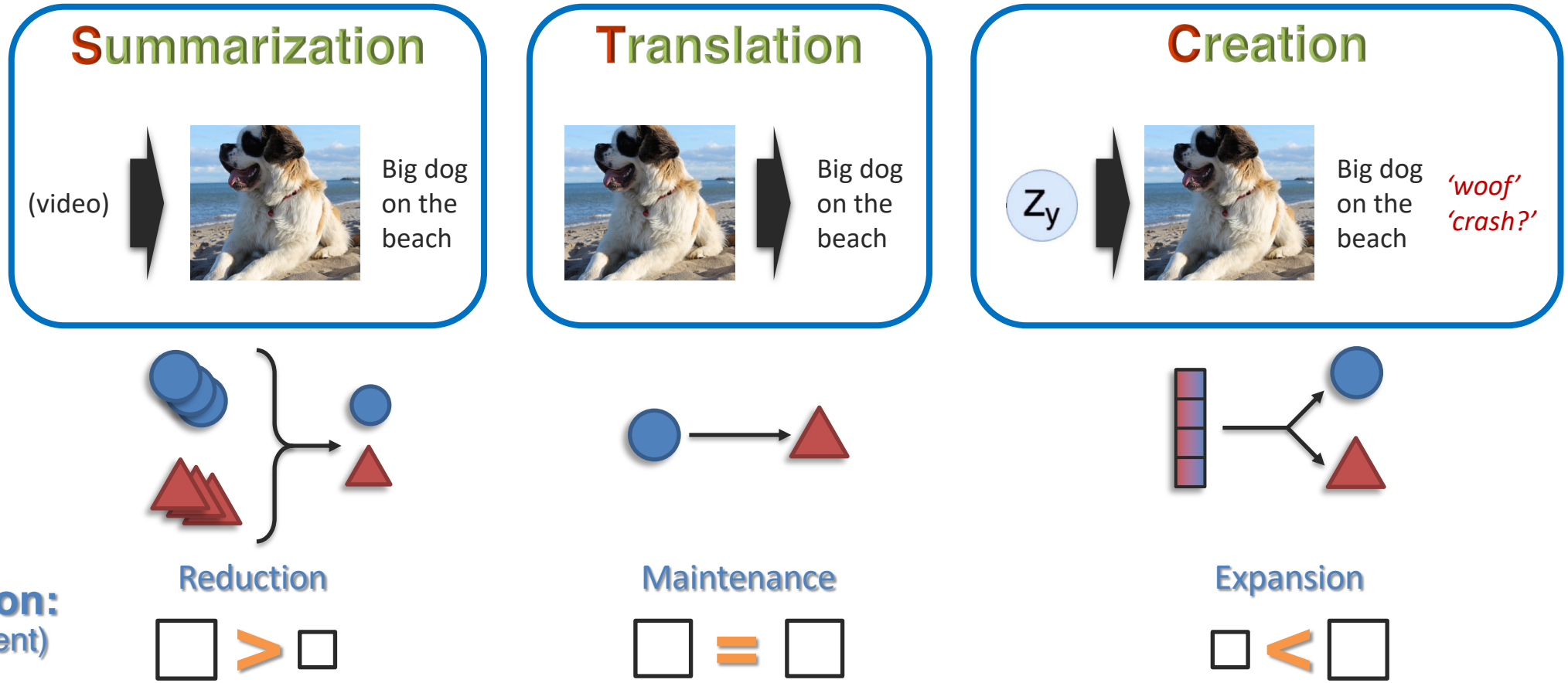
Challenge 4: Generation

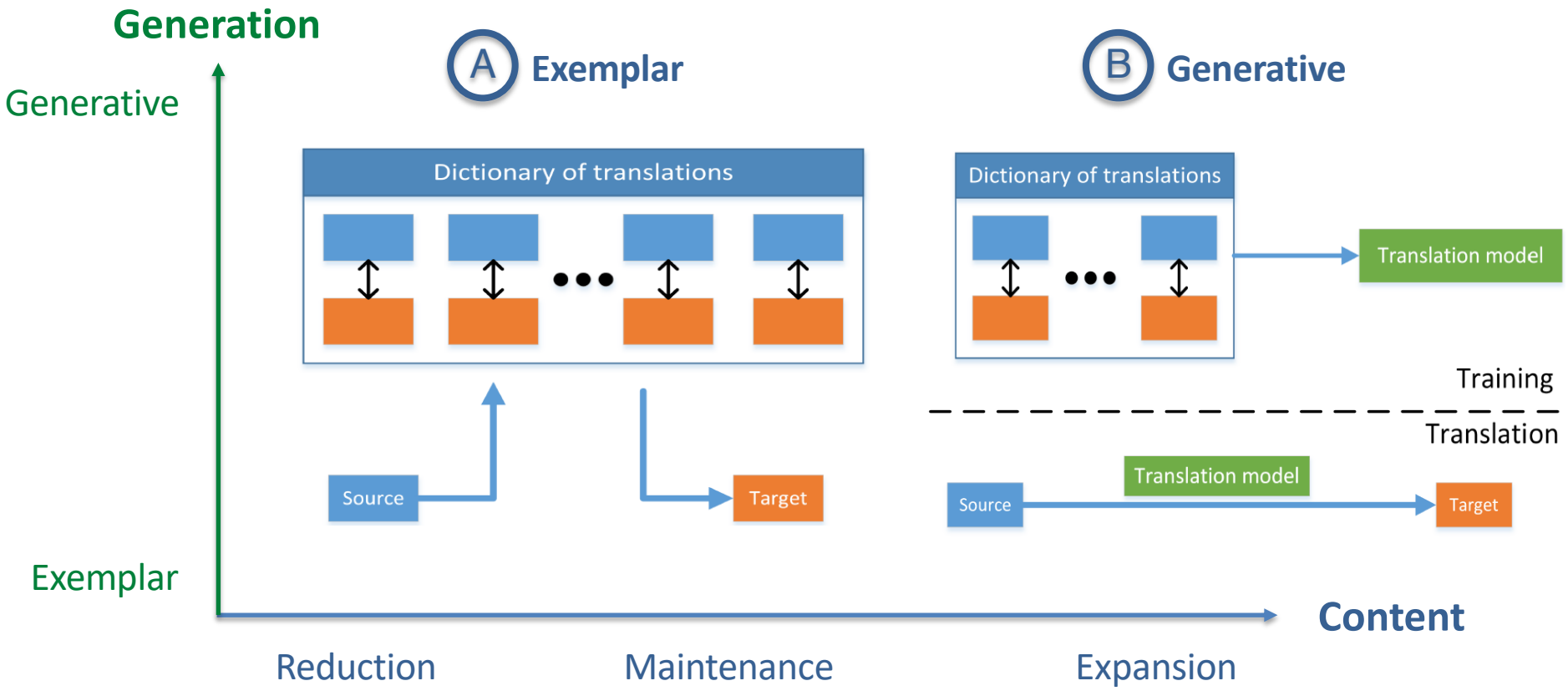
Generation

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



Generation

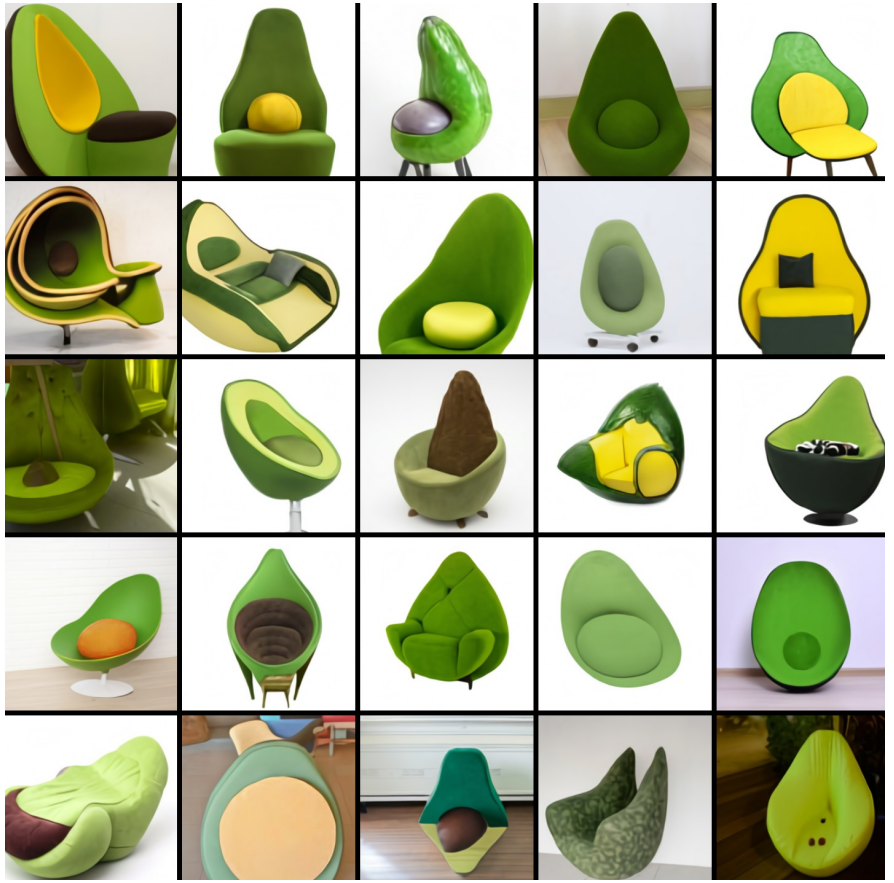
Decoding high-dimensional multimodal data.



Sub-challenge 4a: Translation

