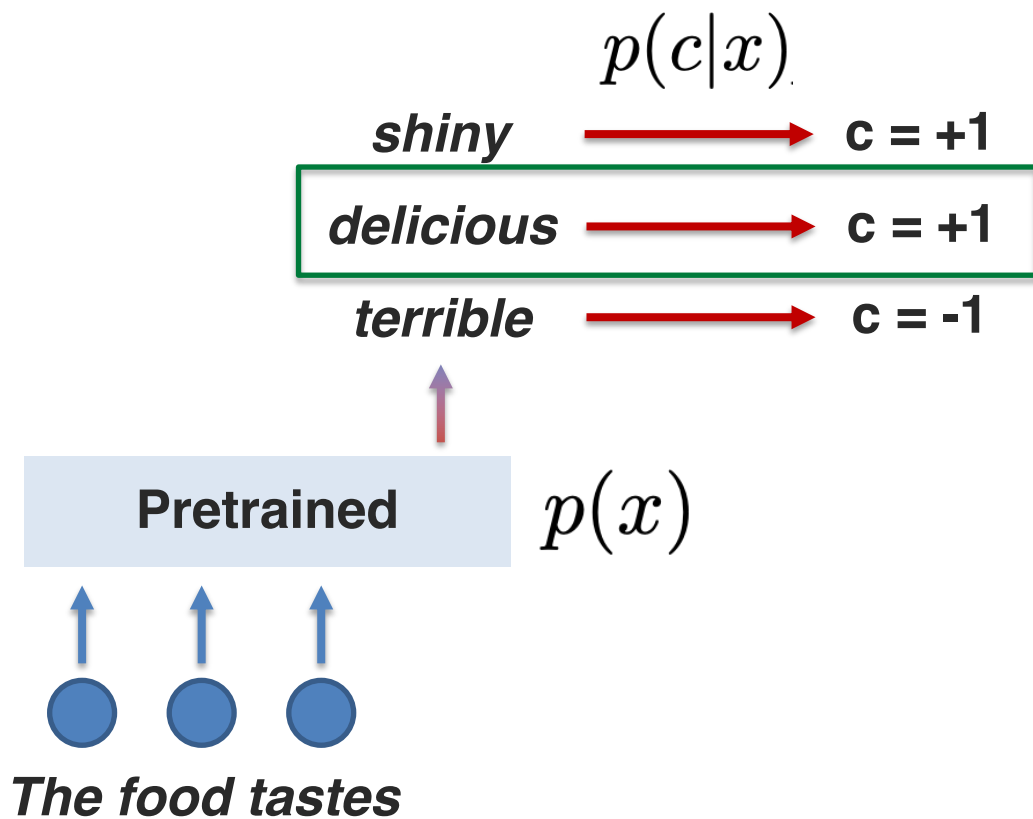


[Yang and Klein, FUDGE: Controlled Text Generation With Future Discriminators. NAACL 2021]

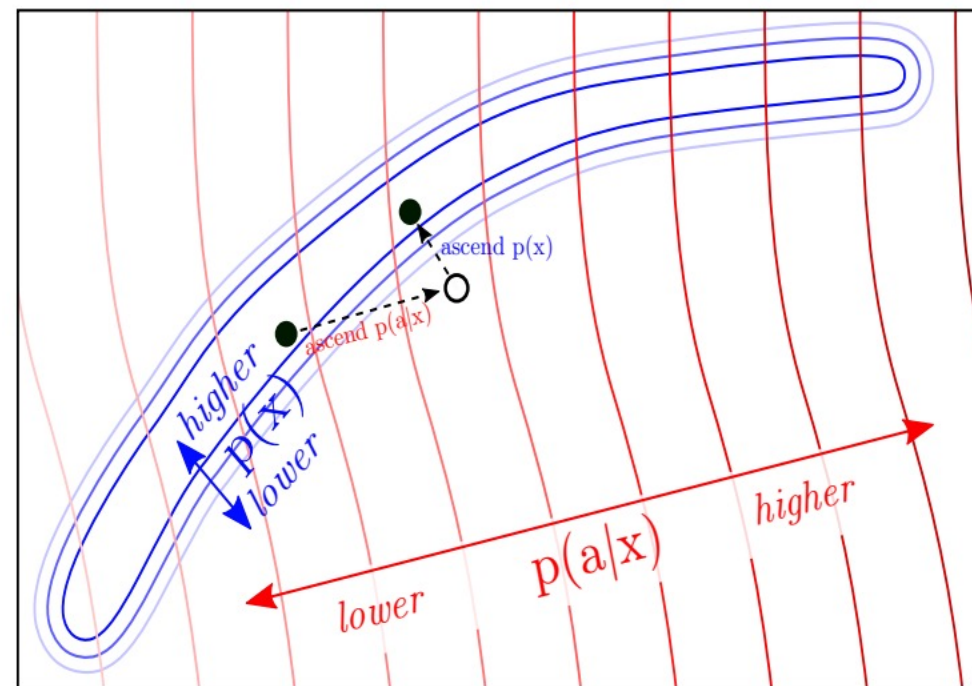
[Lazaridou et al. Multi-agent Communication Meets Natural Language. ACL 2020]

