CS639 Deep Learning for NLP

Multimodal Machine Learning: Vision-Language

Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from LP Morency

https://junjiehu.github.io/cs639-spring26/

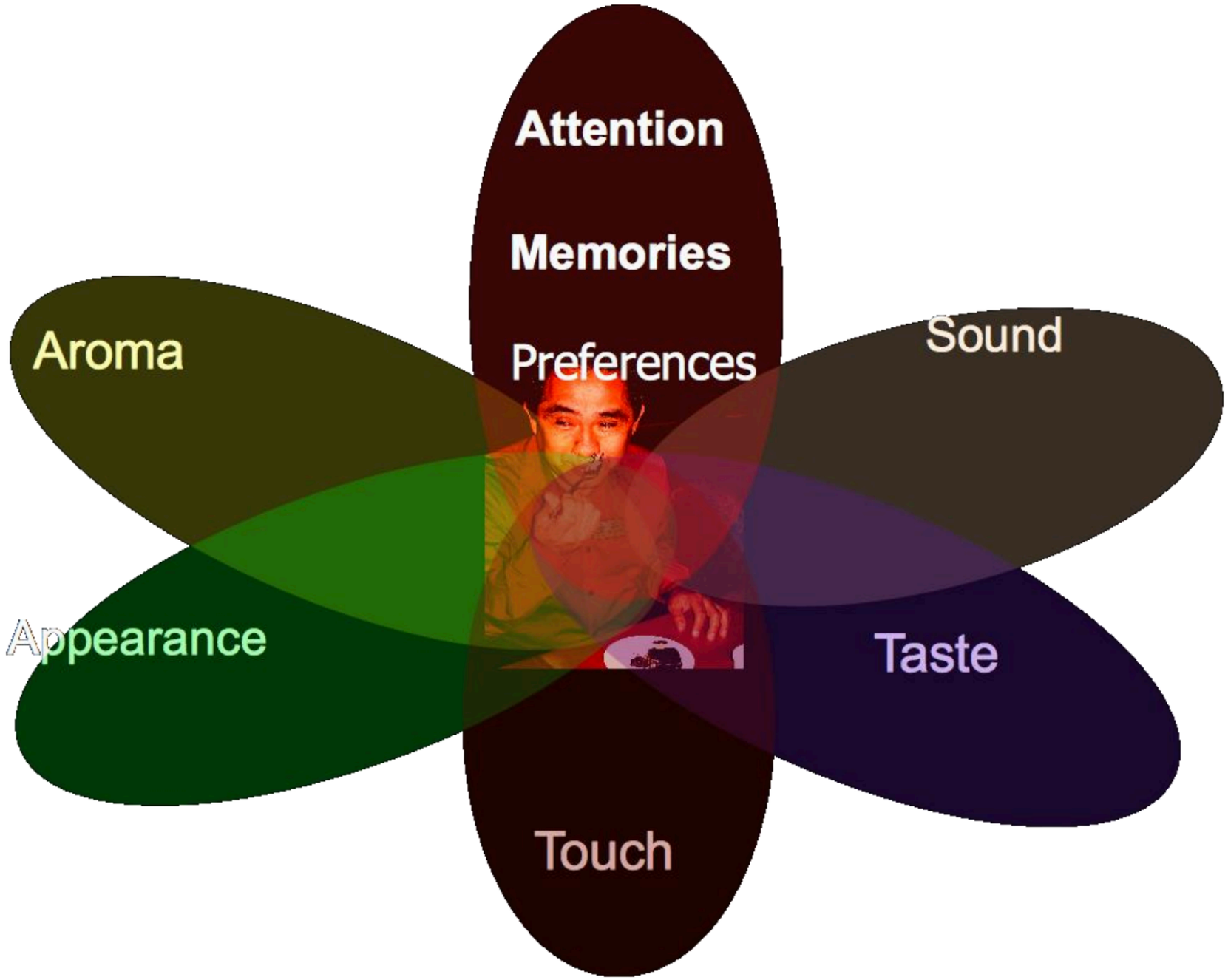
Goal for Today

- 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
  - LLAVA

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

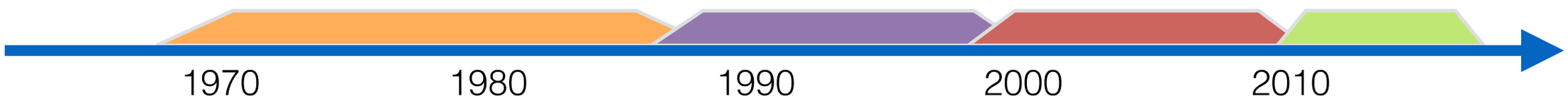
Facial expressions
- FACS action units
- Smile, frowning

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 …)

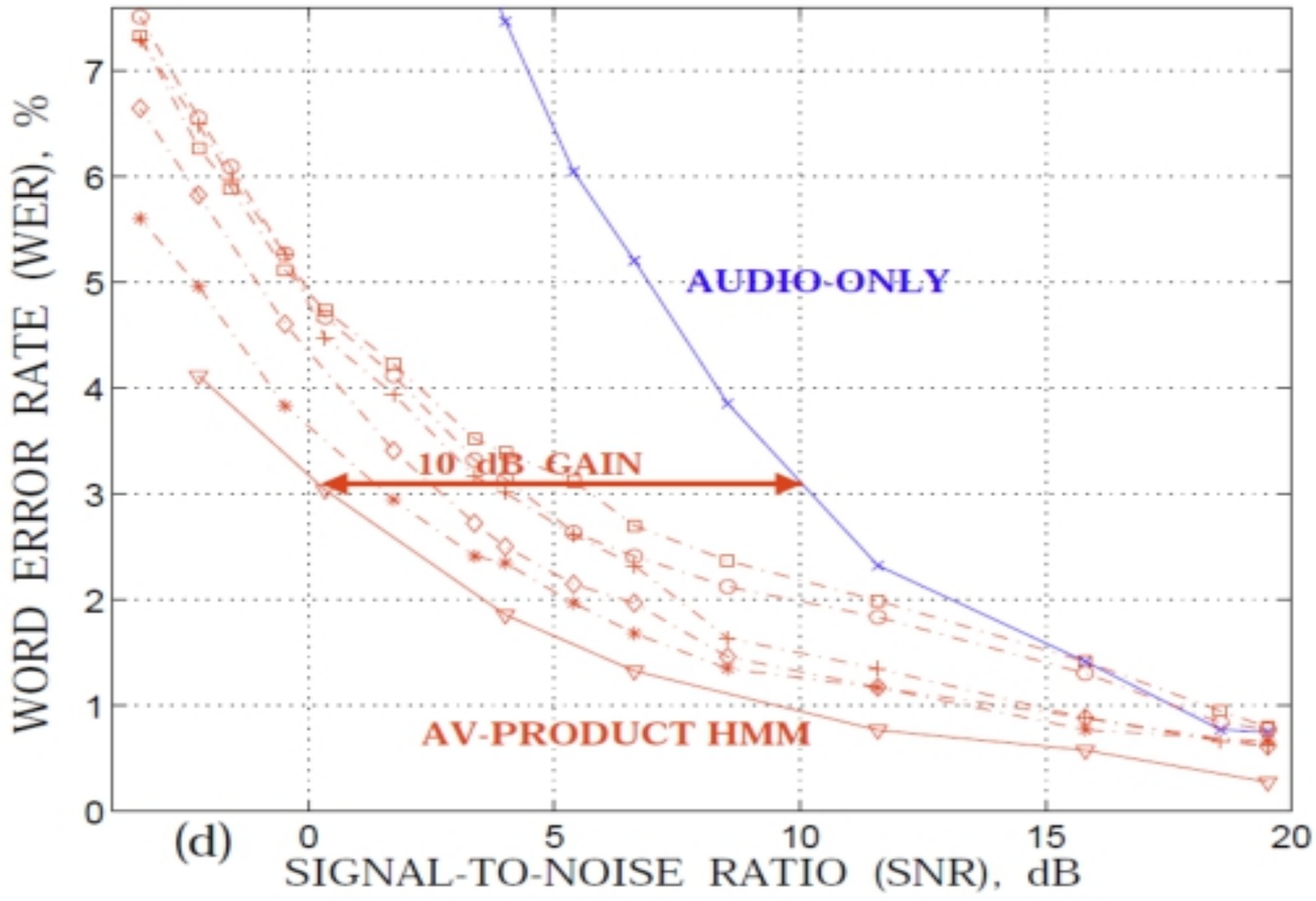


The McGurk Effect (1976)

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

The “Computational” Era (Late 1980s until 2000)

- Audio-Visual Speech Recognition (AVSR)