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

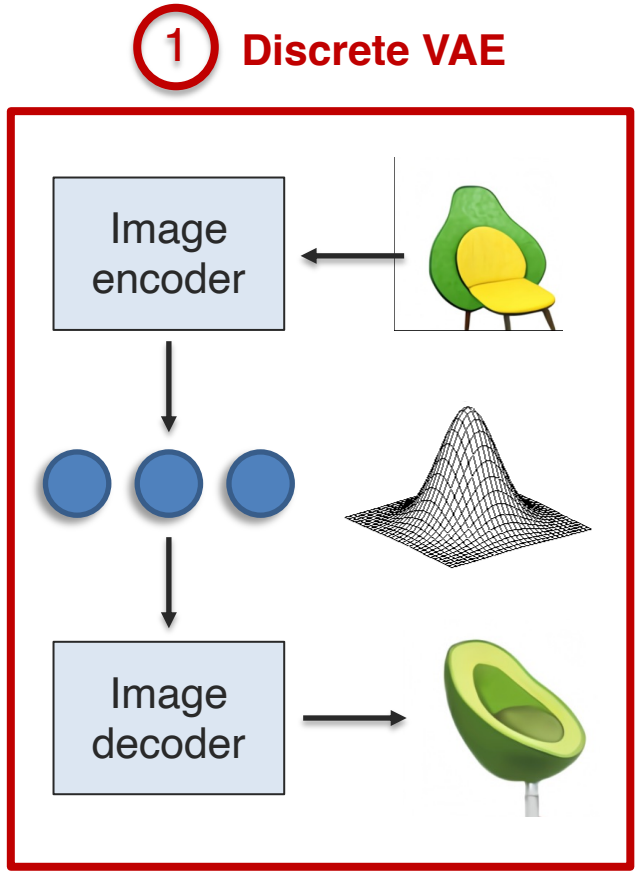
An armchair in the shape of an avocado →



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

Sub-challenge 4a: Translation

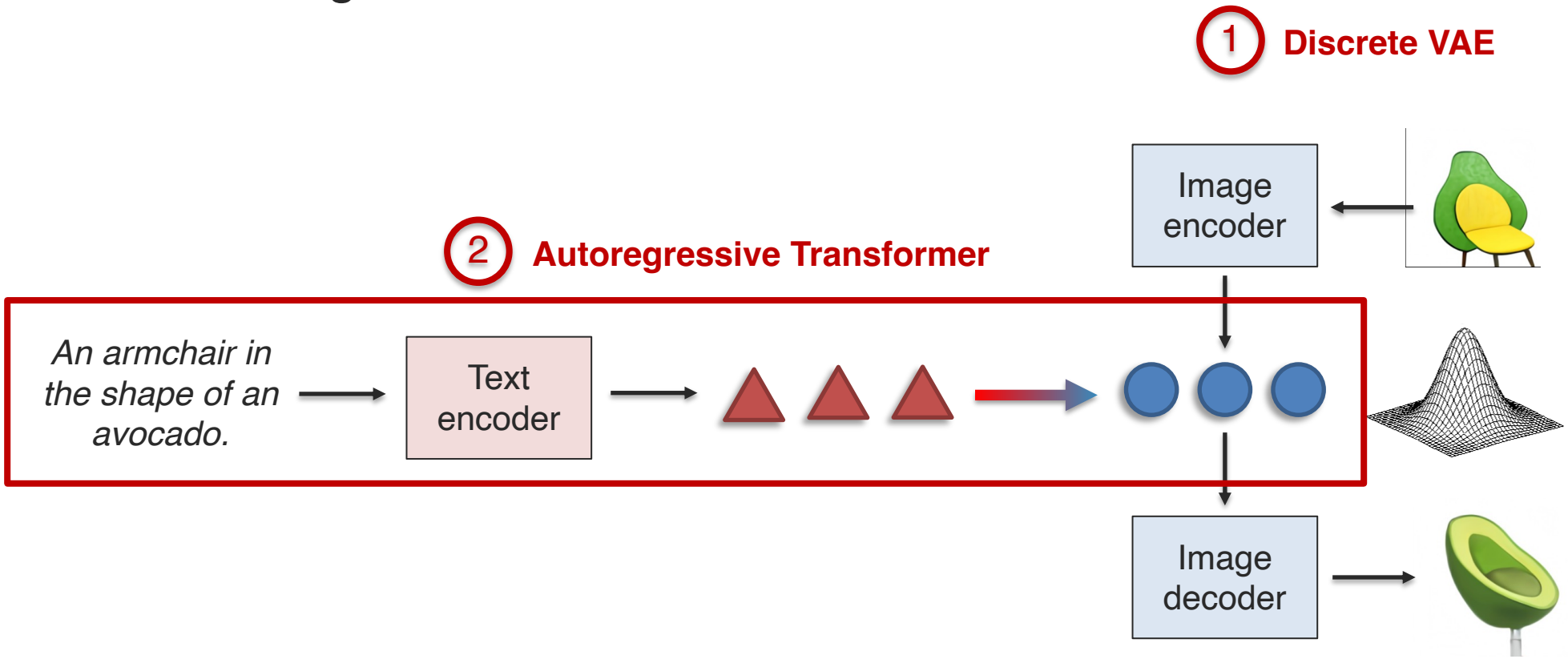
DALL·E: Text-to-image translation at scale