# Conditioning Autoregressive Models

## Conditioning via gradient tuning



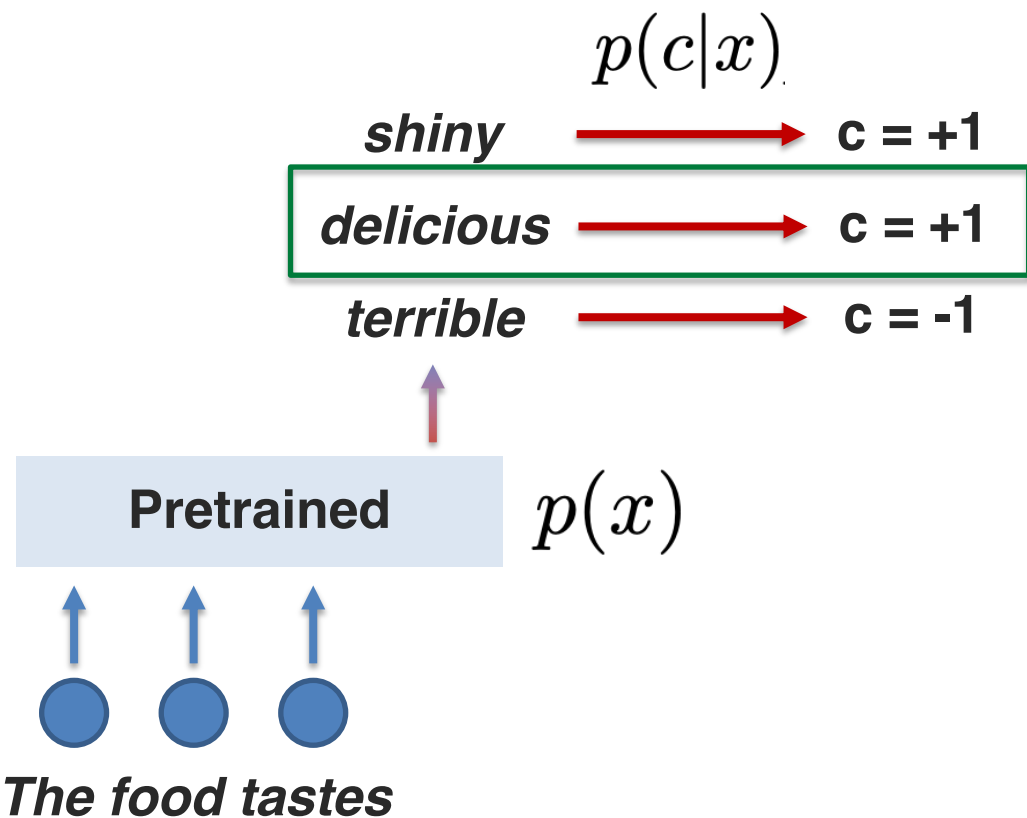
$$p(x|c) \propto p(c|x)p(x)$$



[Dathathri et al., Plug and Play Language Models: A Simple Approach to Controlled Text Generation. ICLR 2020]

# Conditioning Autoregressive Models

## Conditioning via gradient tuning



$$p(x|c) \propto p(c|x)p(x)$$

$H_t$  are final-layer representations at time  $t$

1. Increasing  $p(c|x)$ .

$$\Delta H_t \leftarrow \Delta H_t + \alpha \nabla_{\Delta H_t} \log p(c|H_t + \Delta H_t)$$

2. Increasing  $p(x)$

$$\Delta H_t \leftarrow \Delta H_t + \alpha \lambda \text{KL}(p(x) || p_{\Delta H_t}(x))$$

3. Generate next token using  $H_t + \Delta H_t$

# Summary: Autoregressive Models

---

- Relatively easy to train.
- Slow to sample from.
- Not easy to condition on.

Input Prompt:

Recite the first law of robotics



Output: