CS769 Advanced NLP

Multimodal Machine Learning: Vision-Language

Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from LP Morency

https://junjiehu.github.io/cs769-spring23/

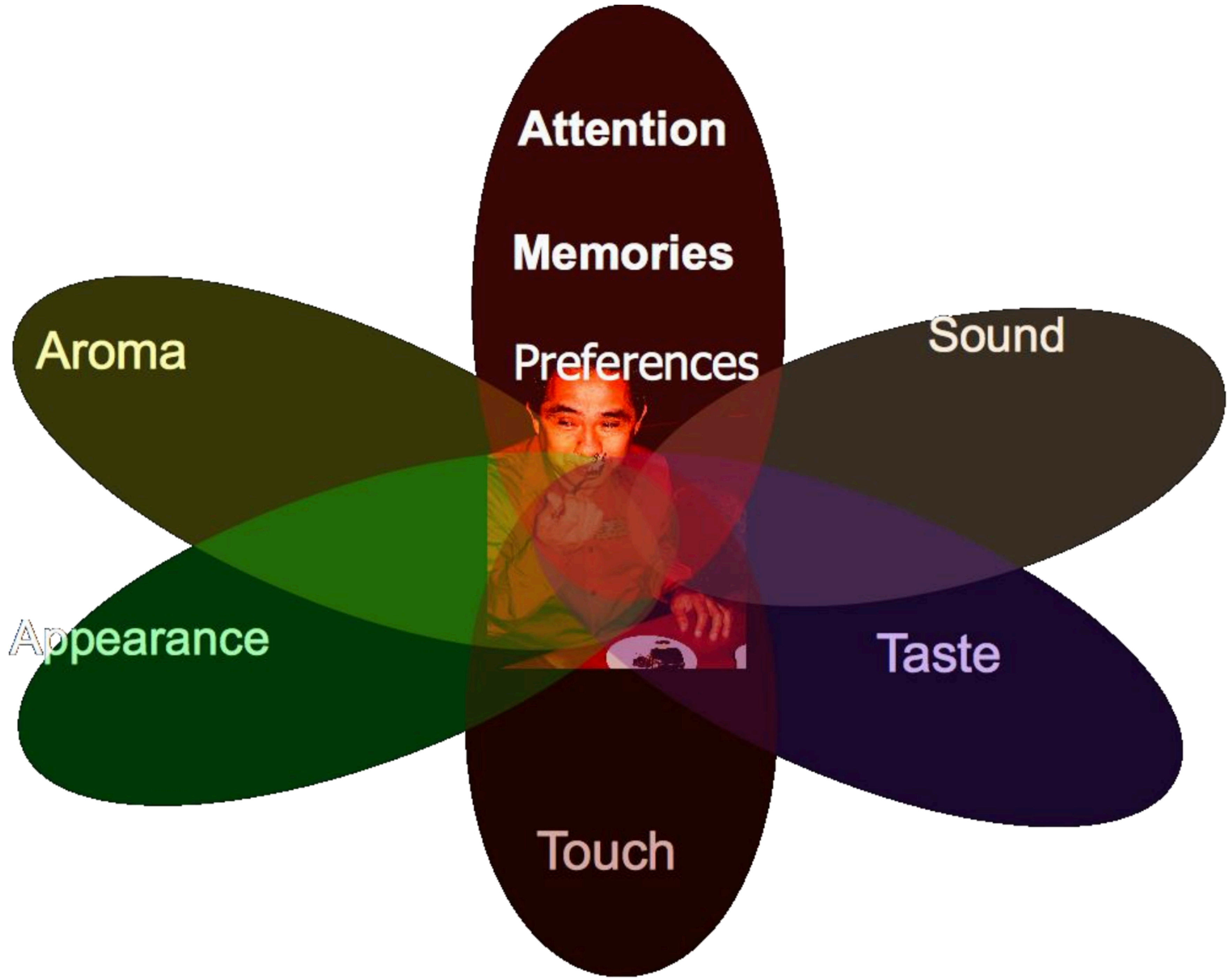
Goal for Today

- What is Multimodal?
  - Historical view, multimodal vs multimedia
- Core technical challenges
  - Representation learning, translation, alignment, fusion, and co-learning
- Recent pre-trained V+L models
  - CLIP
  - DALL-E

Multimodal Machine Learning

What is Multimodal?

Sensory Modalities



Multimodal Communicative Behaviors

Verbal

Lexicon
- Words

Syntax
- Part - of - speech
- Dependencies

Pragmatics
- Discourse acts

Vocal

Prosody
- Intonation
- Voice quality

Vocal expressions
- Laughter, moans

Visual

Gestures
- Head gestures
- Eye gestures
- Arm gestures

Body language
- Body posture
- Proxemics

Eye contact
- Head gaze
- Eye gaze

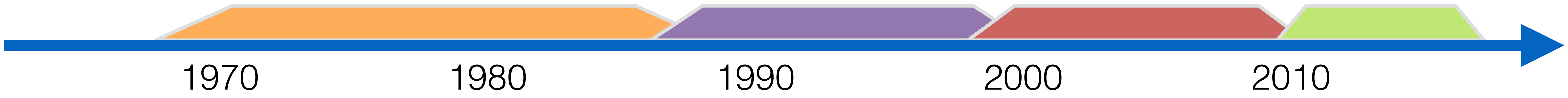
Facial expressions
- FACS action units
- Smile, frowning

Examples of Modalities

- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds, and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

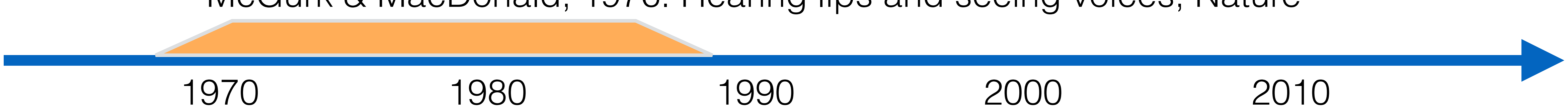
Prior Research on “Multimodal”

- Four eras of multimodal research
  - The “ behavioral” era (1970s until late 1980s)
  - The “ computational” era (late 1980s until 2000)
  - The “ interaction” era (2000 - 2010)
  - The “ deep learning” era (2010s until …)



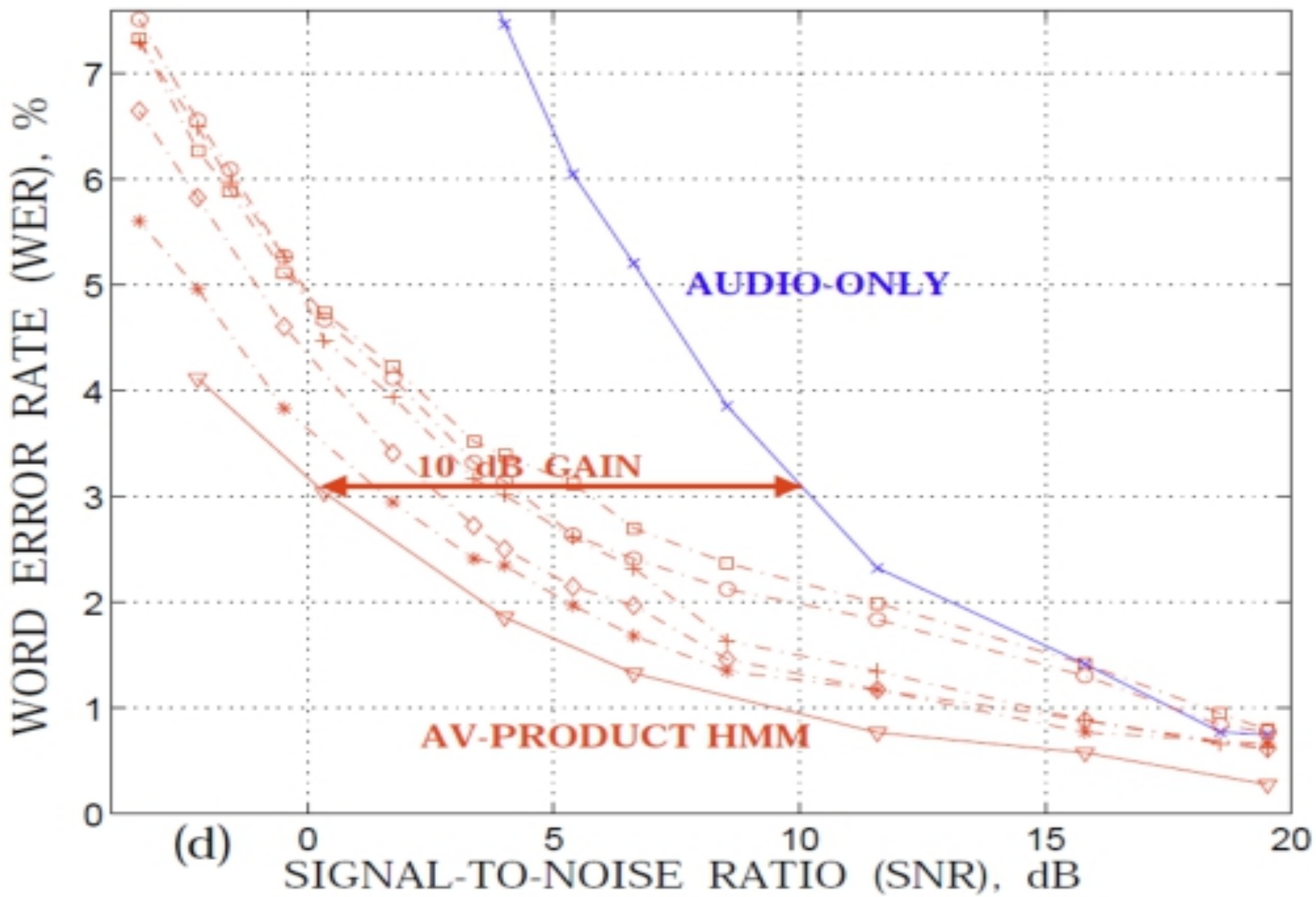
The McGurk Effect (1976)

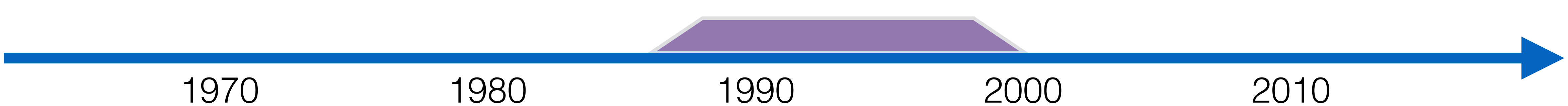
McGurk & MacDonald, 1976. Hearing lips and seeing voices, Nature



The "Computational" Era (Late 1980s until 2000)

- Audio-Visual Speech Recognition (AVSR)





Core Technical Challenges

Core Challenges in “Deep” Multimodal ML (Baltrusaitis et al. 2017)

- Representation
- Alignment
- Fusion
- Translation
- Co-Learning

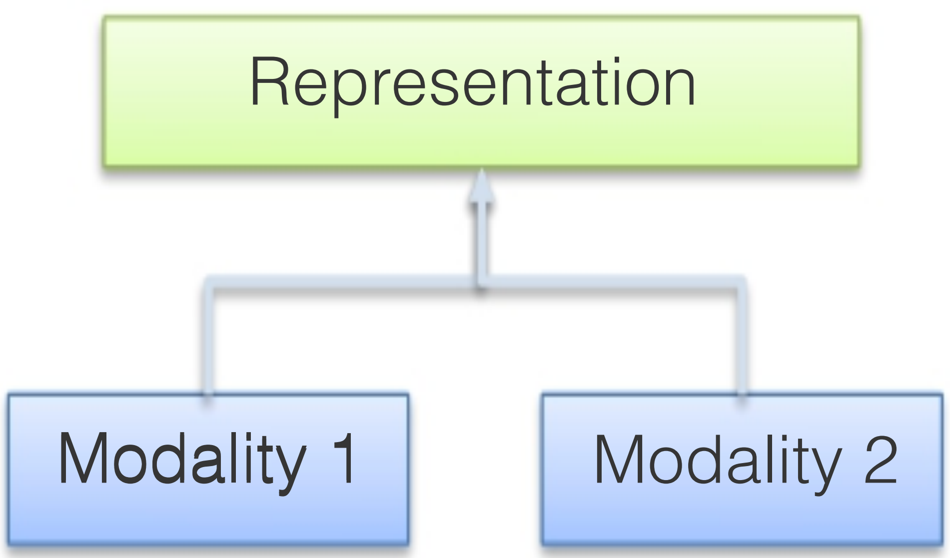
These challenges are non-exclusive.

Baltrusaitis et al. 2017. Multimodal Machine Learning: A Survey and Taxonomy

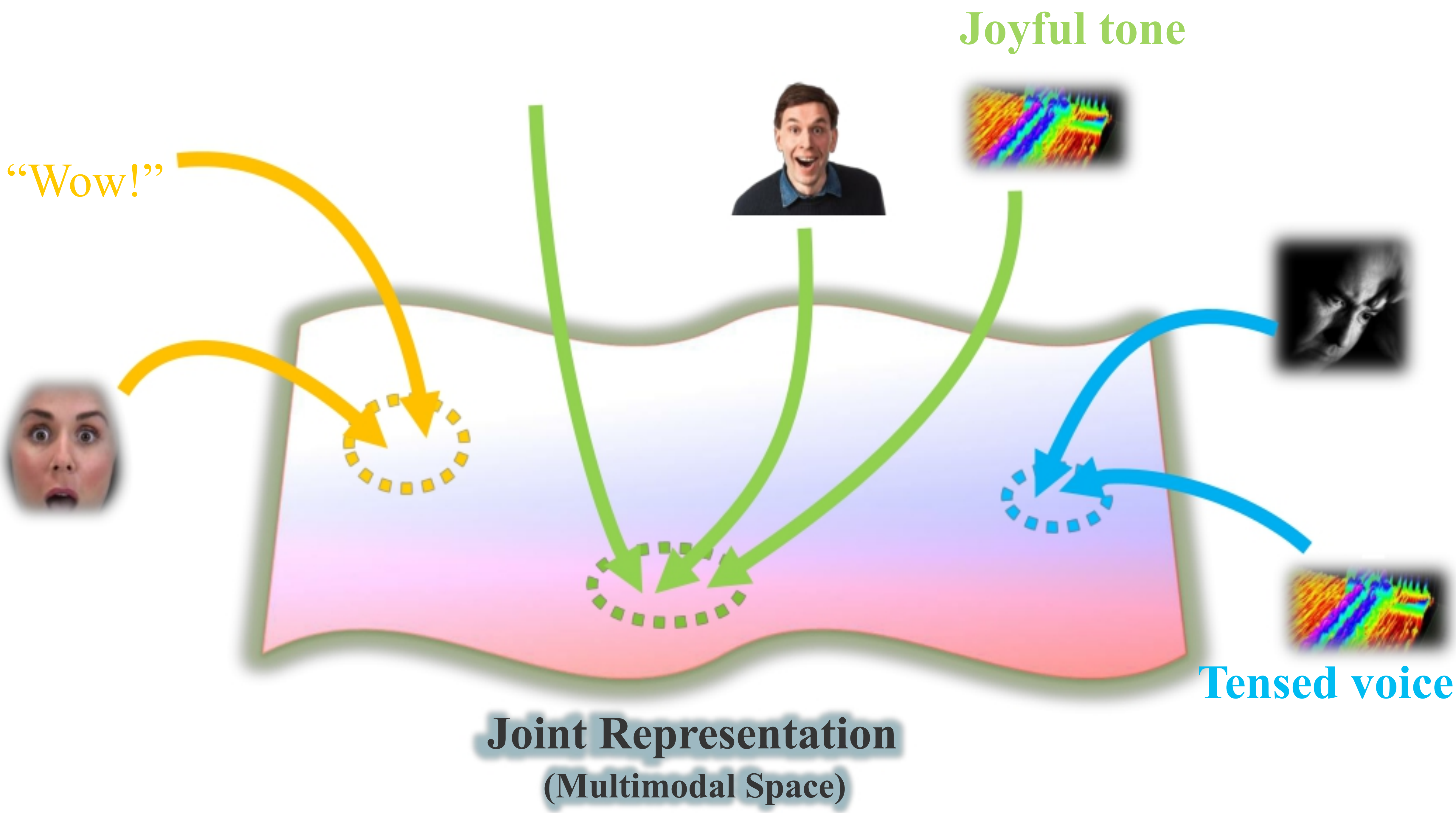
Core Challenge 1: Representation

- **Definition**: Learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy.