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

Sub-challenge 4a: Translation

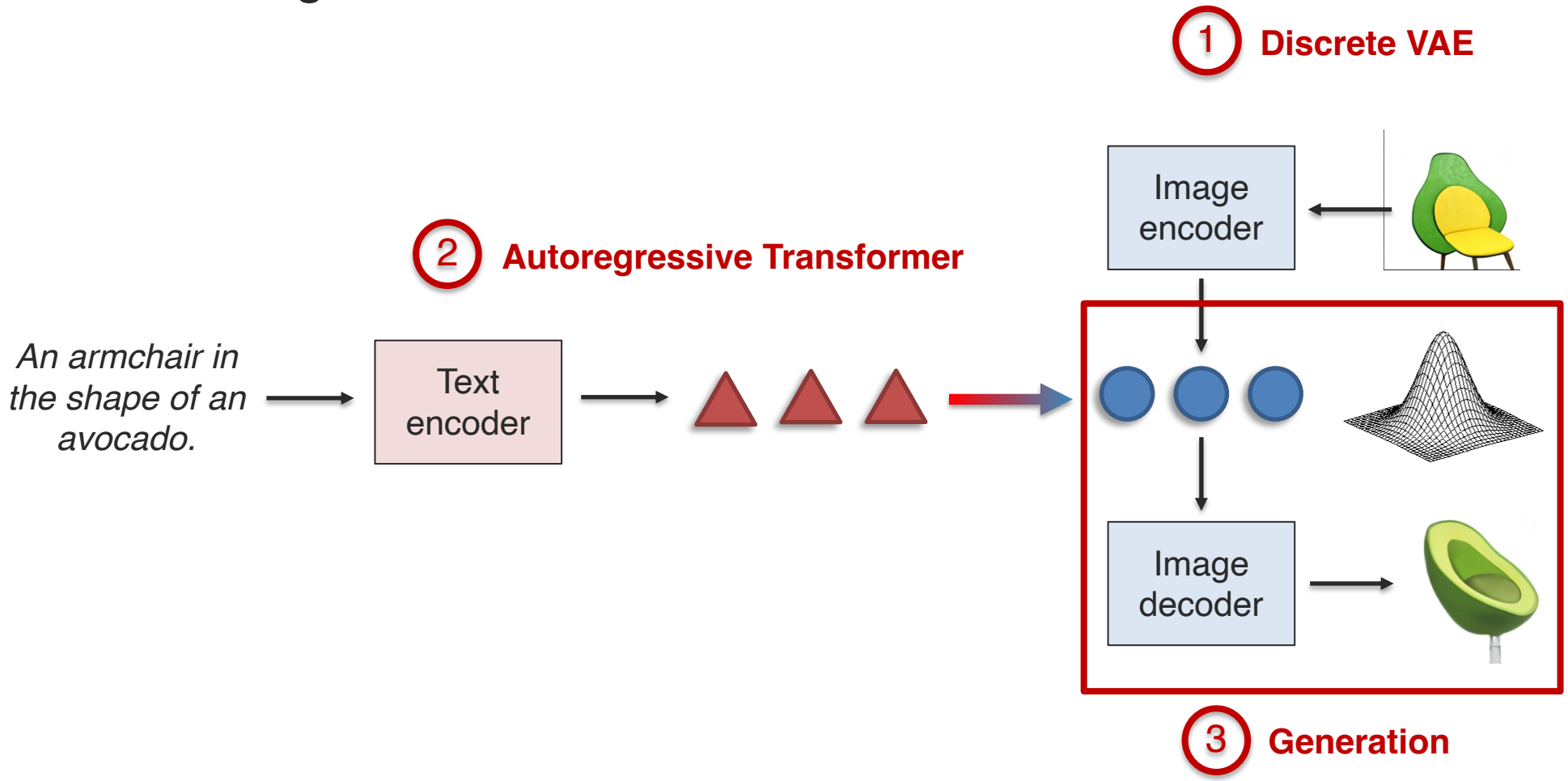
DALL·E: Text-to-image translation at scale



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

Sub-challenge 4a: Translation

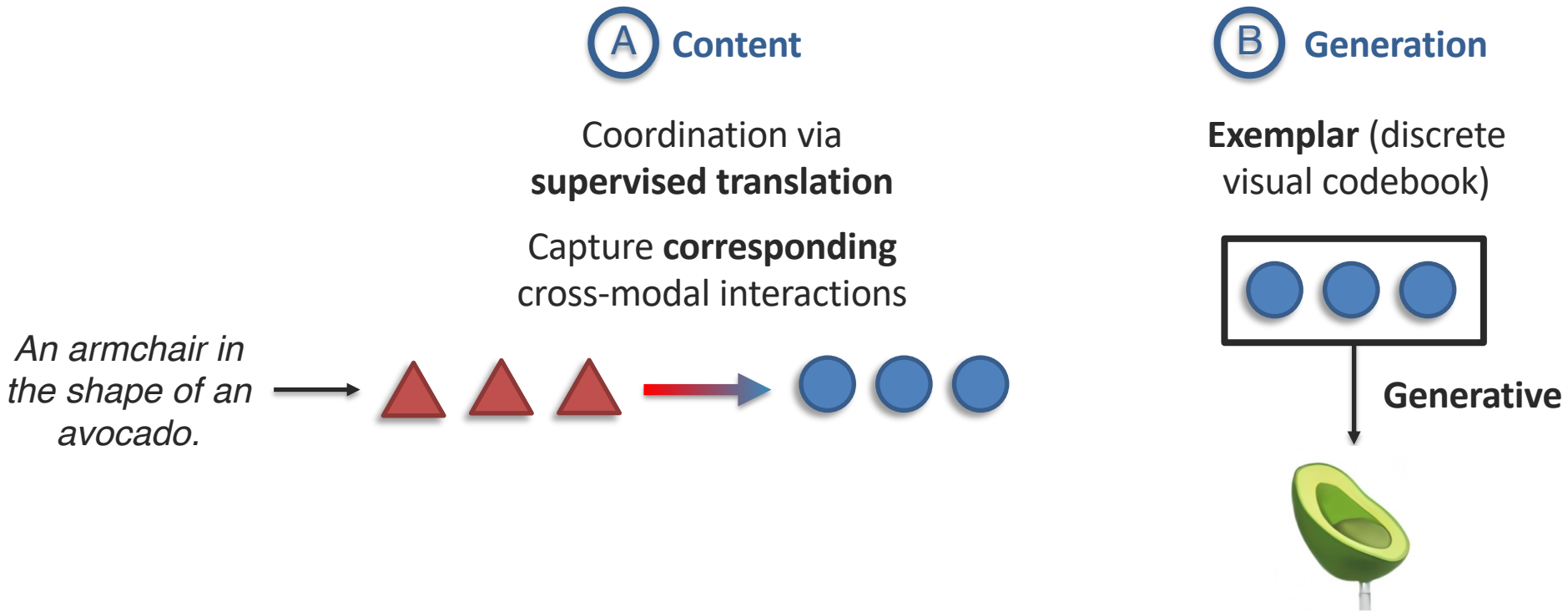
DALL·E: Text-to-image translation at scale