1970 1980 1990 2000 2010

Core Technical Challenges

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

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

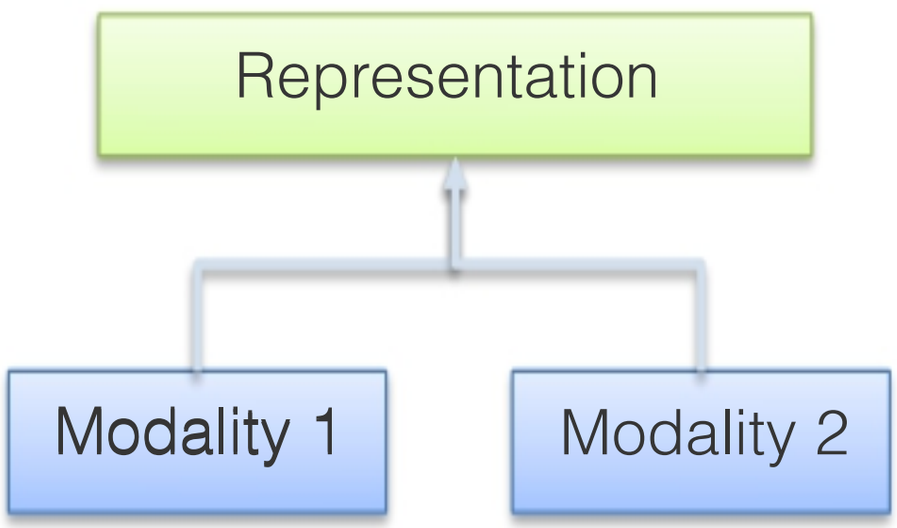
These challenges are non-exclusive.