A Joint representations:

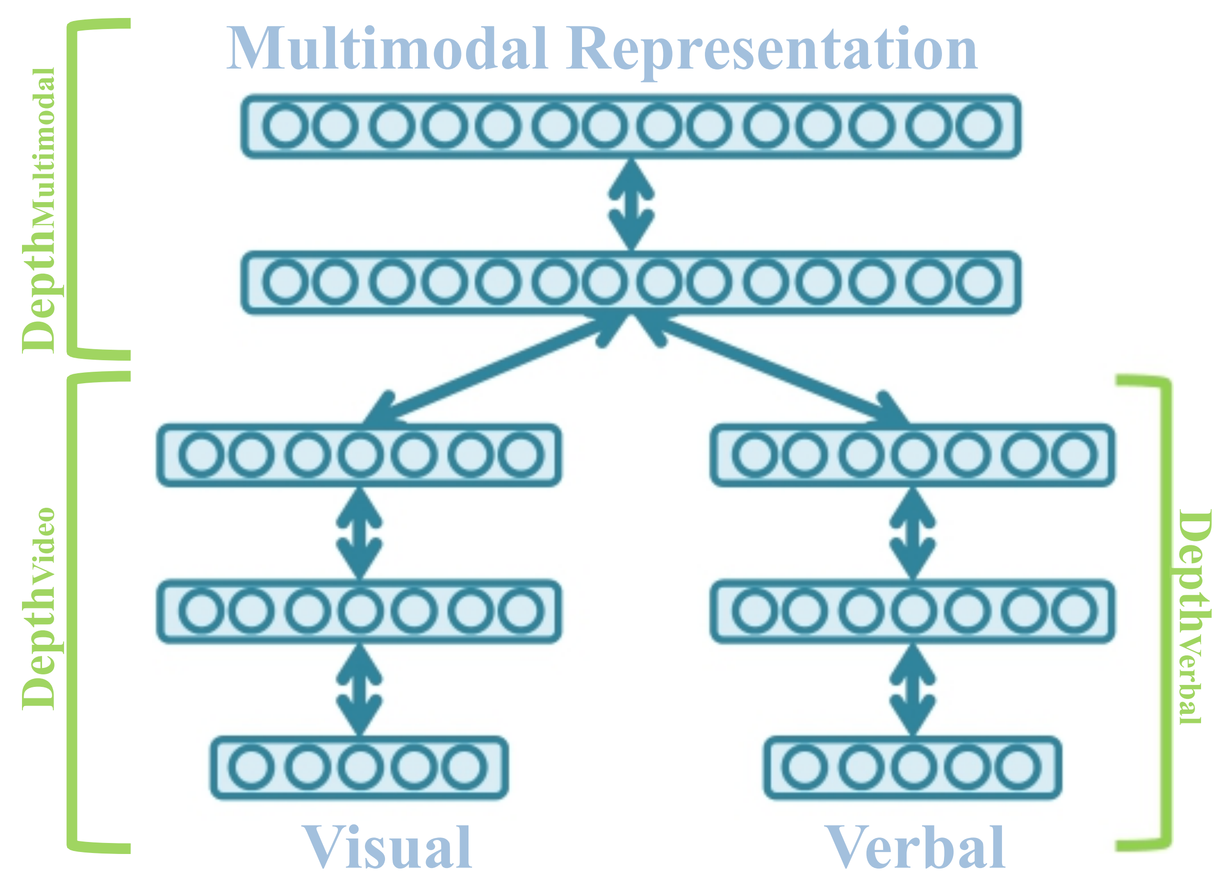


Joint Multimodal Representations



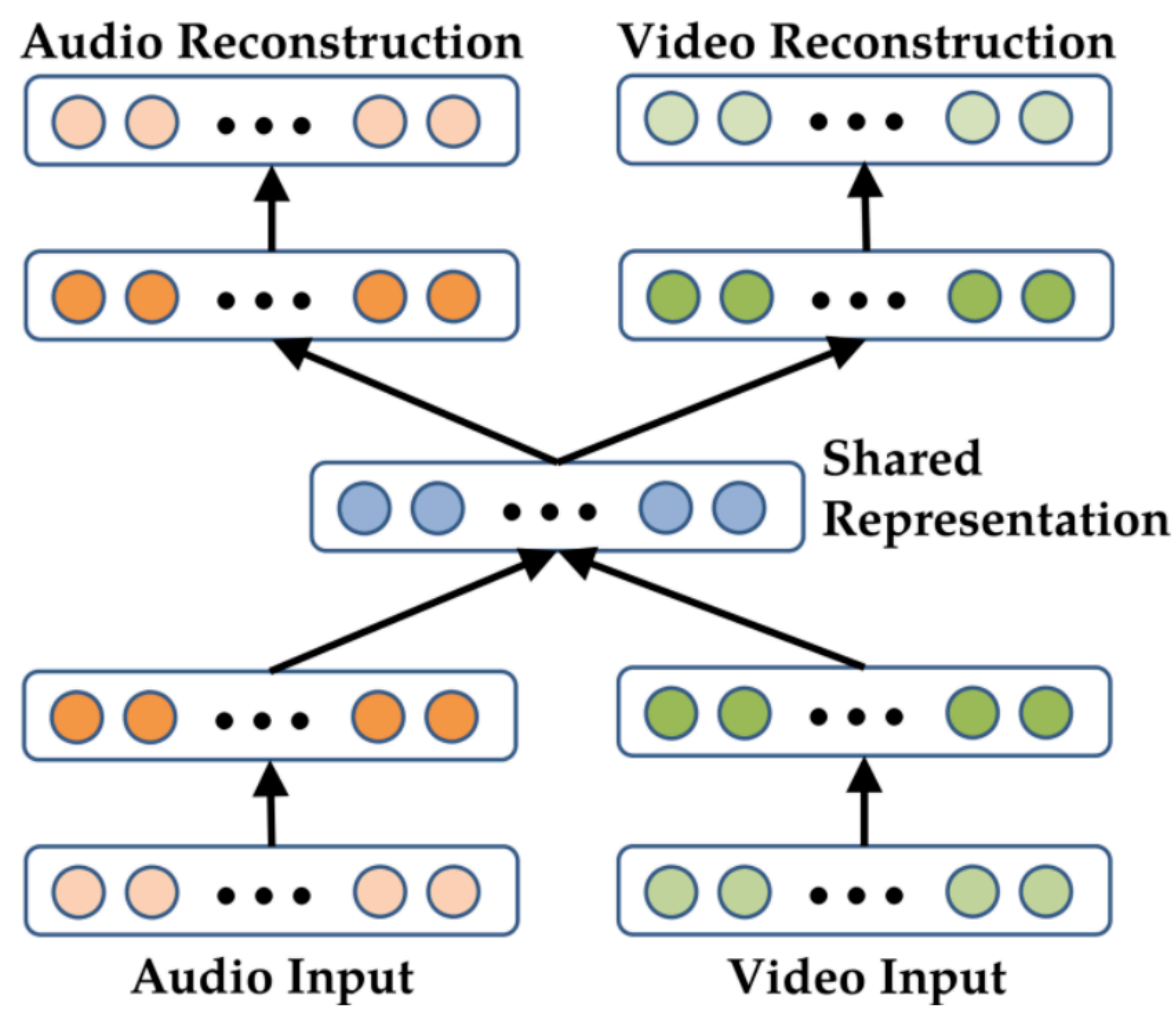
Joint Multimodal Representations

- Audio-visual speech recognition (Ngiam et al. 2011)
    - Bimodal Deep Belief Network
- Image captioning (Srivastava, Salahutdinov, 2012)
    - Multimodal Deep Boltzmann Machine
- Audio-visual emotion recognition (Kim et al. 2013)
    - Deep Boltzmann Machine



Deep Multimodal Autoencoder

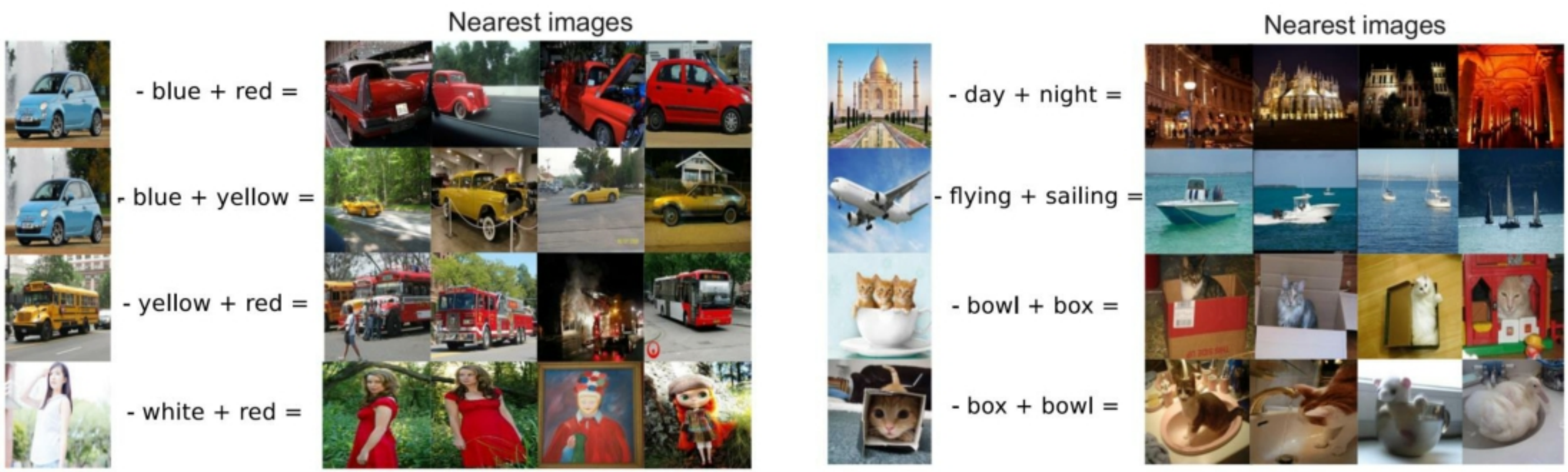
- Bimodal auto-encoder
  - Used for audio-visual speech recognition
- Individual modalities can be pretrained
  - RBMs
  - Denoising Autoencoders
- Train the model to reconstruct the other modality
  - Use both
  - Remove audio
  - Remove video



Ngiam et al., Multimodal Deep Learning, 2011.

Multimodal Vector Space Arithmetic

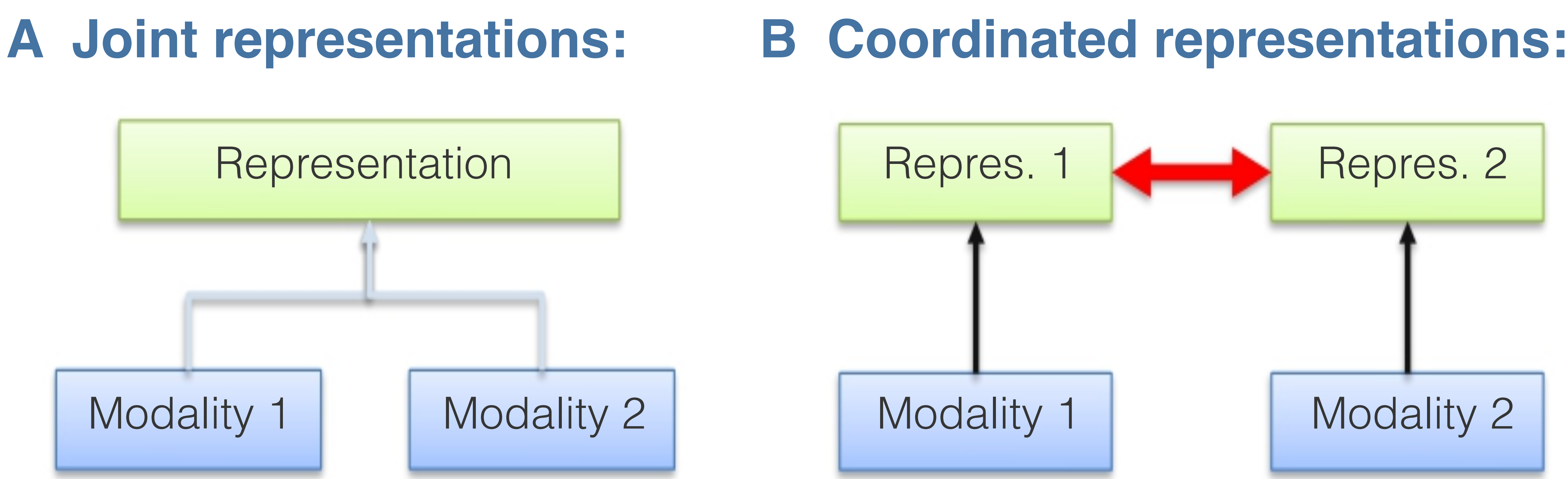
- Obtain a vector by the image embedding of a blue car - word embedding of “blue” + word embedding of “red”
- Retrieve the nearest images



Kiros et al. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language models.

Core Challenge 1: Representation

- **Definition**: Learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy.