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

Sub-challenge 4a: Translation

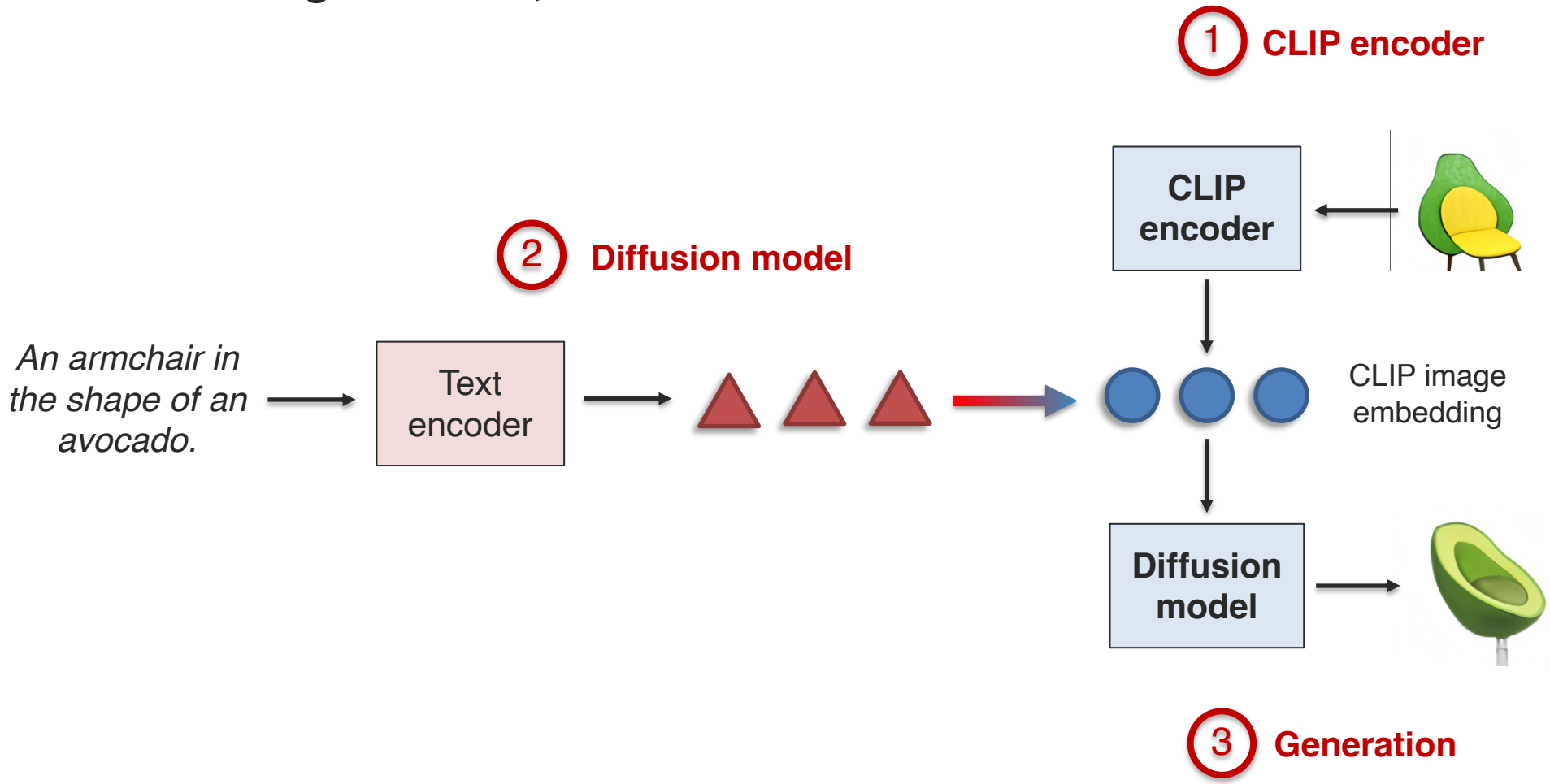
DALL·E: Text-to-image translation at scale



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

Sub-challenge 4a: Translation

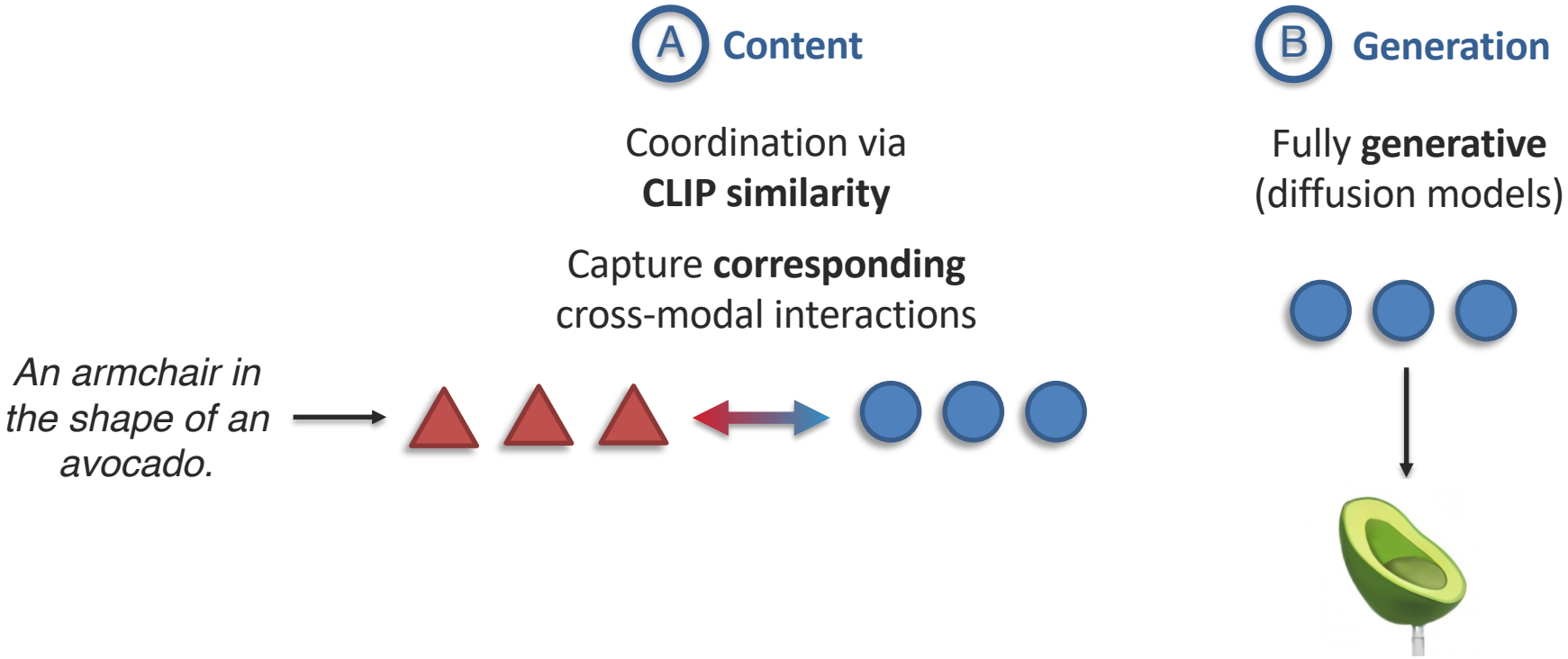
DALL·E 2: Combining with CLIP, diffusion models



[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]

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]

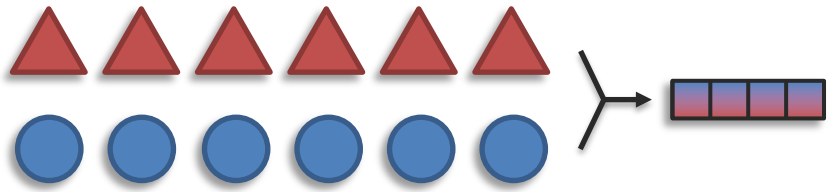
Sub-challenge 4b: Summarization

Video summarization

A Content

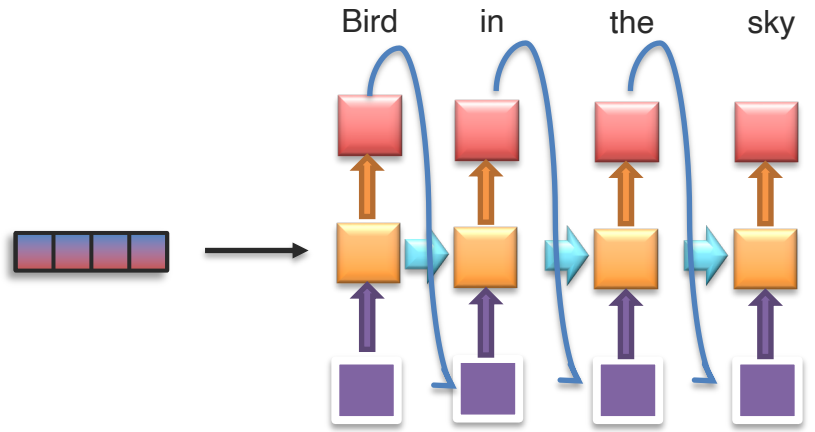
Fusion via
joint representation

Capture **complementary**
cross-modal interactions



B Generation

Generative ≈ abstractive summarization

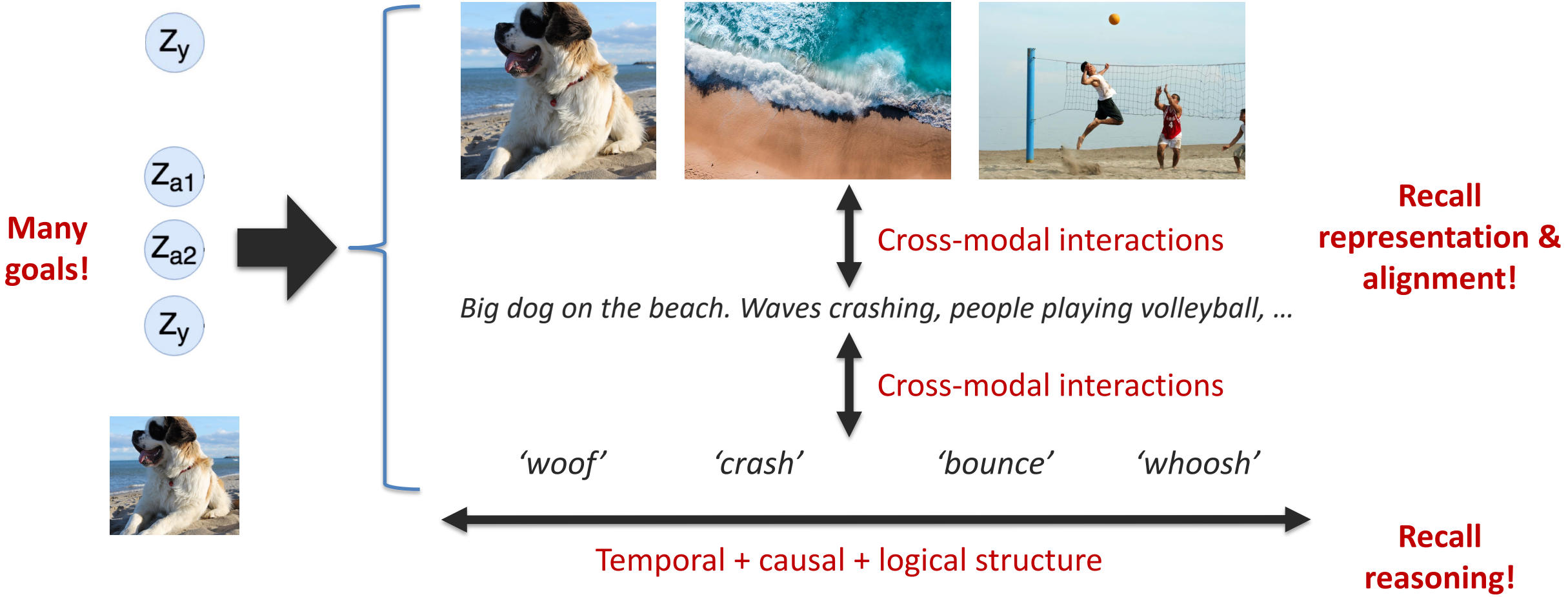


[Palaskar et al., Multimodal Abstractive Summarization for How2 Videos. ACL 2019]

Sub-challenge 4c: Creation

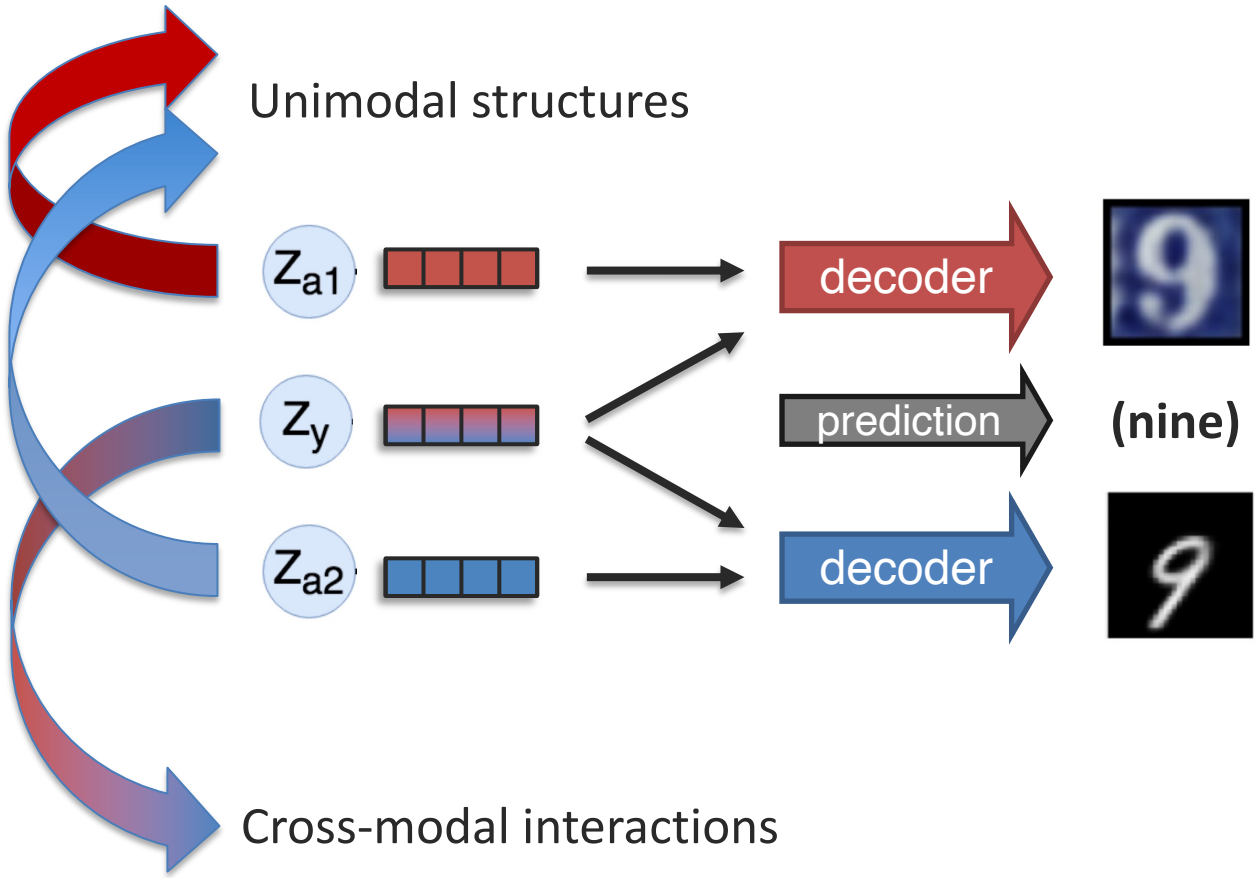
Open challenges

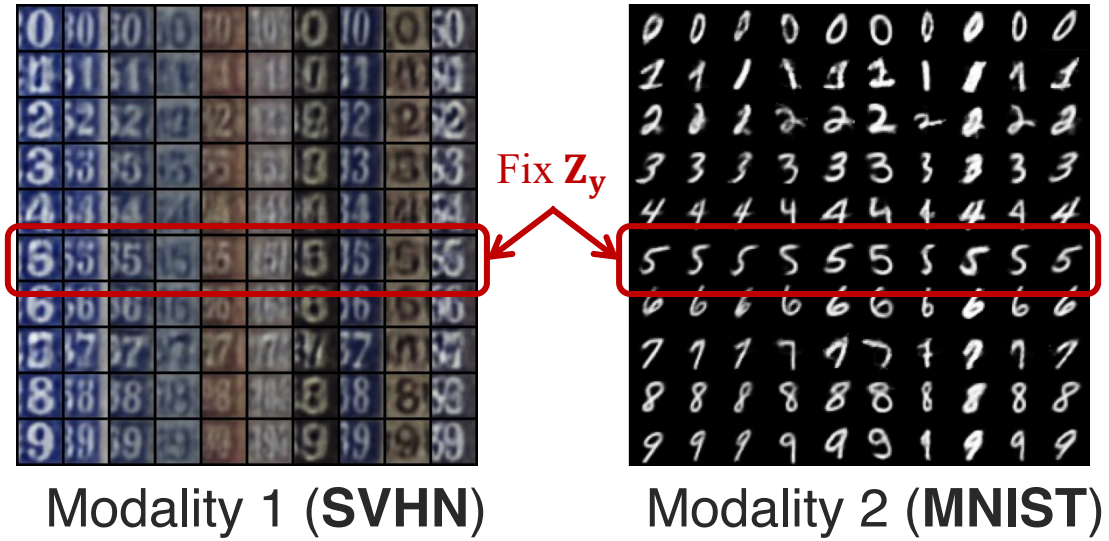
Definition: Simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities.