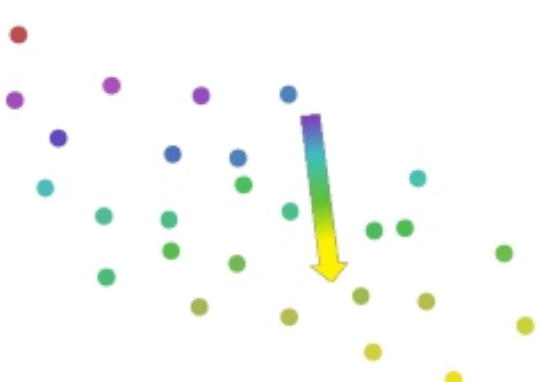


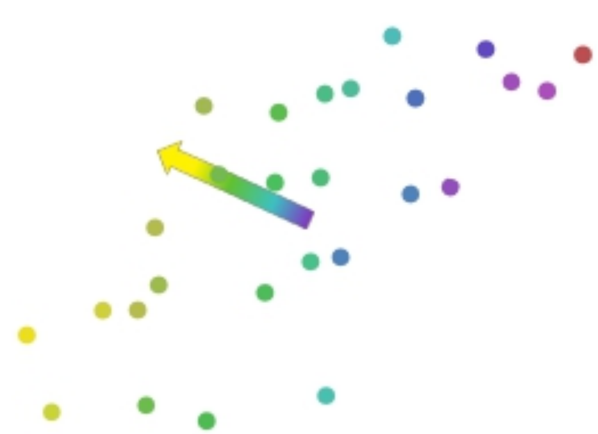
Coordinated Representation: Deep CCA

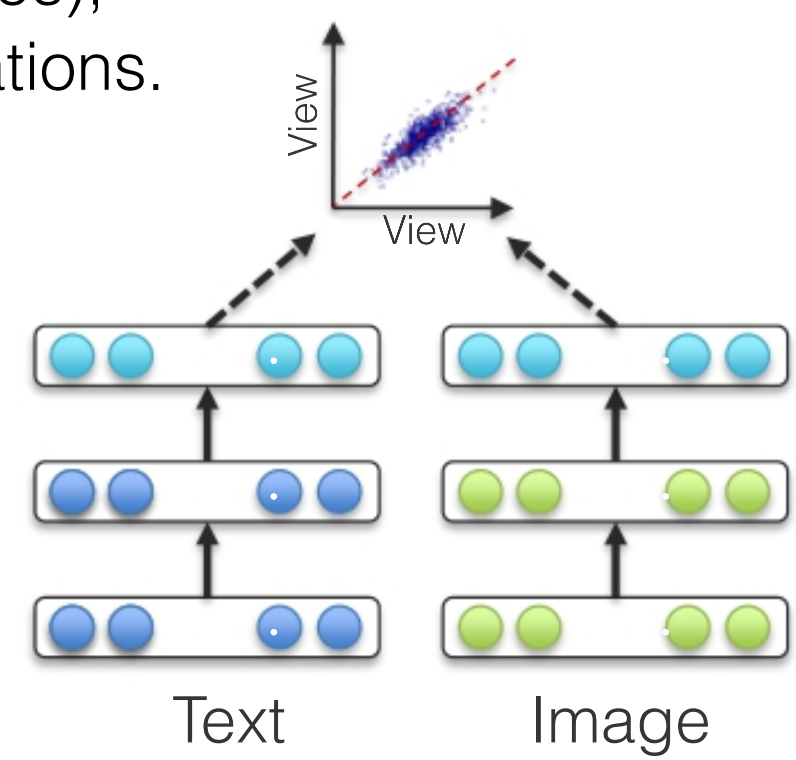
- Learn linear projections that are maximally correlated:

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1, \theta_2)}{\operatorname{argmax}} \operatorname{corr}(f_1(X_1; \theta_1), f_2(X_2; \theta_2)).$$

where f_1 and f_2 are two encoders (e.g., for texts, images), corr computes the correlation between two representations.

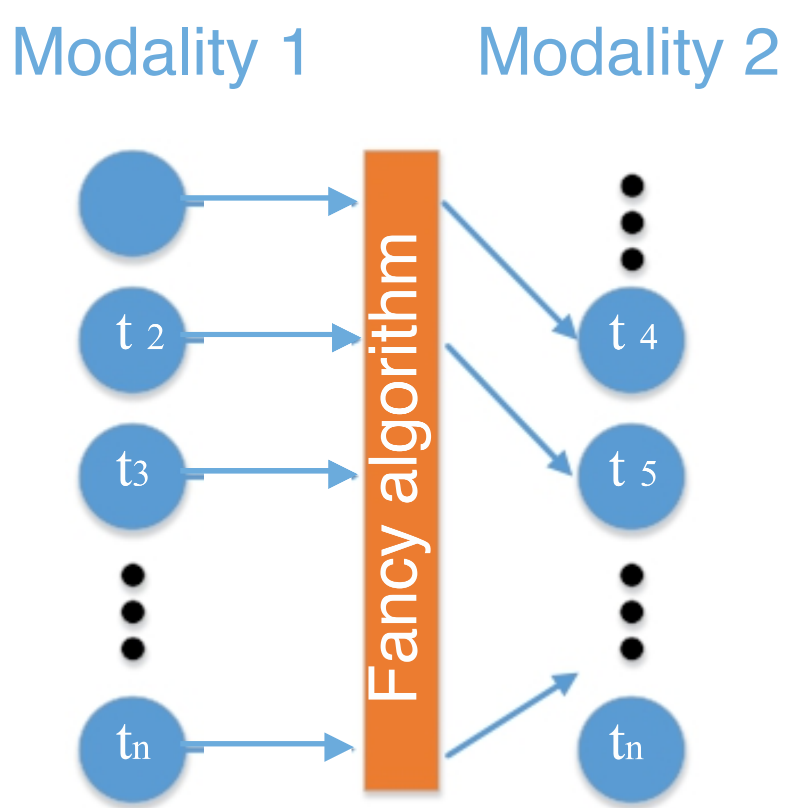






Core Challenge 2: Alignment

- Definition: Identify the direct relations between (sub)elements from two or more different modalities



A Explicit Alignment

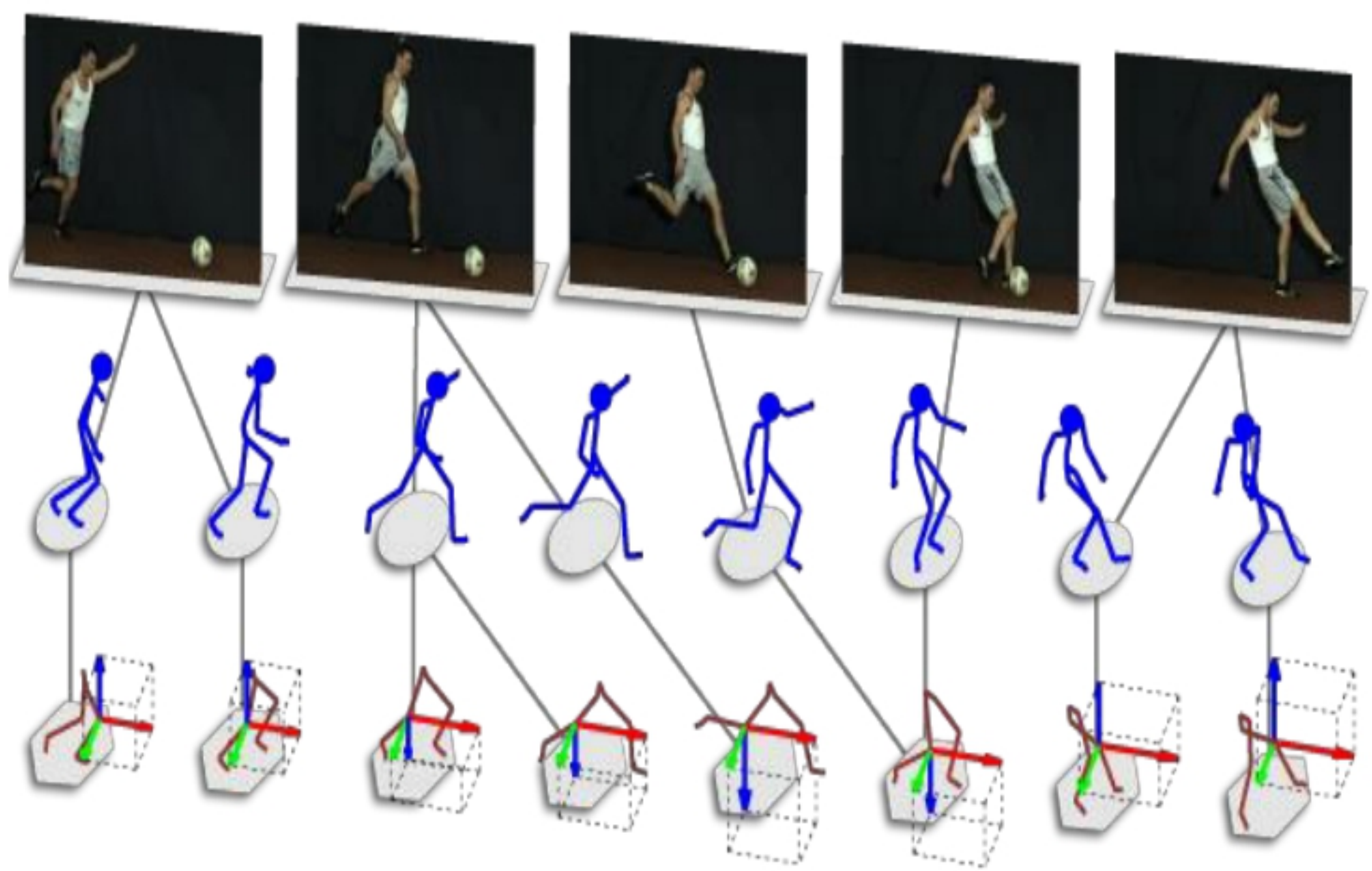
The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

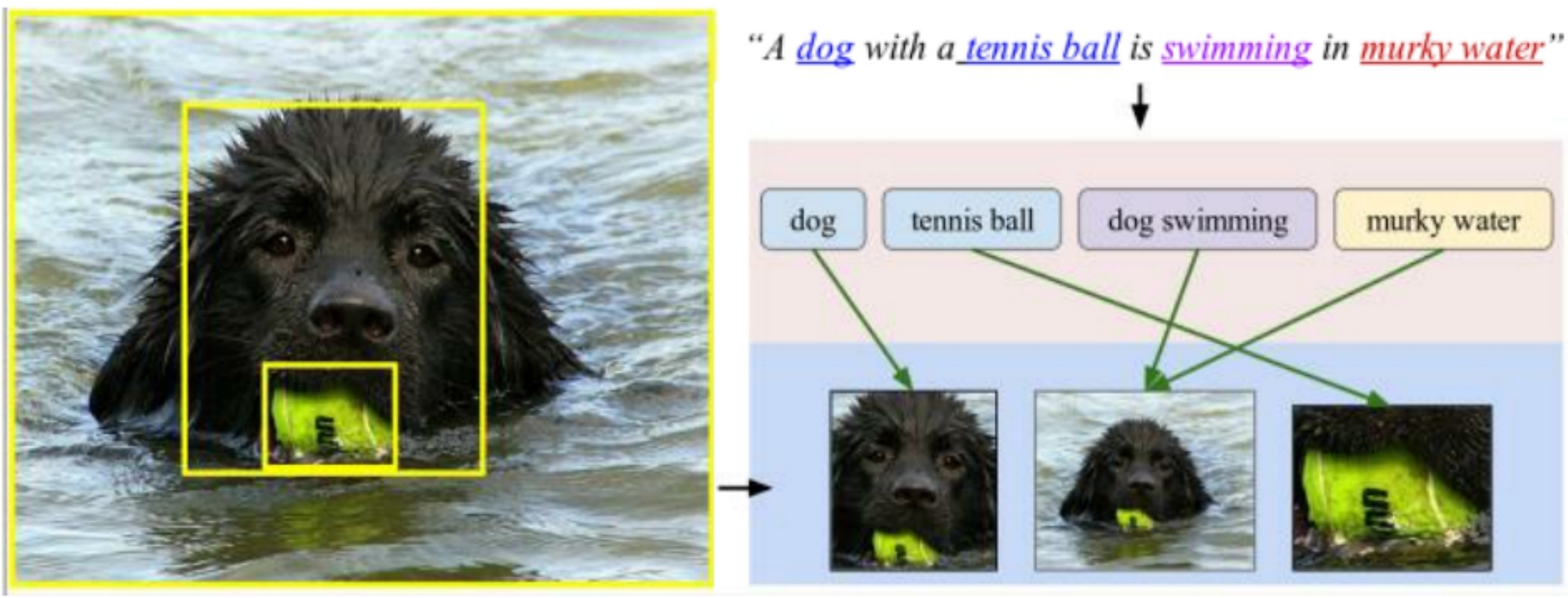
Example: Temporal Sequence Alignment

- Application:
  - Re-aligning asynchronous data
  - Finding similar data across modalities
  - Event reconstruction from multiple sources



Implicit Alignment

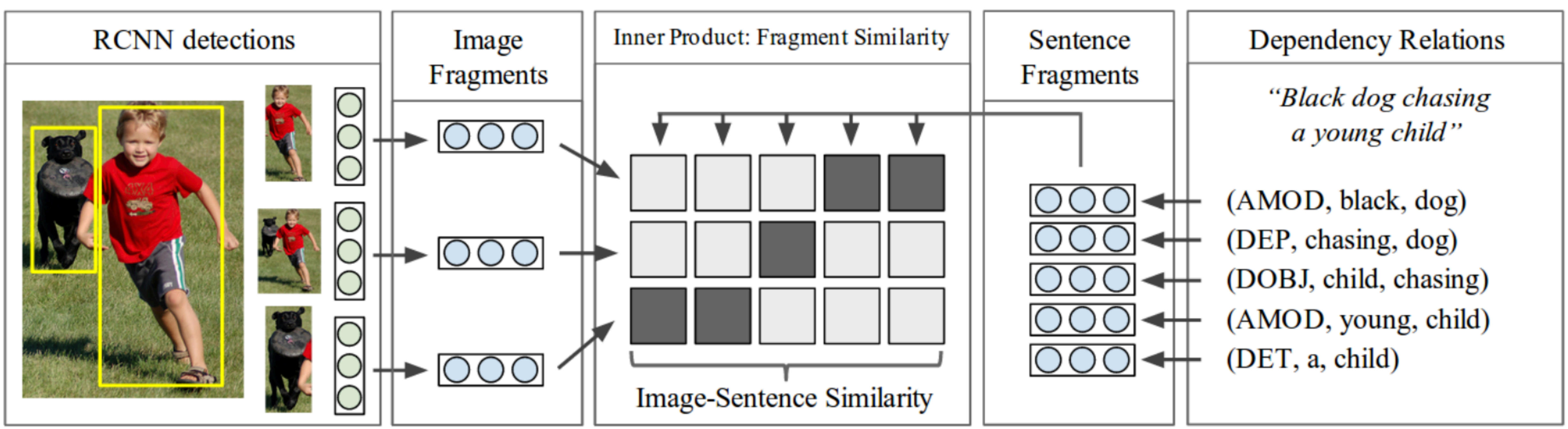
- Vision-language alignment, a.k.a. visual grounding.



Karpathy et al. 2014 Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Implicit Alignment

- Use object detection (RCNN) tools to extract bounding boxes, and encode each bounding box
- Use dependency parsing to extract dependency relations (Relation-head-tail triple), and encode each relation
- Compute the similarity and optimize the alignment objectives.

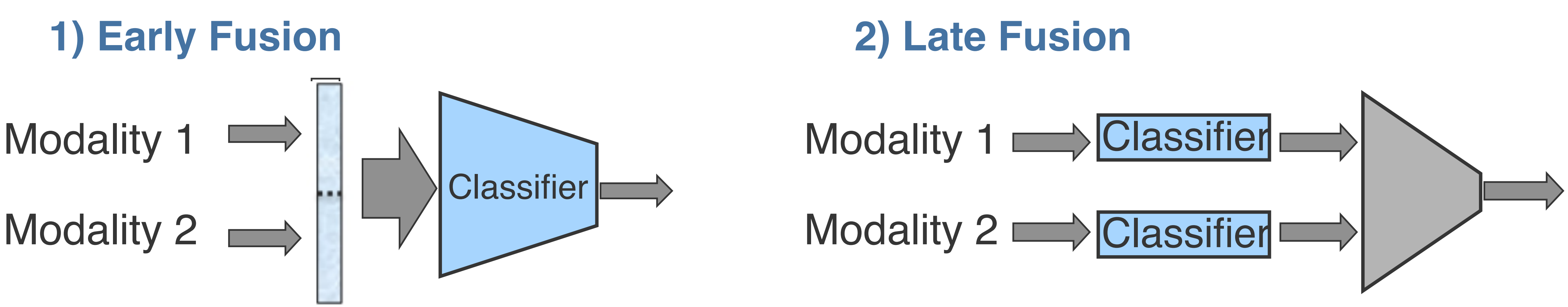


Karpathy et al. 2014 Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Core Challenge 3: Fusion

- **Definition**: To join information from two or more modalities to perform a prediction task.

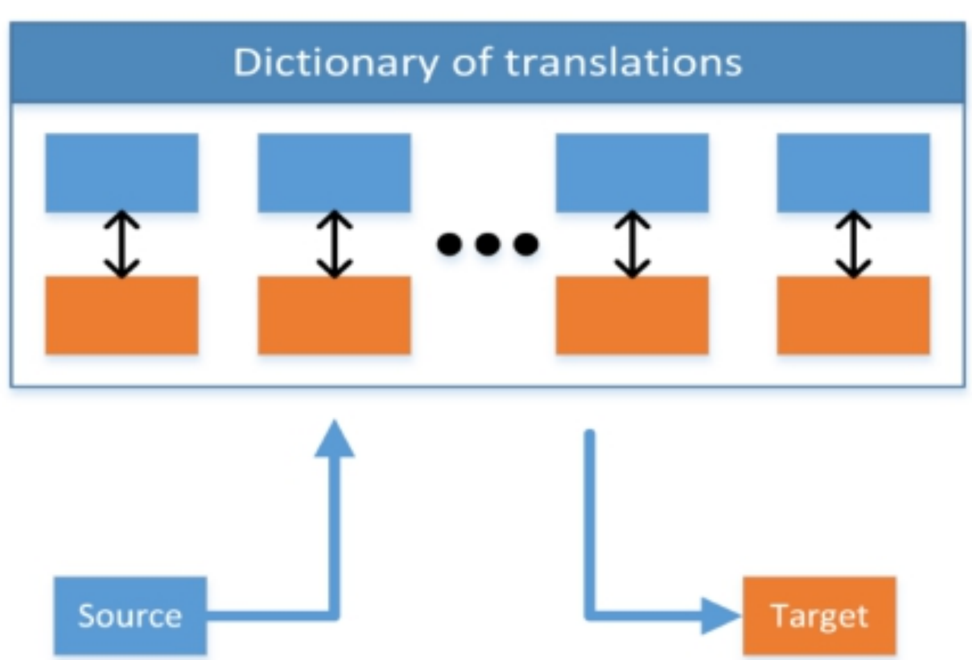
A Model-Agnostic Approaches



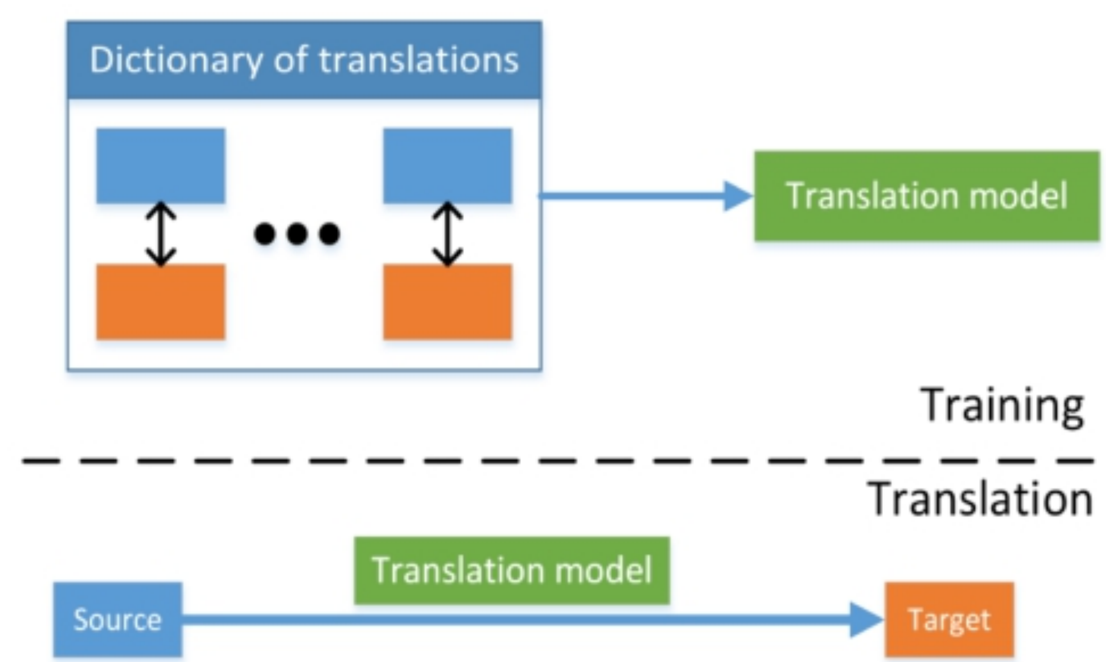
Core Challenge 4: Translation

- **Definition**: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective

A Example-based



A Model-based



Text+Audio to Vision Translation

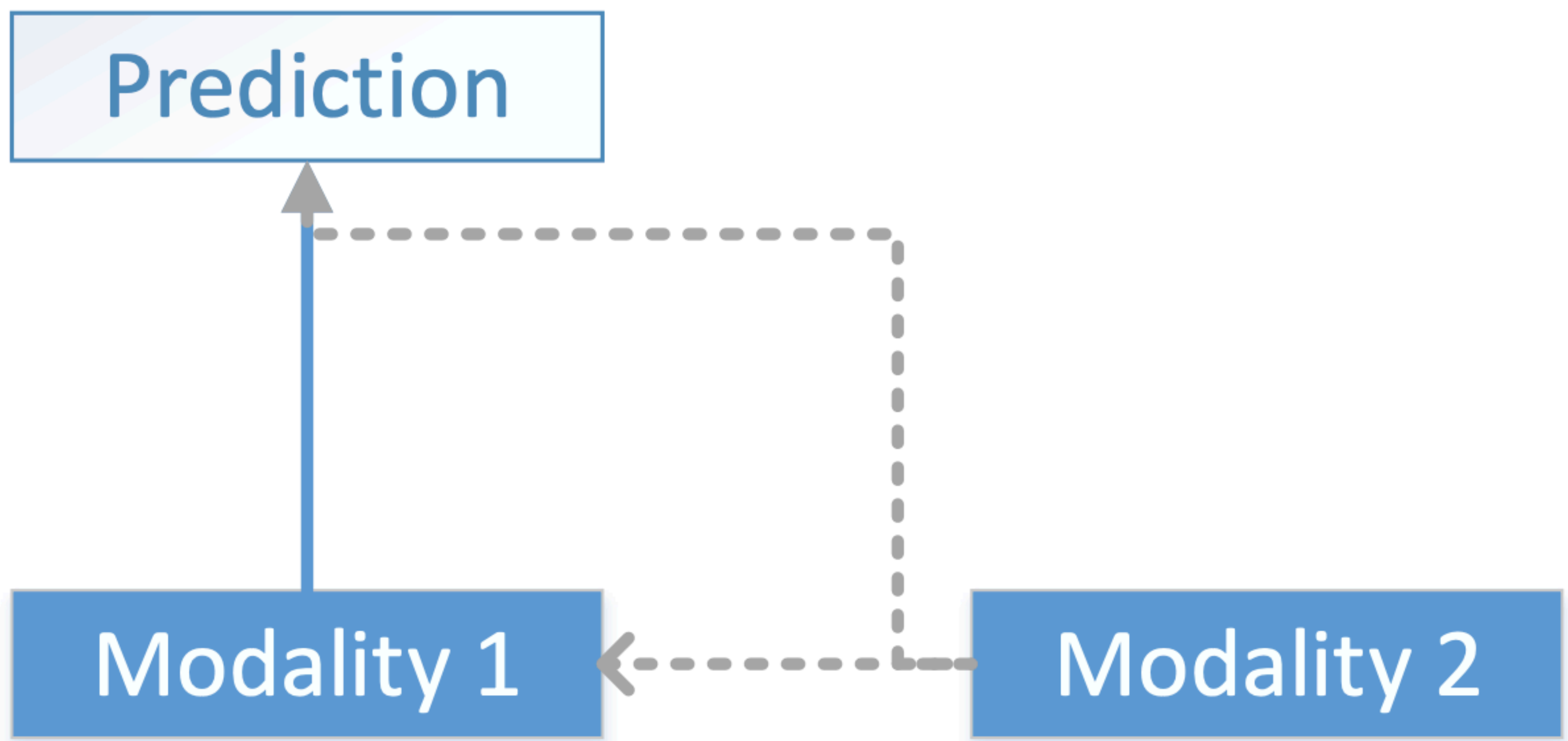
Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

Marsella et al., Virtual character performance from speech, SIGGRAPH/ Eurographics Symposium on Computer Animation, 2013

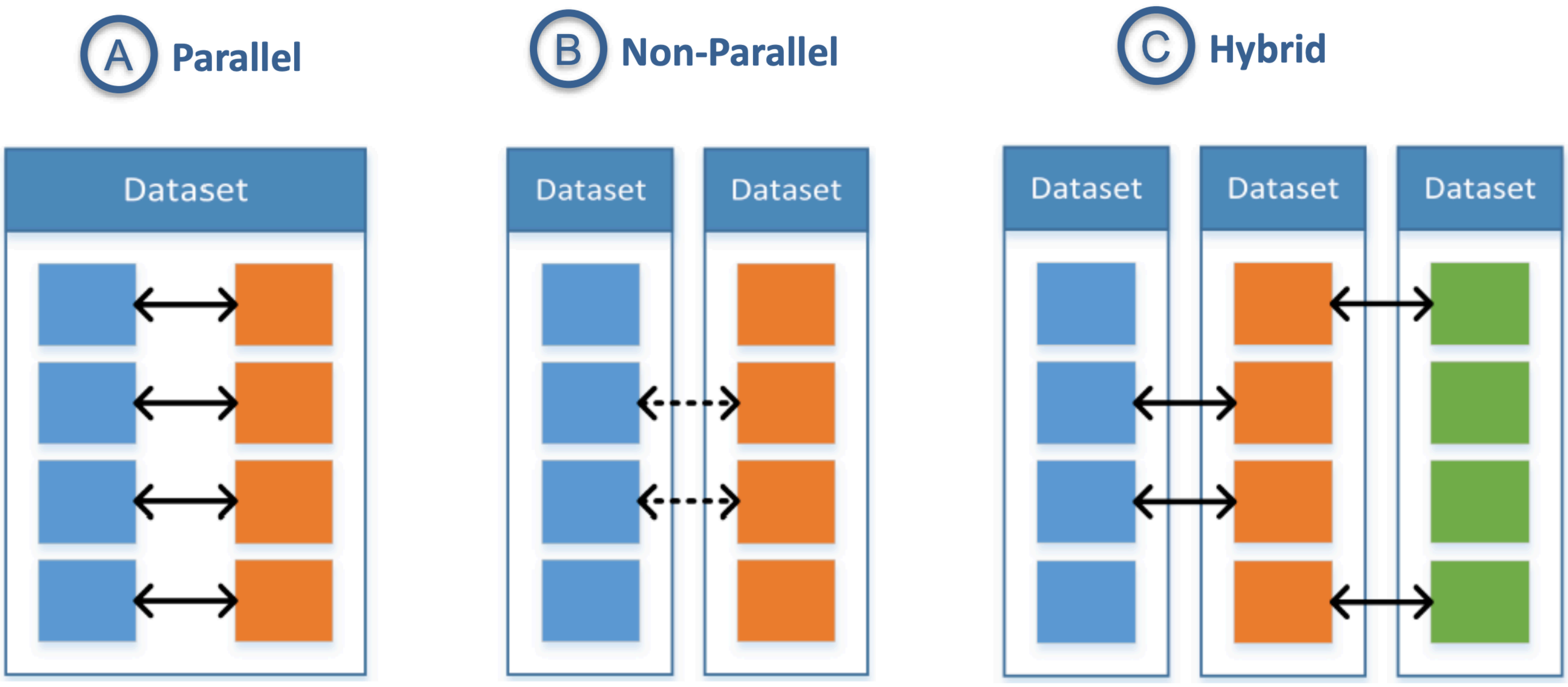
Core Challenge 5: Co-Learning

- **Definition**: Transfer knowledge between modalities, including their representations and predictive models.



Core Challenge 5: Co-Learning

- Three data settings.



Taxonomy of Multimodal Research

Representation

- Joint
  - *Neural networks*
  - *Graphical models*
  - *Sequential*
- Coordinated
  - *Similarity*
  - *Structured*

Translation

- Example-based
  - *Retrieval*
  - *Combination*
- Model-based
  - *Grammar-based*
  - *Encoder-decoder*
  - *Online prediction*

Alignment

- Explicit
  - *Unsupervised*
  - *Supervised*
- Implicit
  - *Graphical models*
  - *Neural networks*

Fusion

- Model agnostic
  - *Early fusion*
  - *Late fusion*
  - *Hybrid fusion*
- Model-based
  - *Kernel-based*
  - *Graphical models*
  - *Neural networks*

Co-learning

- Parallel data
  - *Co-training*
  - *Transfer learning*
- Non-parallel data
  - *Zero-shot learning*
  - *Concept grounding*
  - *Transfer learning*
- *Hybrid data*
  - *Bridging*