Sub-challenge 4c: Creation

Some initial attempts: factorized generation

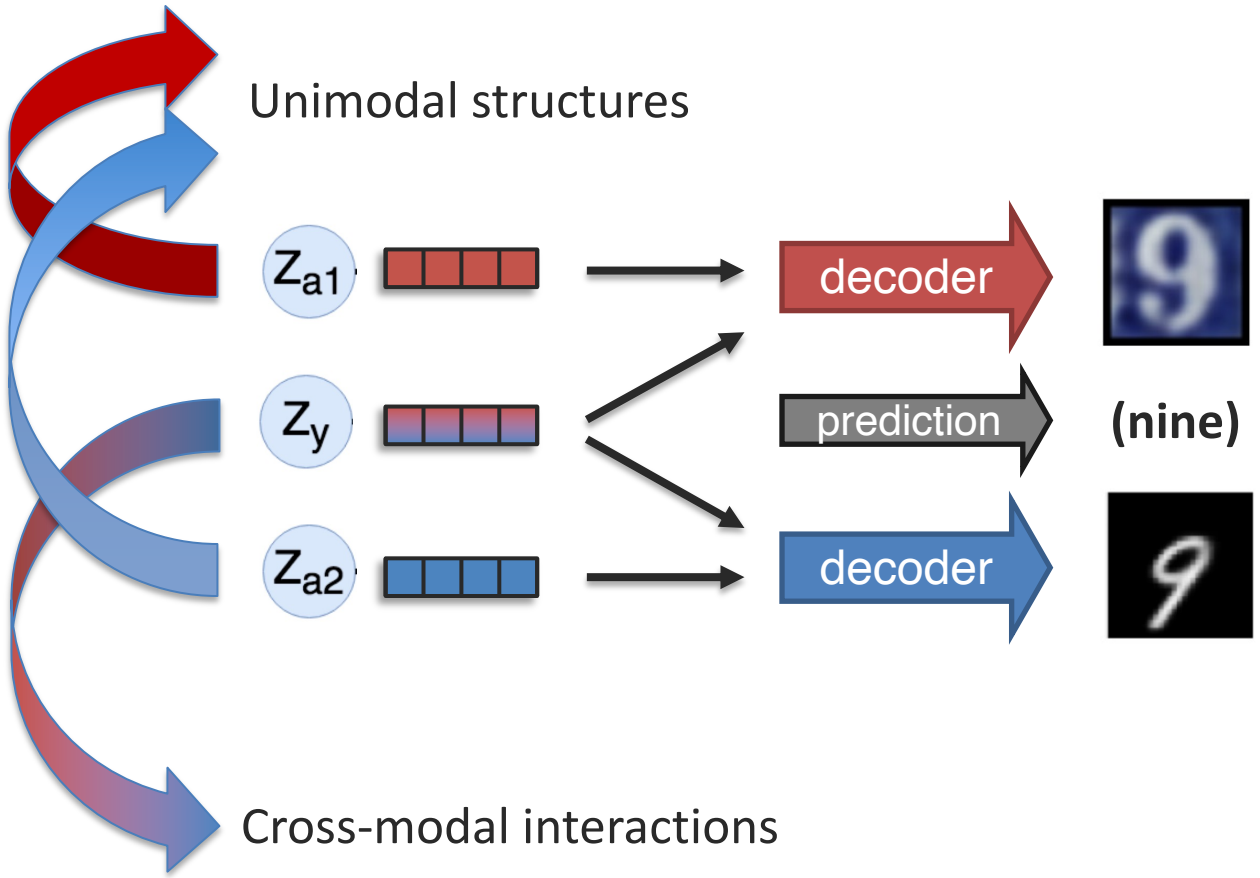


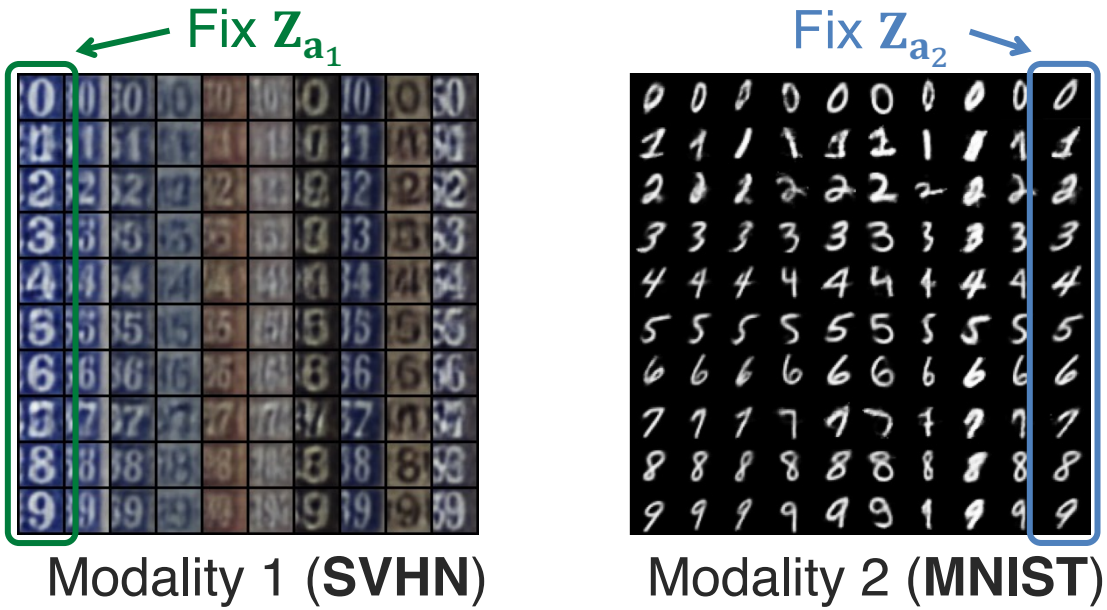


[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Sub-challenge 4c: Creation

Some initial attempts: factorized generation

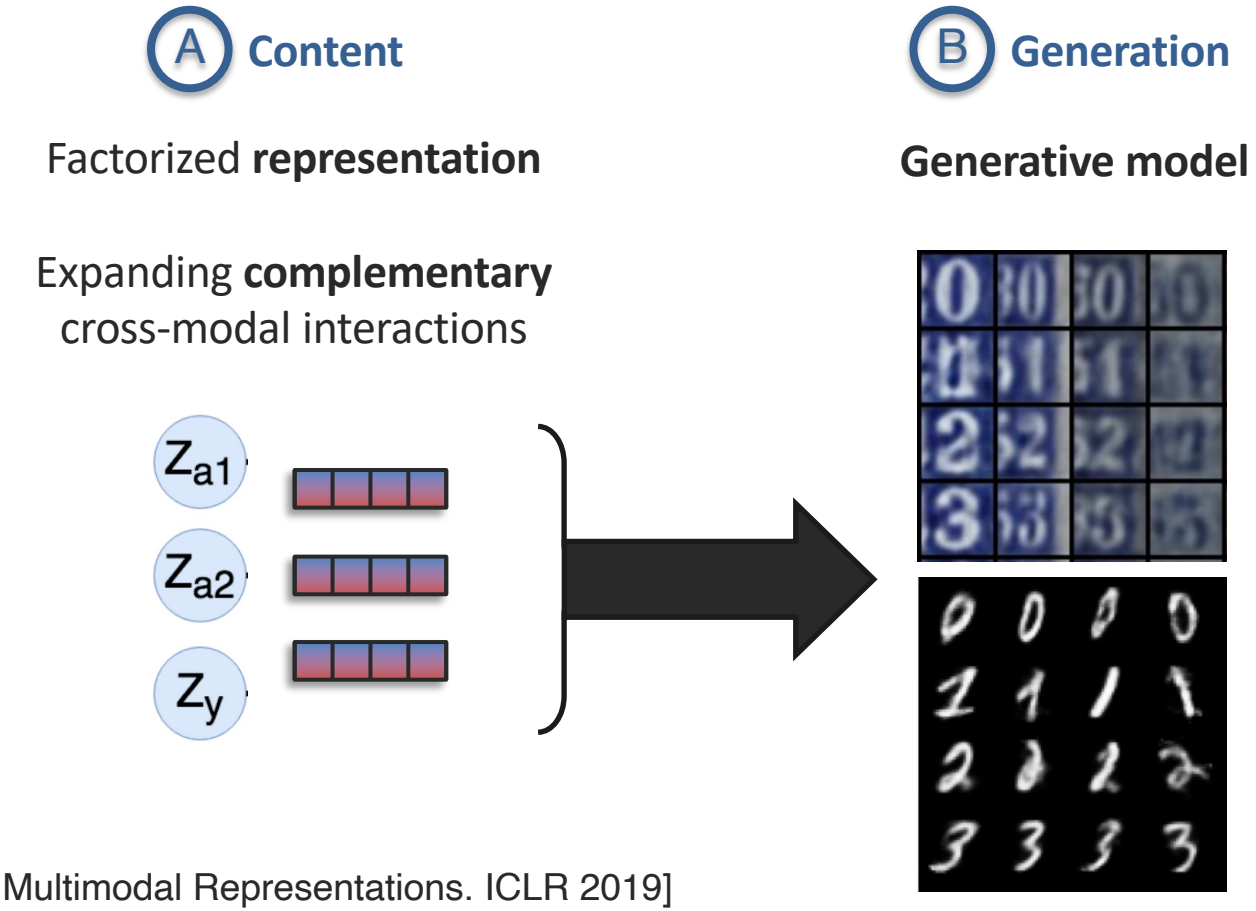




[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Sub-challenge 4c: Creation

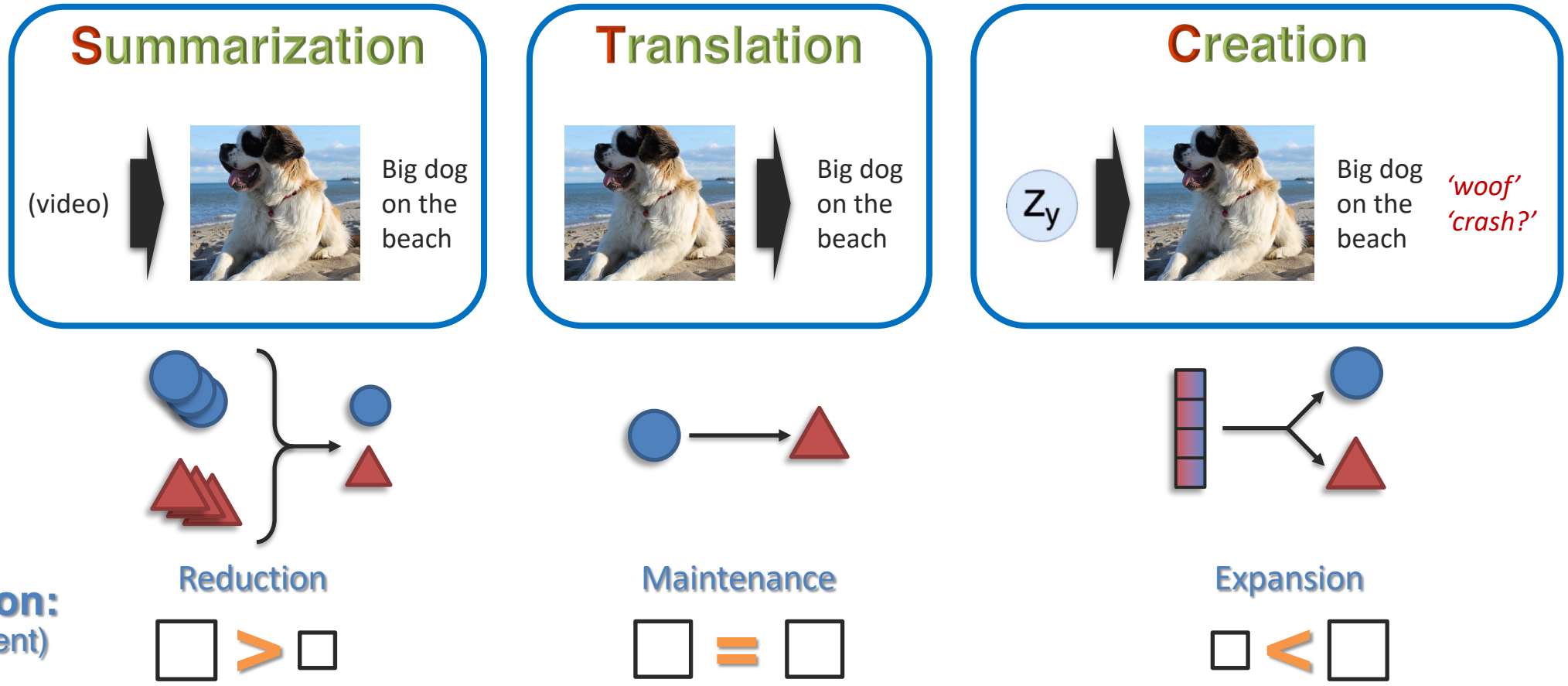
Some initial attempts: factorized generation



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

Summary: Generation

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



Model Evaluation & Ethical Concerns

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

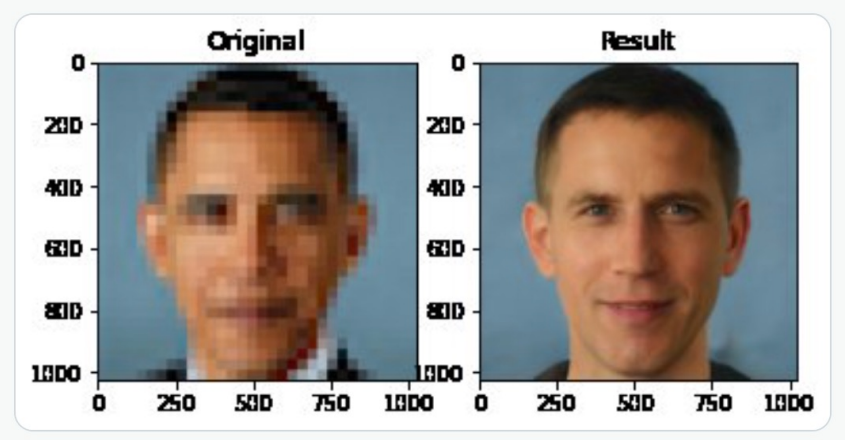
[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]
[Carlini et al., Extracting Training Data from Large Language Models. USENIX 2021]
[Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019]

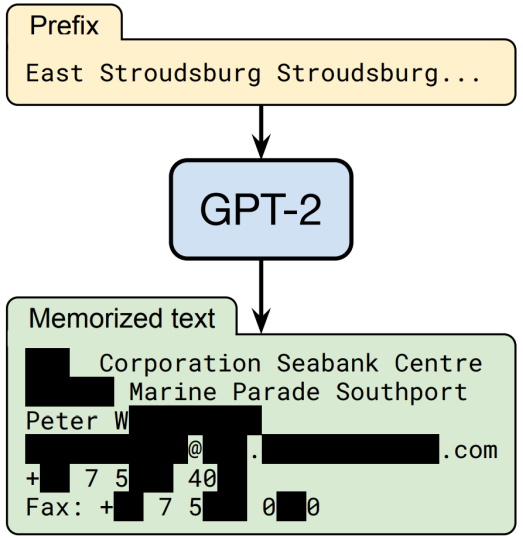
Model Evaluation & Ethical Concerns

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]