Baltrusaitis et al. 2017. Multimodal Machine Learning: A Survey and Taxonomy

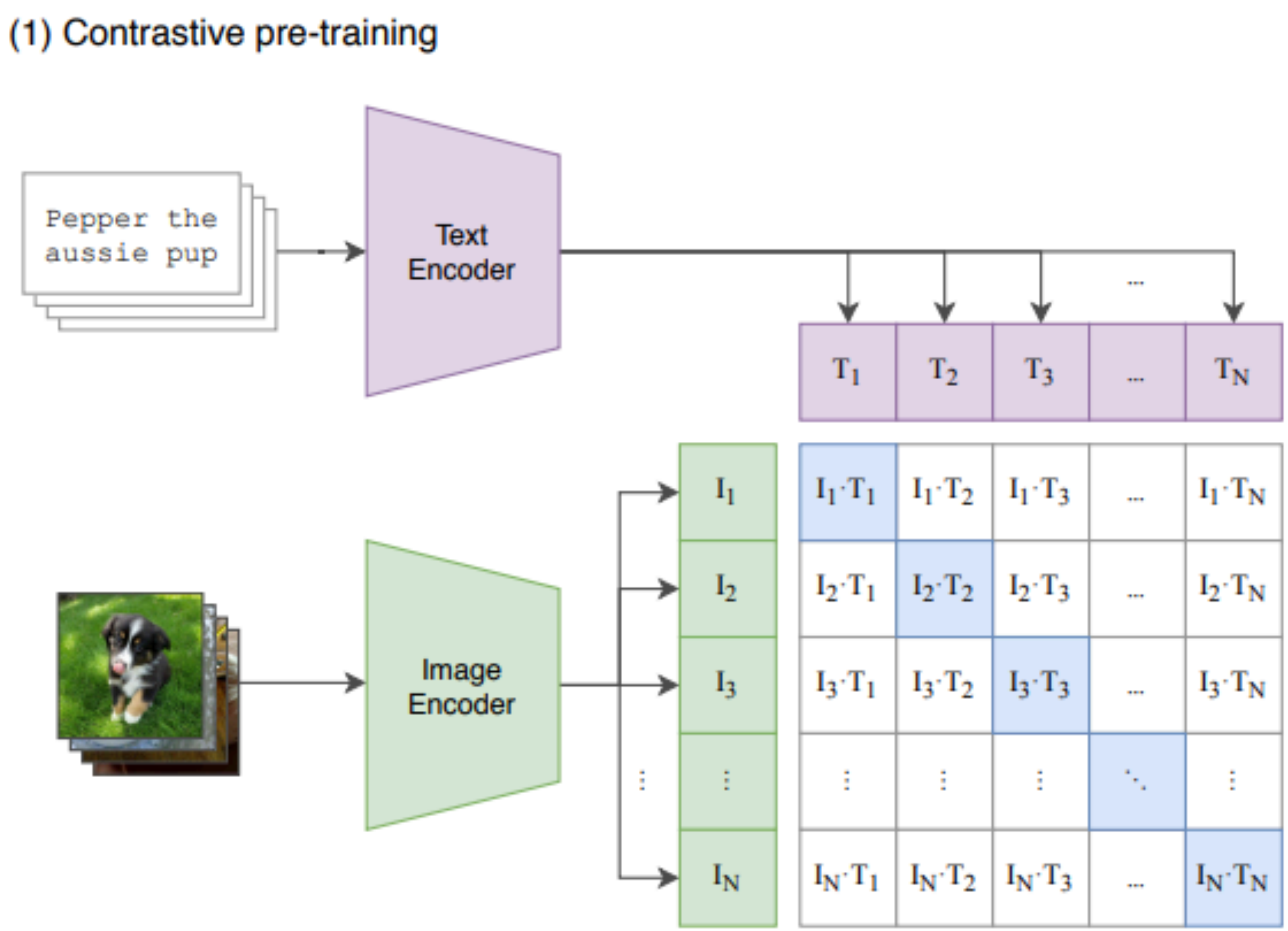
Multimodal Applications

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

Recent Pre-trained Vision-Language Models

CLIP

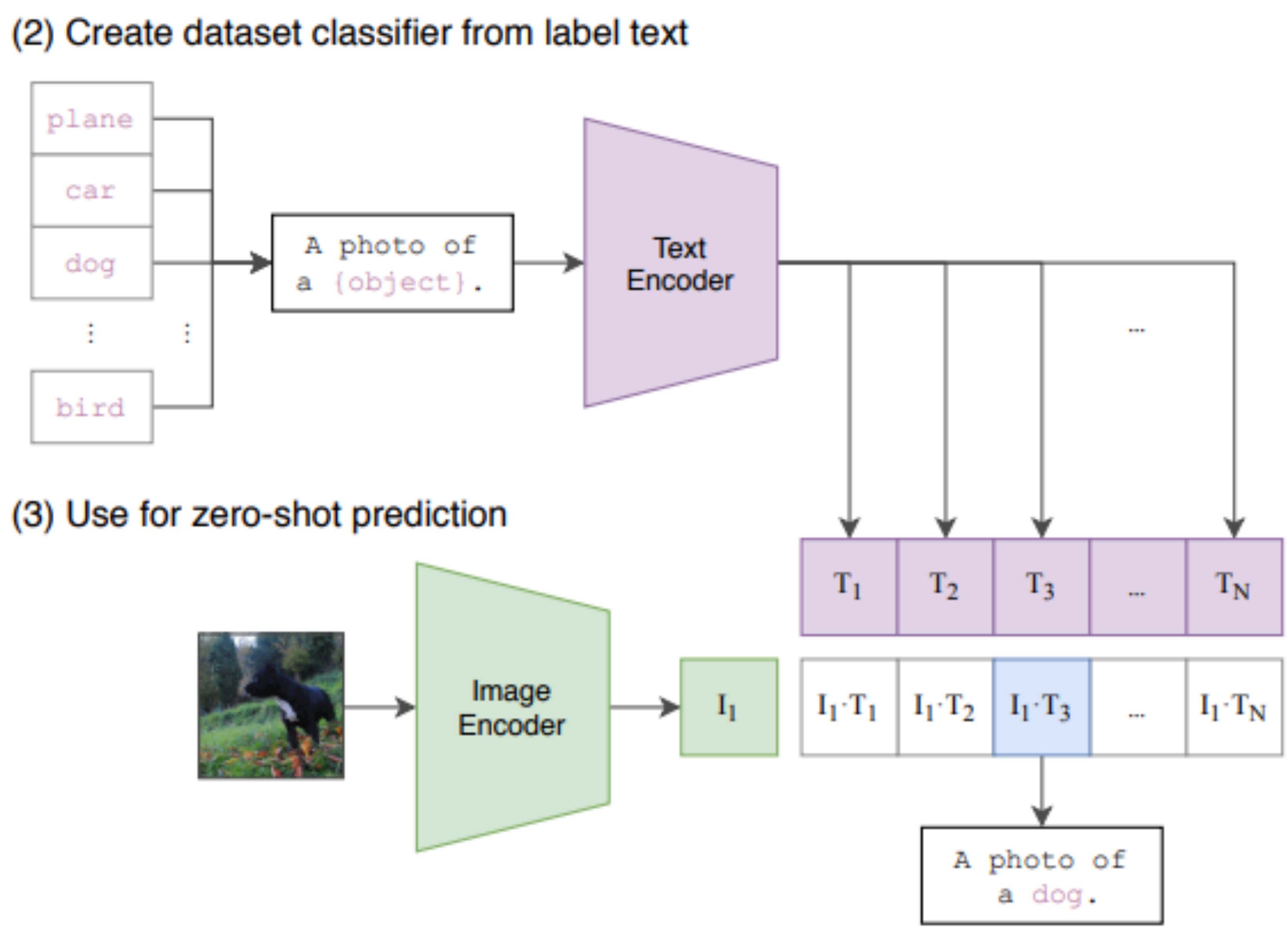
- Pre-train V+L models using image captioning data (i.e., image-text pairs) by contrastive loss



Radford et al. 2021 Learning Transferable Visual Models From Natural Language Supervision

CLIP: Zero-shot Image Classification

- Use a template + class label string to create a sentence

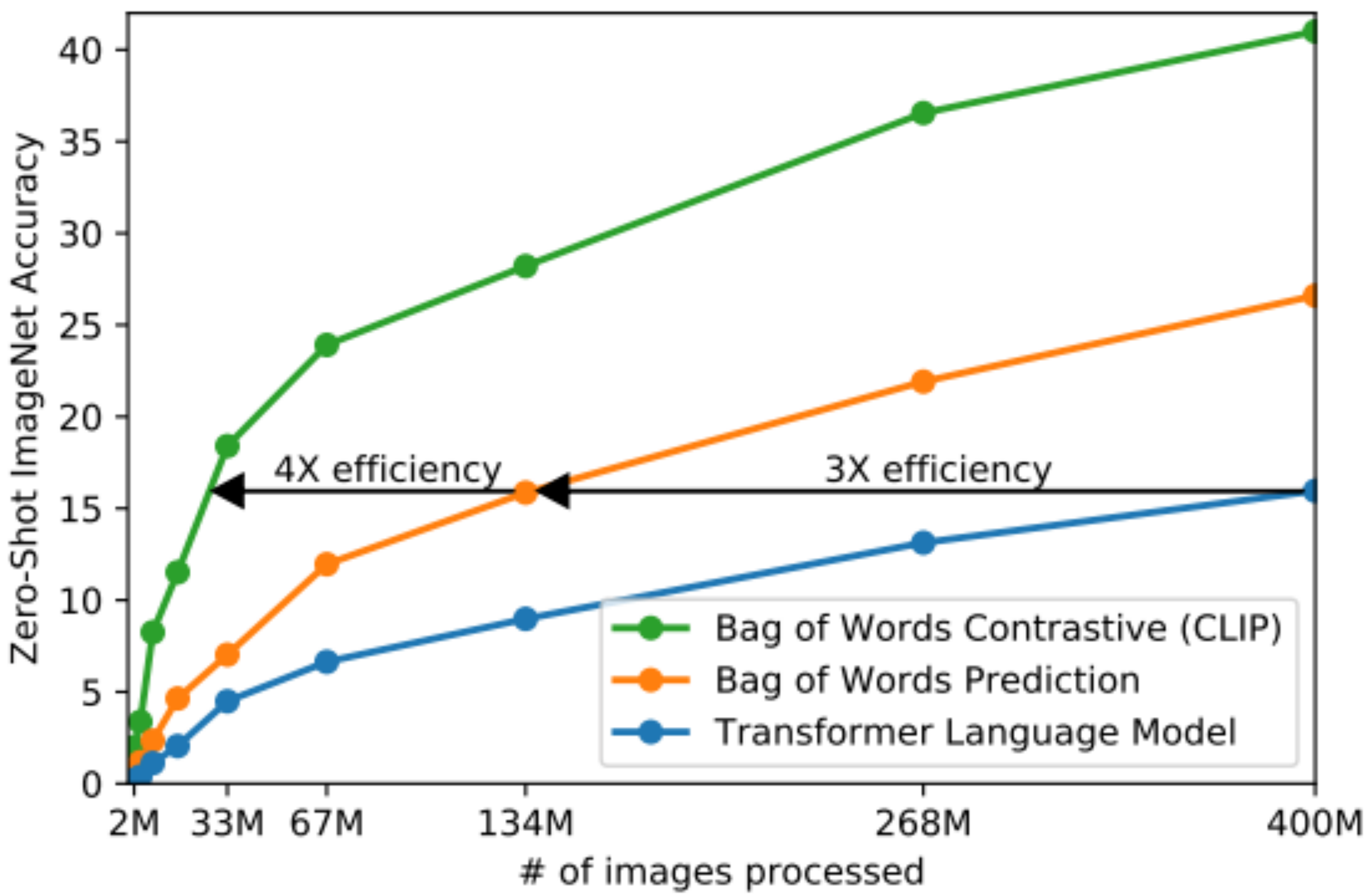


CLIP: pseudocode

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

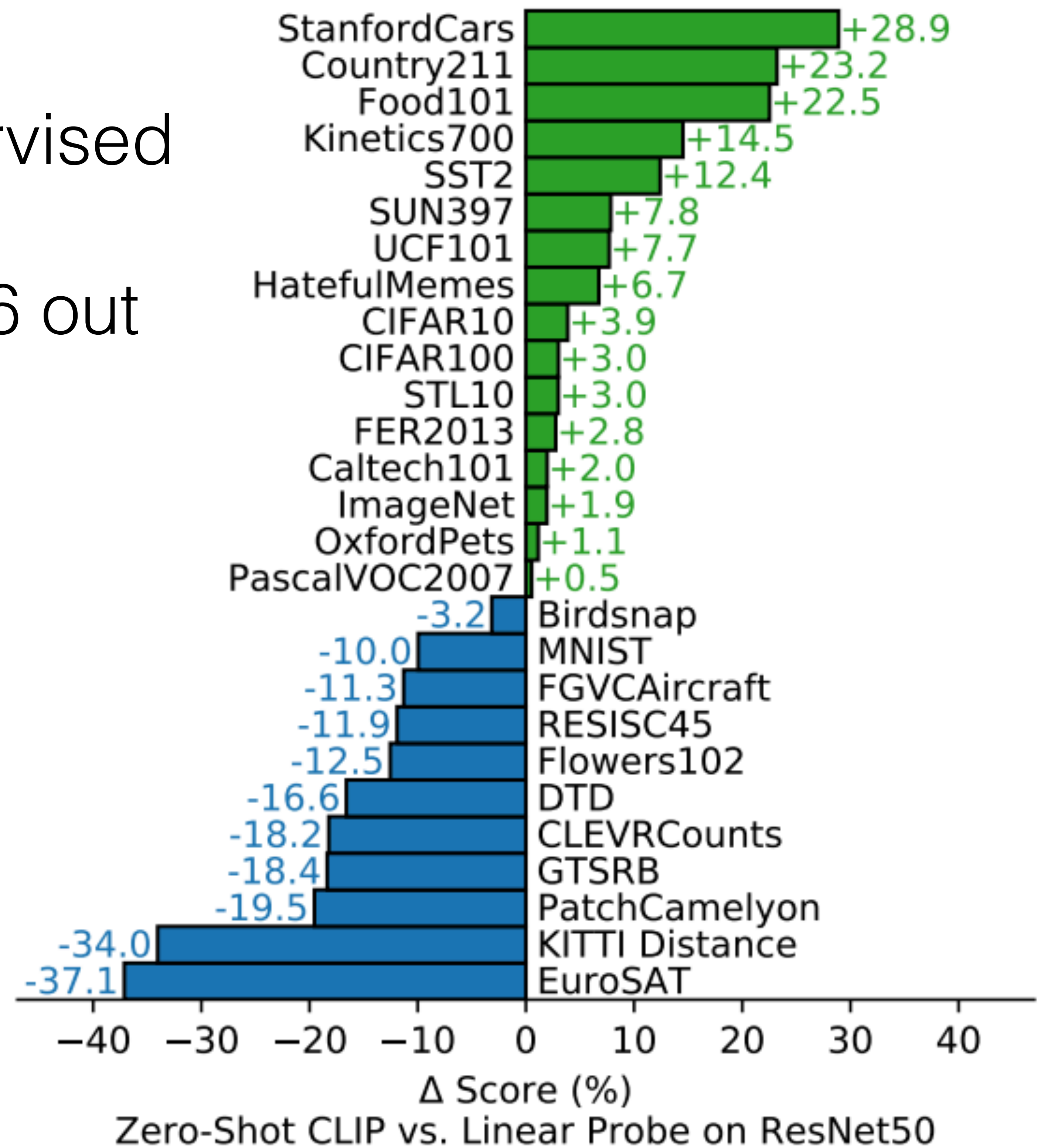
Efficiency of BoW Representations

- CLIP w/ BoW representations work better than transformer language model on zero-shot ImageNet prediction



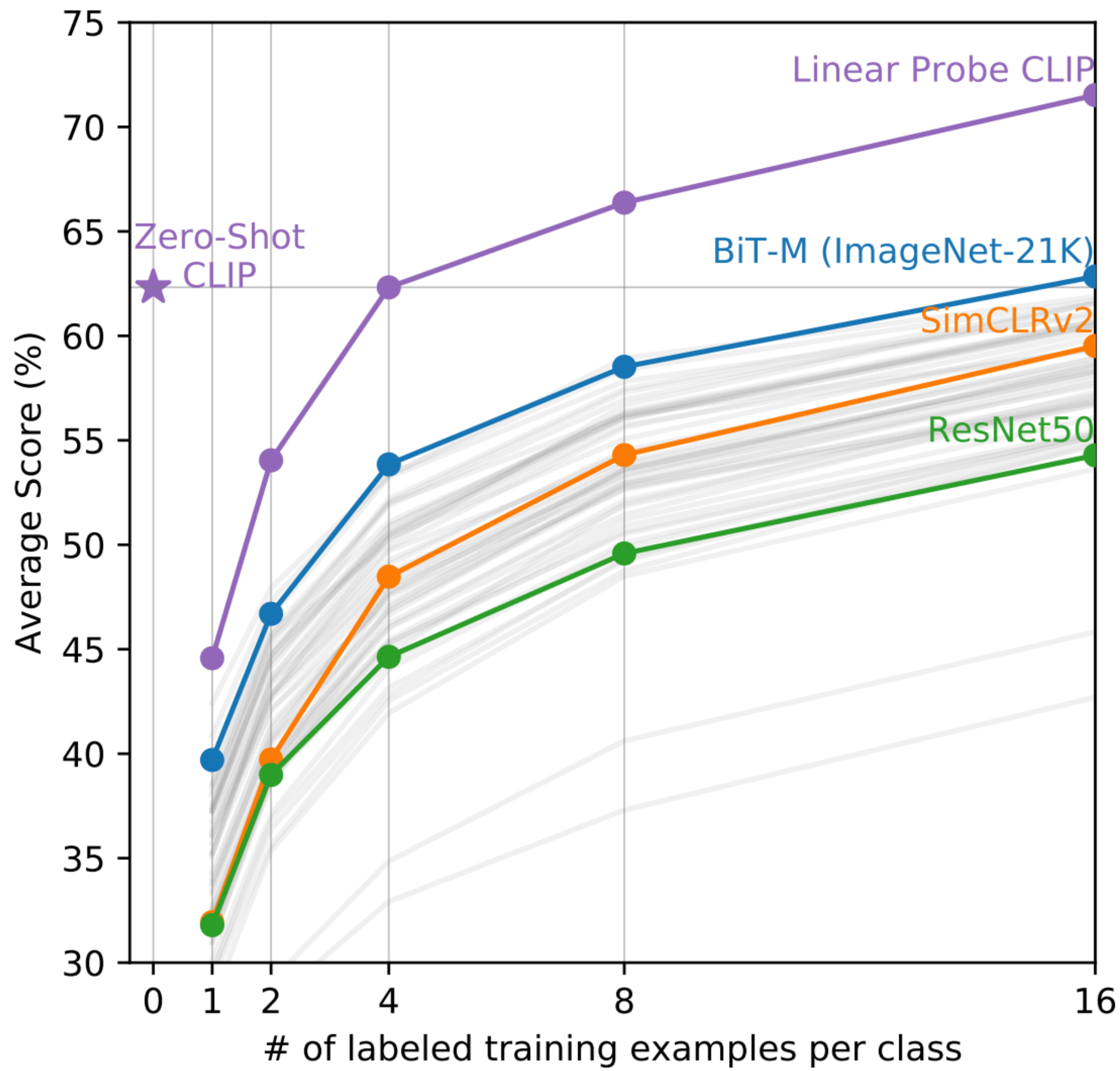
Zero-shot Image Classification

- Zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 out of 27 datasets (including ImageNet).



Few-shot Performance

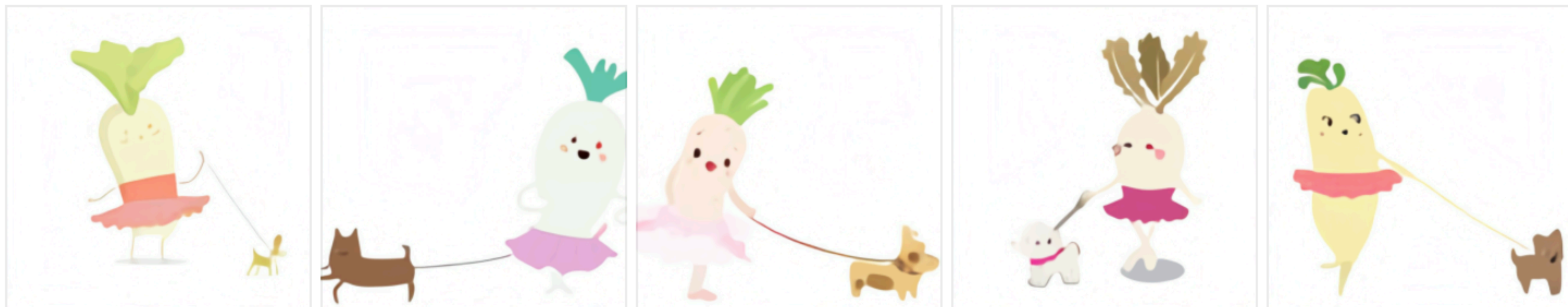
- Zero-shot CLIP outperforms other few-shot baselines
- Few-shot CLIP further improves w/ a few labeled data.



DALL-E: Text-to-Image Generation

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

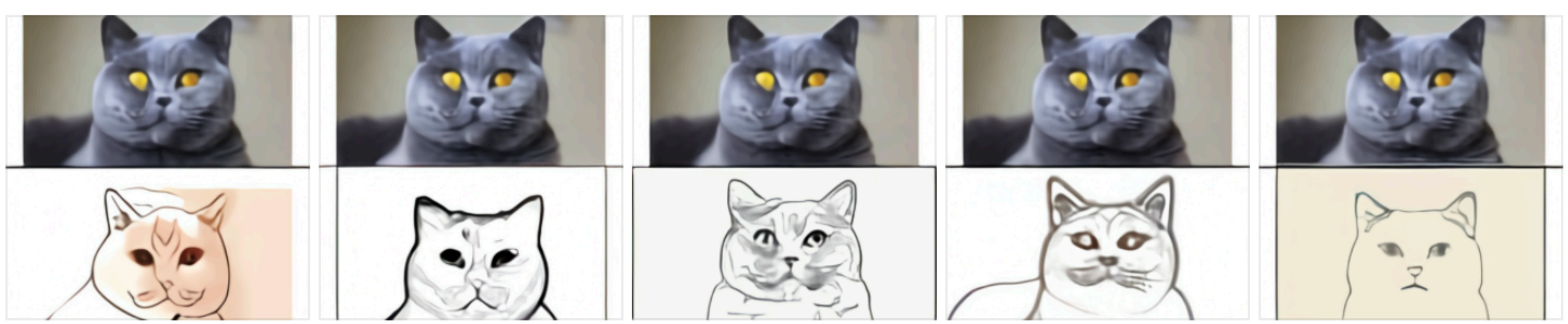


TEXT PROMPT a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

TEXT & IMAGE PROMPT the exact same cat on the top as a sketch on the bottom

AI-GENERATED IMAGES



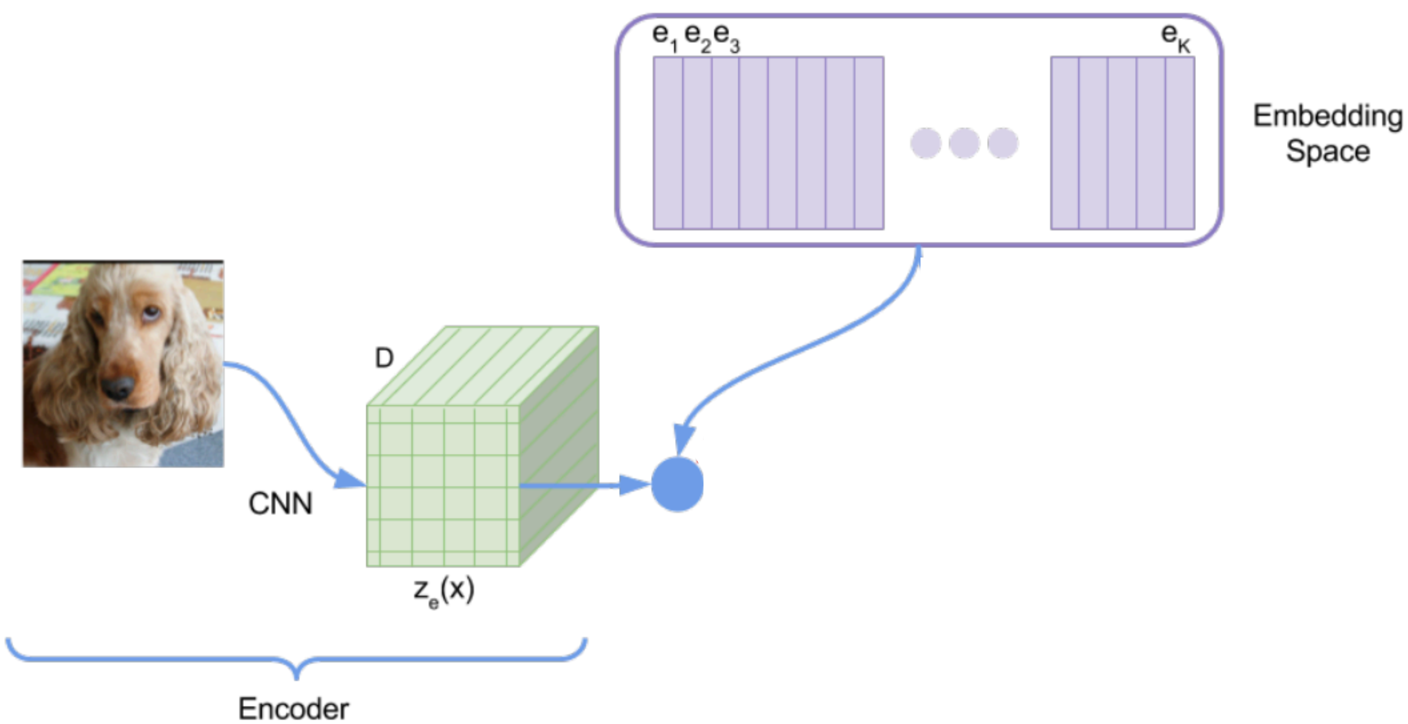
DALL-E

- **Stage 1**: Train a discrete VAE on **only images** (encode RGB images to image tokens (latent variable), and decode image tokens back to RGB images)
- **Stage 2**: Train a language model (LM) to generate a combined sequence of **both text tokens and image tokens**

DALL-E: dVAE Training

- **Stage 1**: Train a discrete variational autoencoder (dVAE or VQ-VAE, Oord et al. 2018) to compress each 256x256 RGB image into 32x32 grid of image tokens.

Each image token finds the nearest vector from a 8196 codebook (vocabulary)

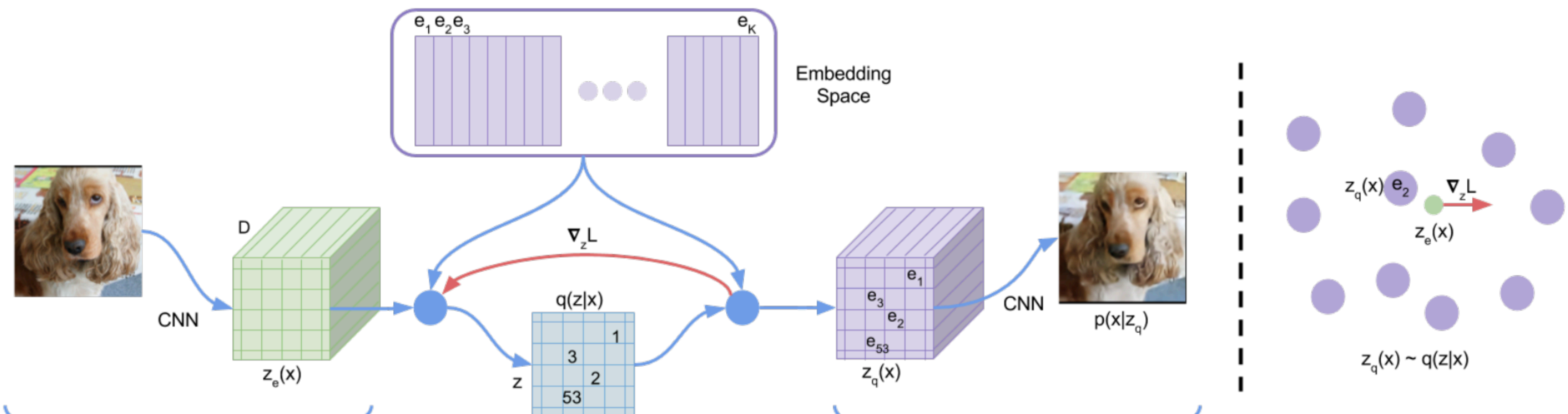


Ramesh et al. 2021. Zero-shot Text-to-Image Generation
Oord et al. 2018. Neural Discrete Representation Learning

DALL-E: dVAE Training

- **Stage 1**: Train a discrete variational autoencoder (dVAE or VQ-VAE, Oord et al. 2018) to compress each 256x256 RGB image into 32x32 grid of image tokens.

Each image token finds the nearest vector from a 8196 codebook (vocabulary)

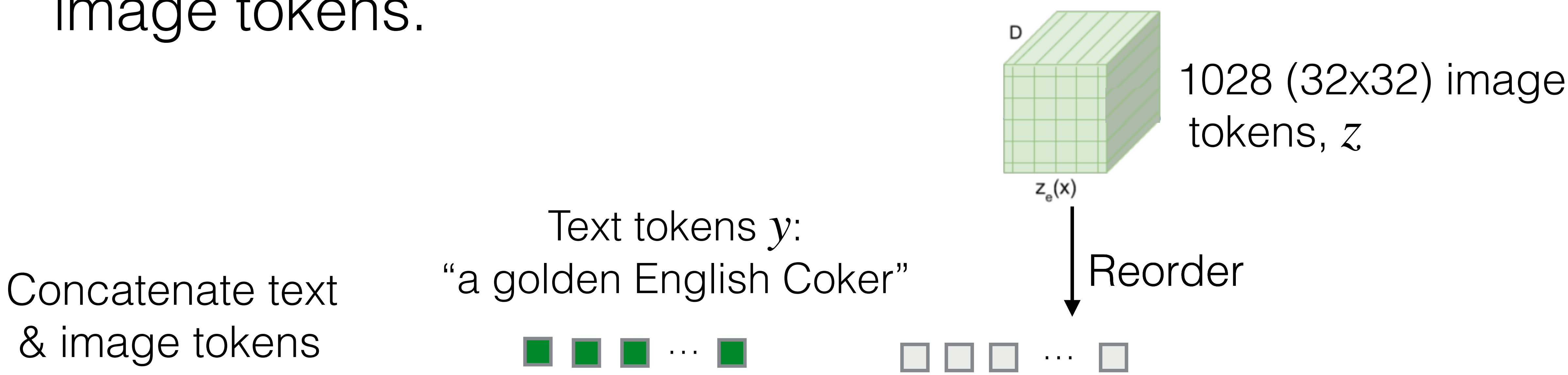


Discrete latent variables:
Gumbel-softmax — c.f. LVM lecture

VAE training:
Maximize Evidence Lower Bound

DALL-E: Language Model Training

- **Stage 2**: Concatenate up to 256 text tokens with the 32x32 (=1024) image tokens, and train an autoregression transformer to model the joint distribution of the text and image tokens.



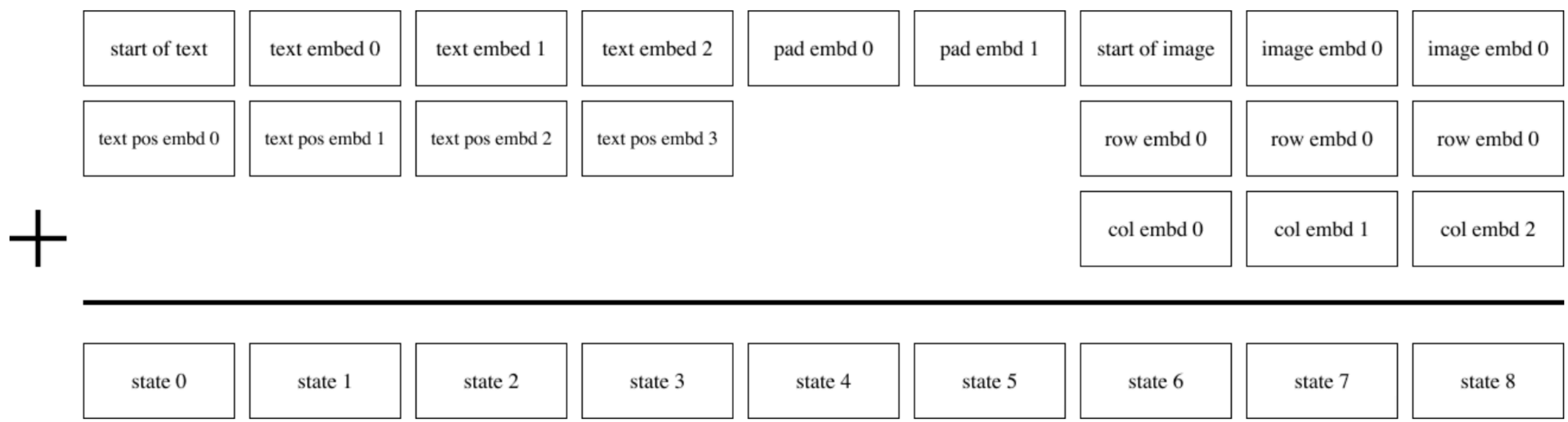
Autoregressive LM training:
Maximum Likelihood Estimation

$$\max_{\psi} p_{\psi}(y, z)$$

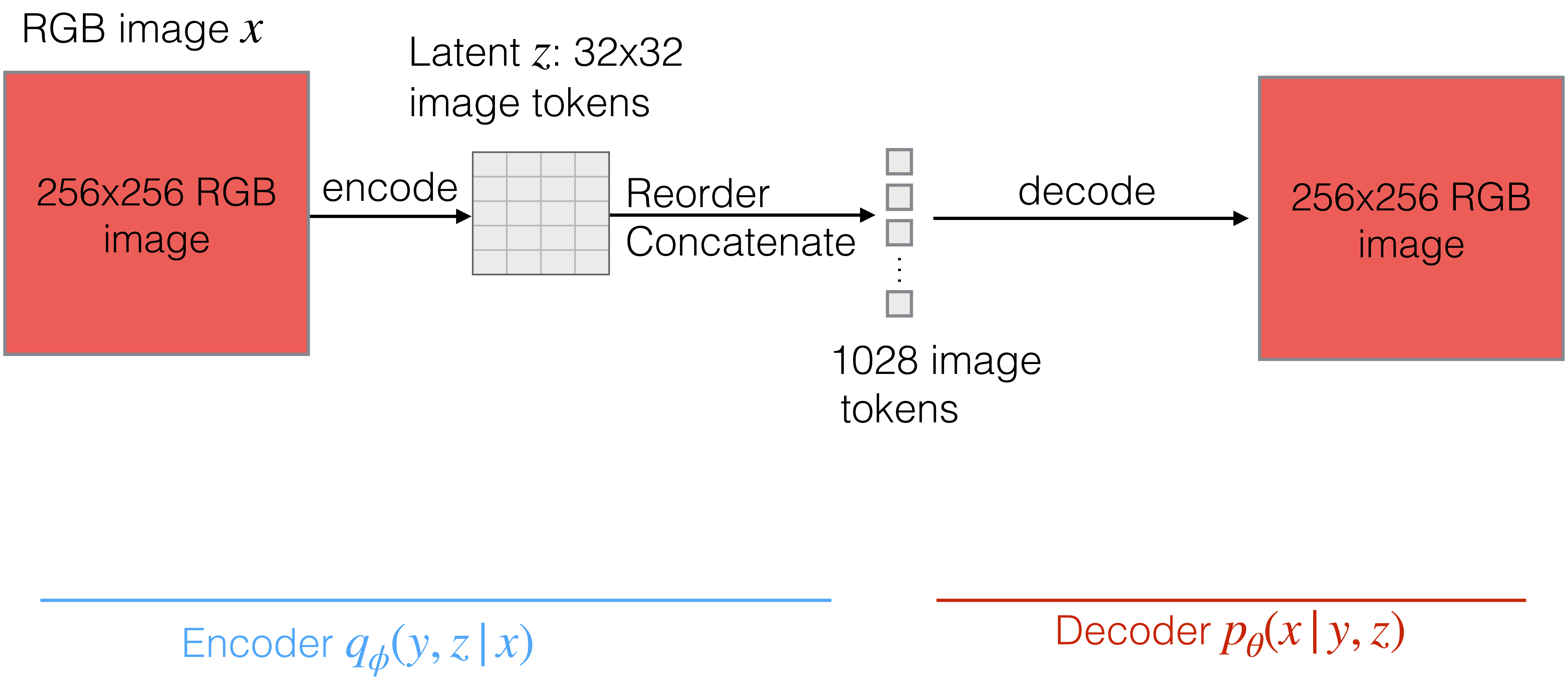
Ramesh et al. 2021. Zero-shot Text-to-Image Generation

DALL-E: Language Model Training

- Representation of the combined text + image token sequence

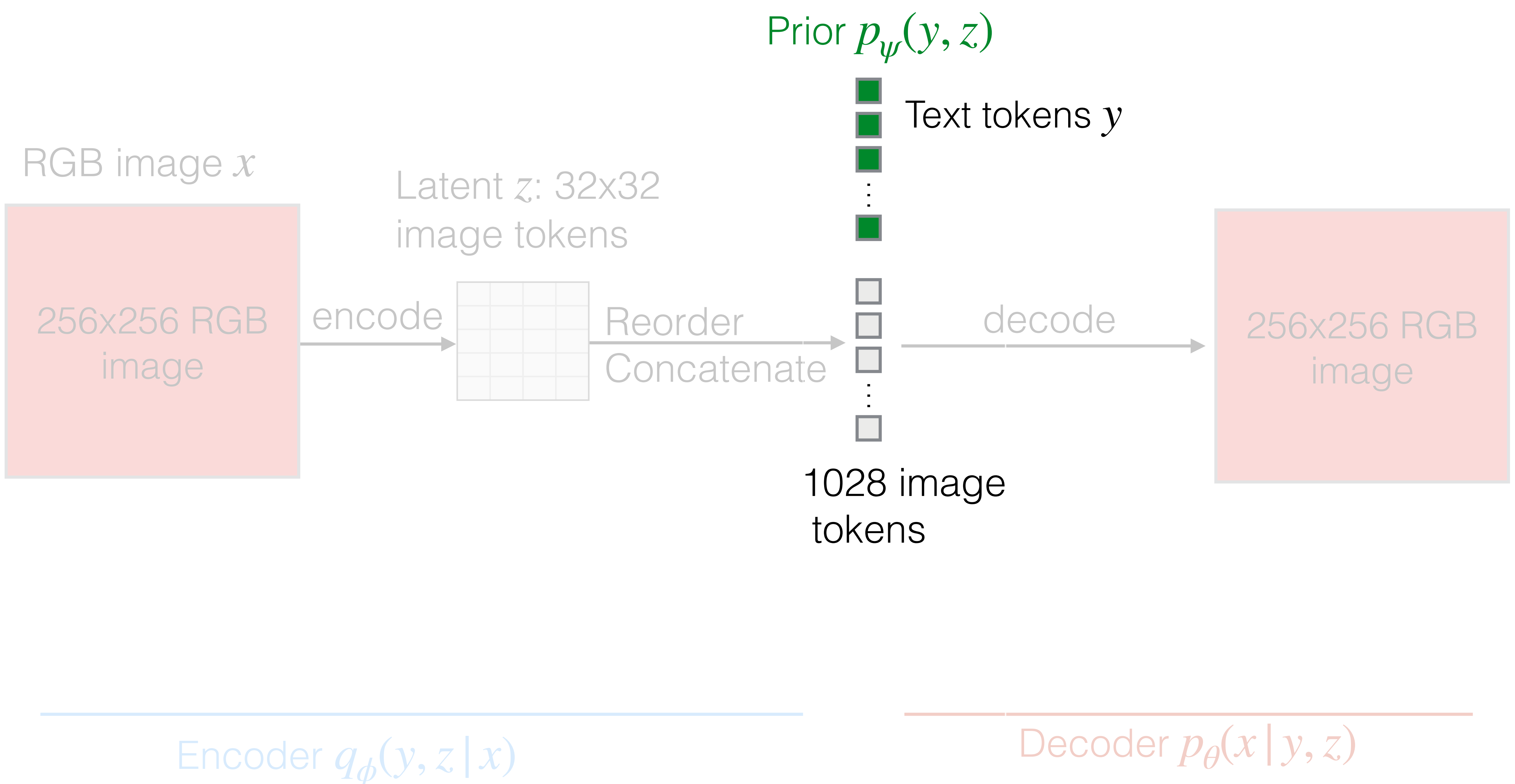


DALL-E: Stage 1



Ramesh et al. 2021. Zero-shot Text-to-Image Generation

DALL-E: Stage 2



Ramesh et al. 2021. Zero-shot Text-to-Image Generation

DALL-E: Overall Training Procedure

Maximize Evidence Lower Bound (ELB)— LVM lecture

$$\ln p_{\theta,\psi}(x, y) \geqslant \mathop{\mathbb{E}}_{z \sim q_\phi(z \,|\, x)} \Big(\ln p_\theta(x \,|\, y, z) - \beta \, D_{\mathrm{KL}}(q_\phi(y, z \,|\, x), p_\psi(y, z)) \Big),$$

Stage 1 updates p_θ, q_ϕ and fixes p_ψ

Stage 2 fixes p_θ, q_ϕ and updates p_ψ

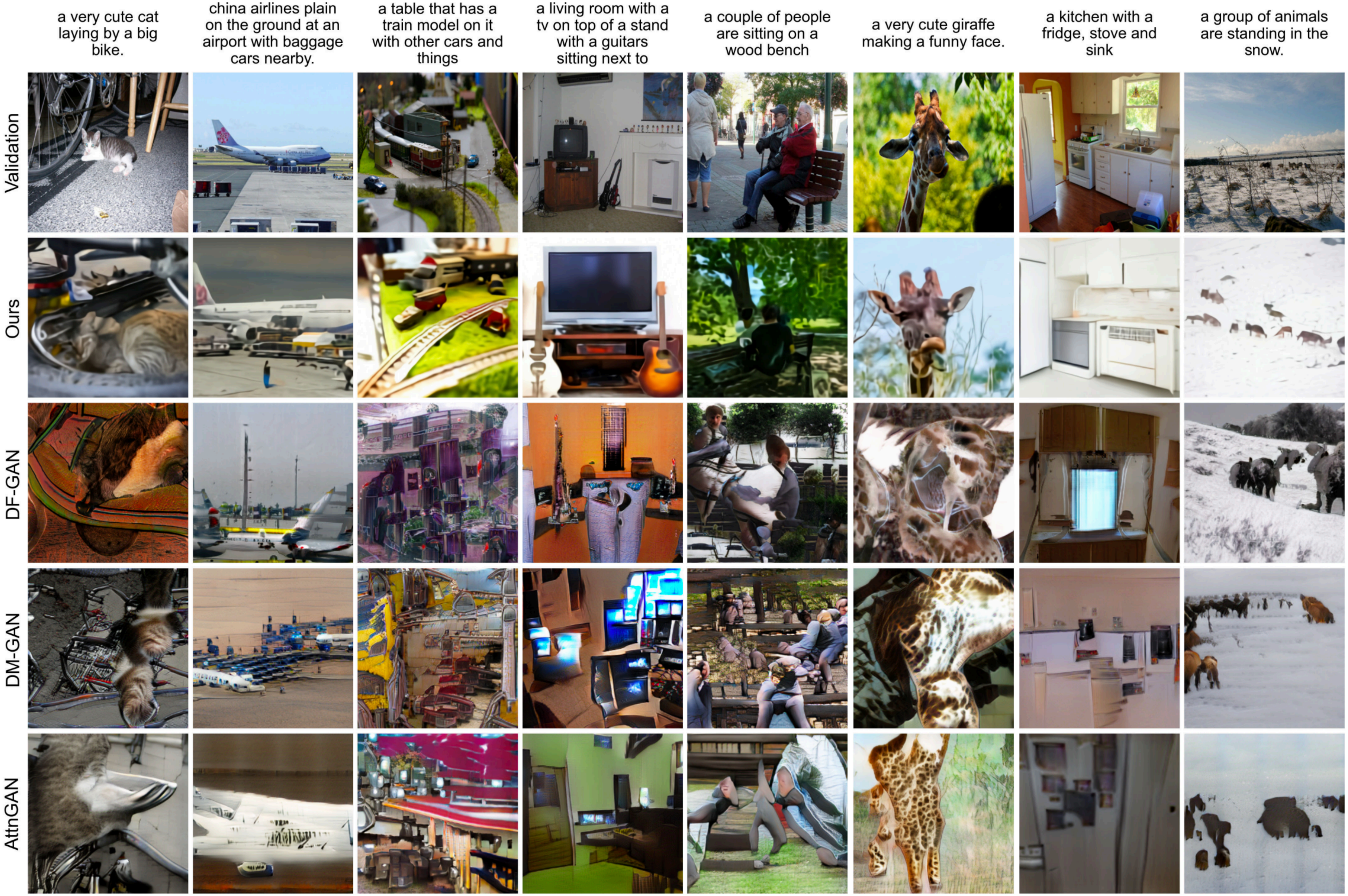
- x: the RGB image (256x256)
- z: the 32x32 (=1024) image tokens
- y: the text up to 256 tokens
- q_ϕ is the distribution over text tokens and the 32x32 image tokens generated by dVAE encoder given the RGB image x
- p_θ is the distribution over the RGB image generated by dVAE decoder given the image tokens and text tokens
- p_ψ is the prior distribution over the text and image tokens.

Ramesh et al. 2021. Zero-shot Text-to-Image Generation

DALL-E: Test Time

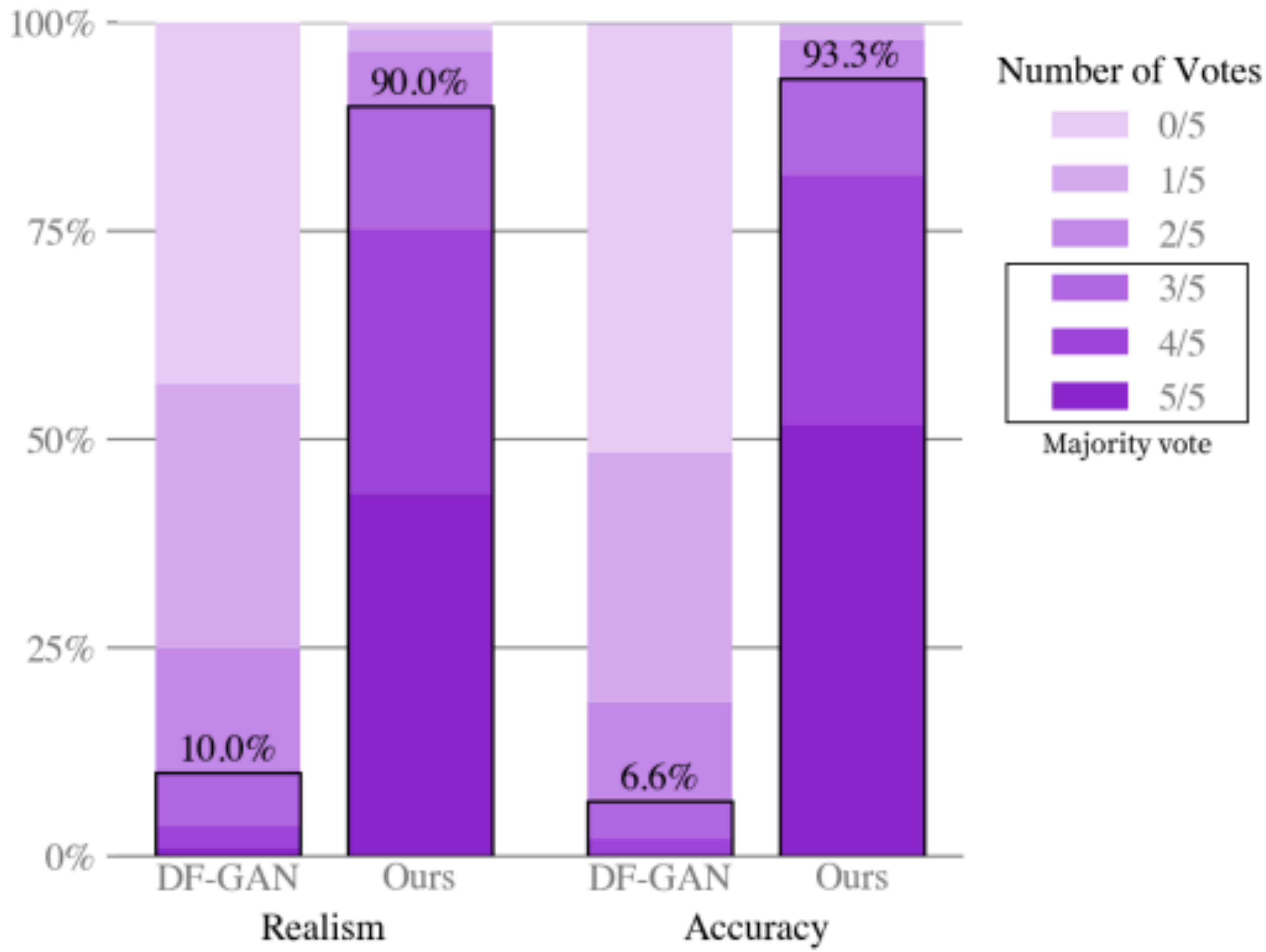
- Given a text prompt y, use the prior distribution (LM) to sample a sequence of 1028 image tokens
- Re-order 1028 image tokens to 32x32 shape
- Use dVAE's decoder to generate a RGB image from the image tokens.

Text-to-Image Generation



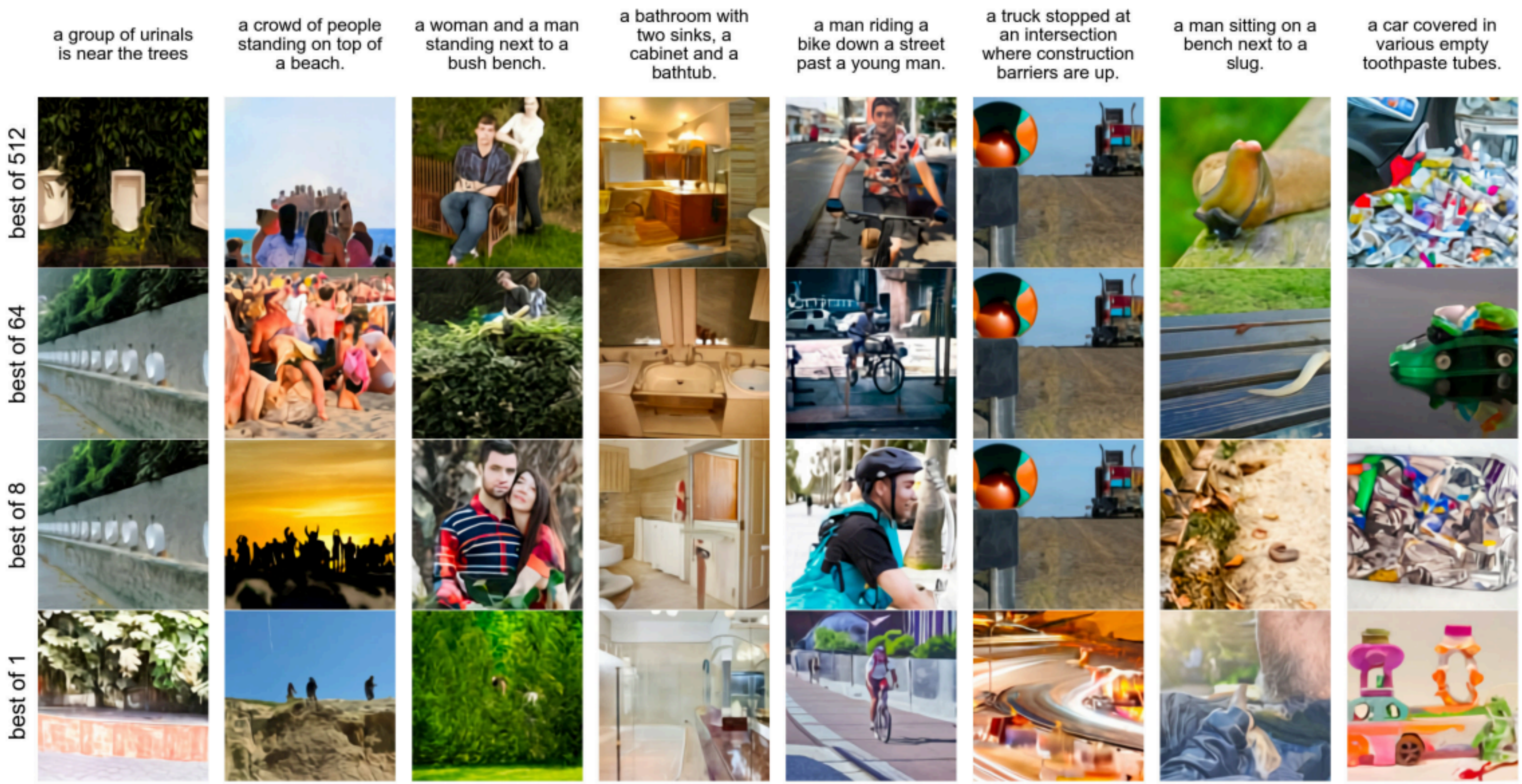
Human Eval on "Realism" and "Accuracy"

- DALL-E outperforms DF-GAN



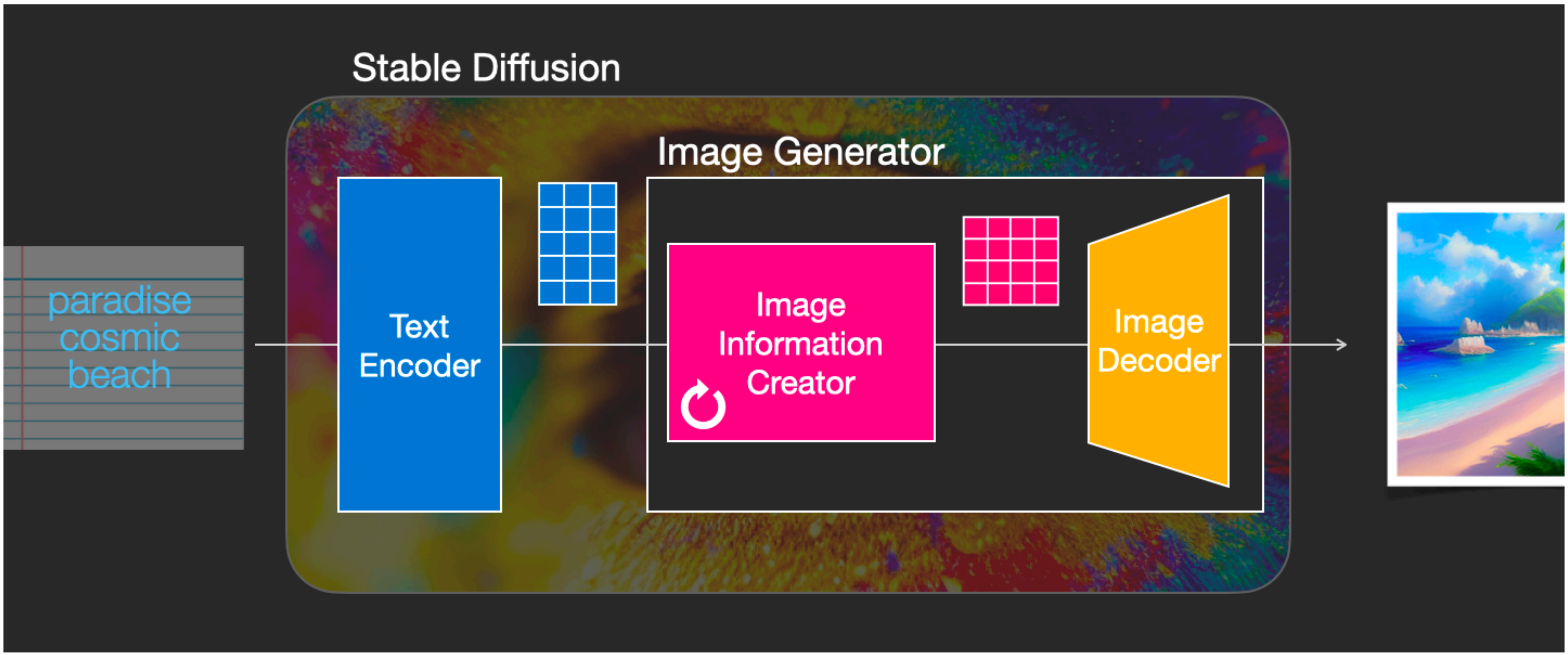
Sample, then Re-rank

- Sample K (e.g., K=1, 8, 64, 512) images from DALL-E, re-rank by CLIP, and pick the best output.



More Text-to-Image Generation Models

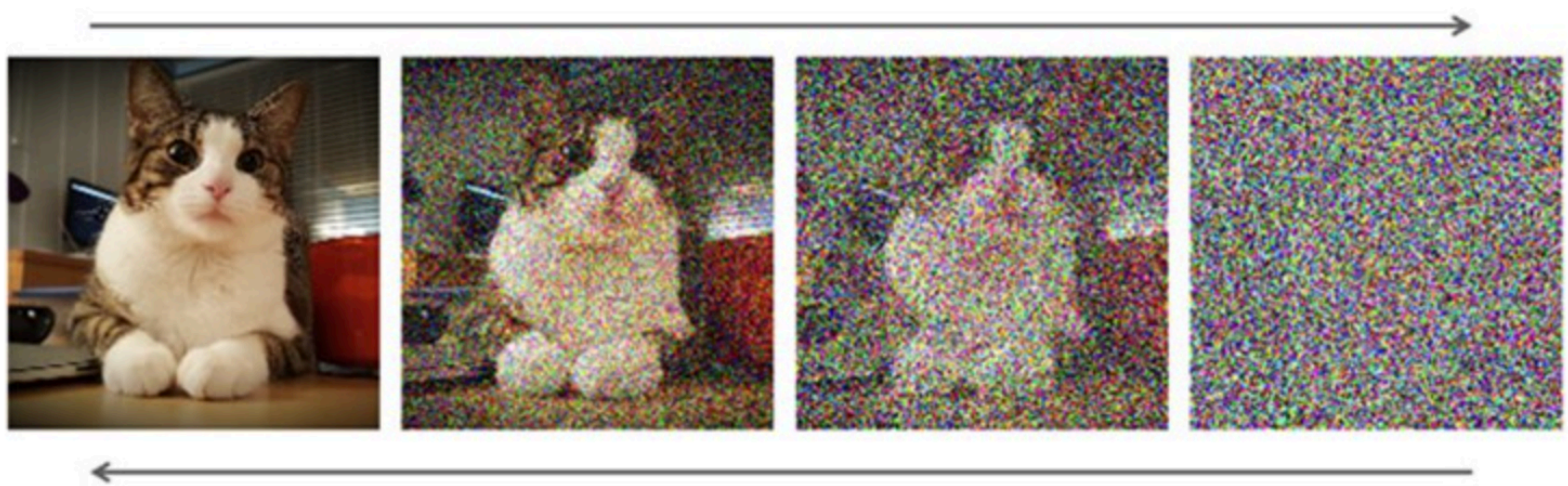
- Stable diffusion: Latent diffusion model

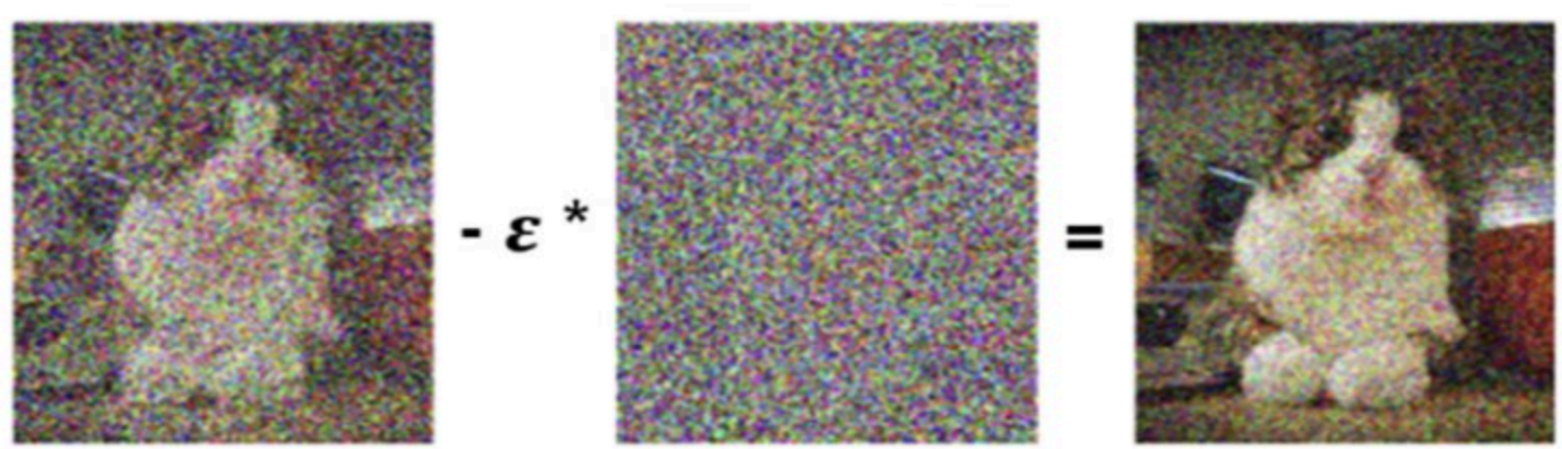


https://jalammar.github.io/illustrated-stable-diffusion/

More Text-to-Image Generation Models

- Stable diffusion: Add noise & remove noise





Questions?