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

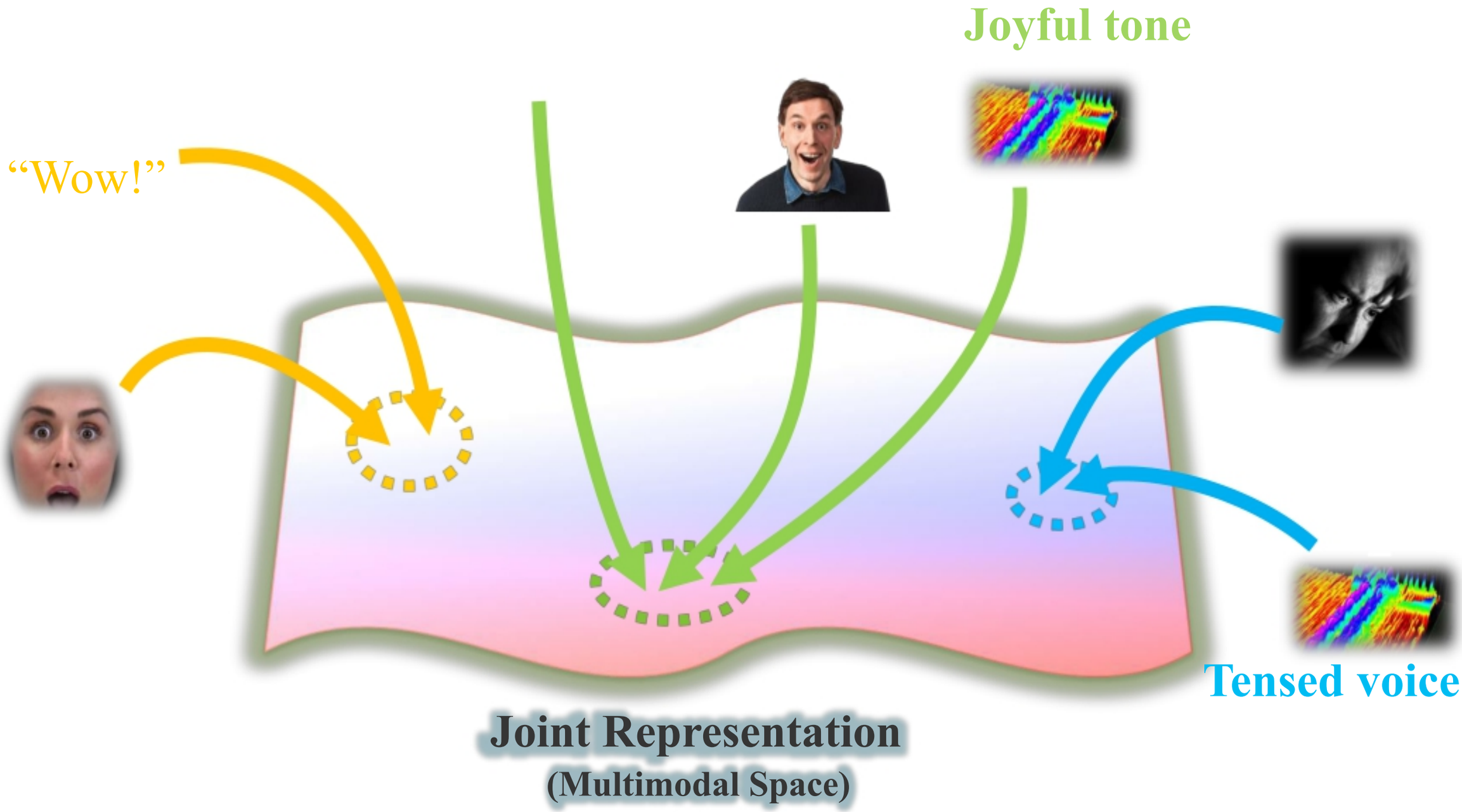
Core Challenge 1: Representation

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

A Joint representations:

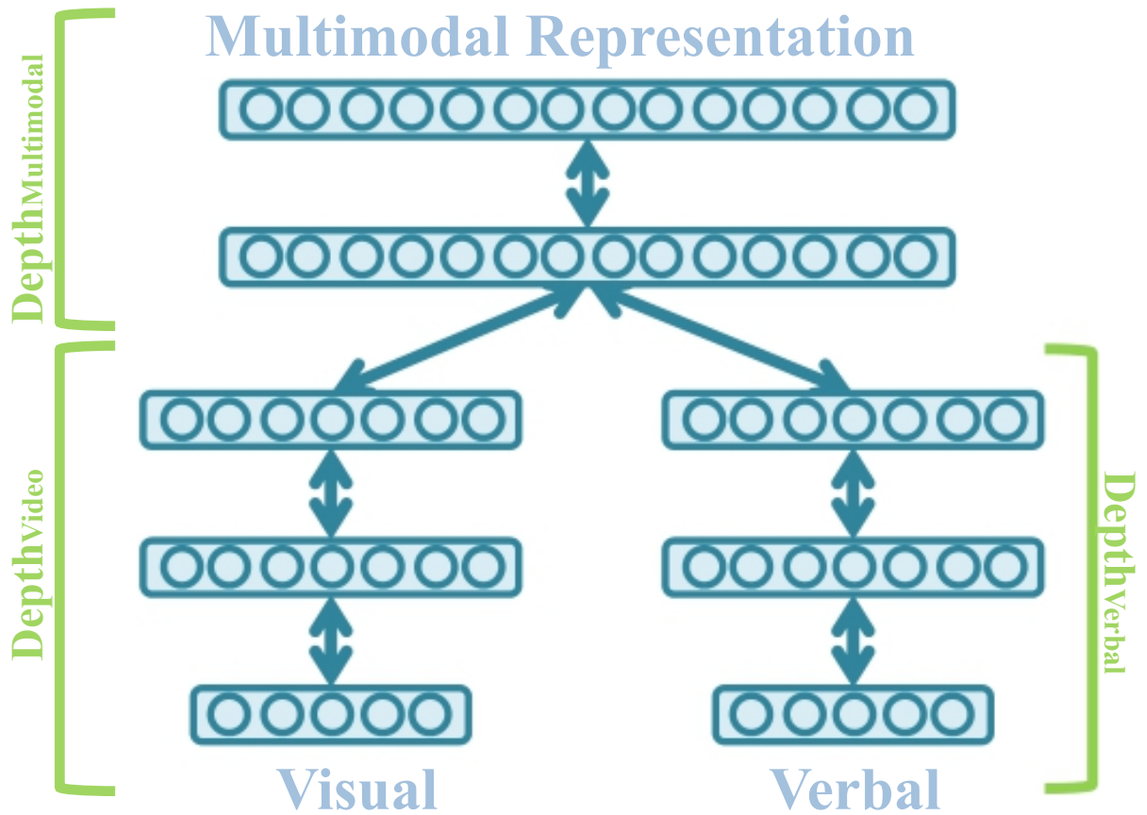


Joint Multimodal Representations



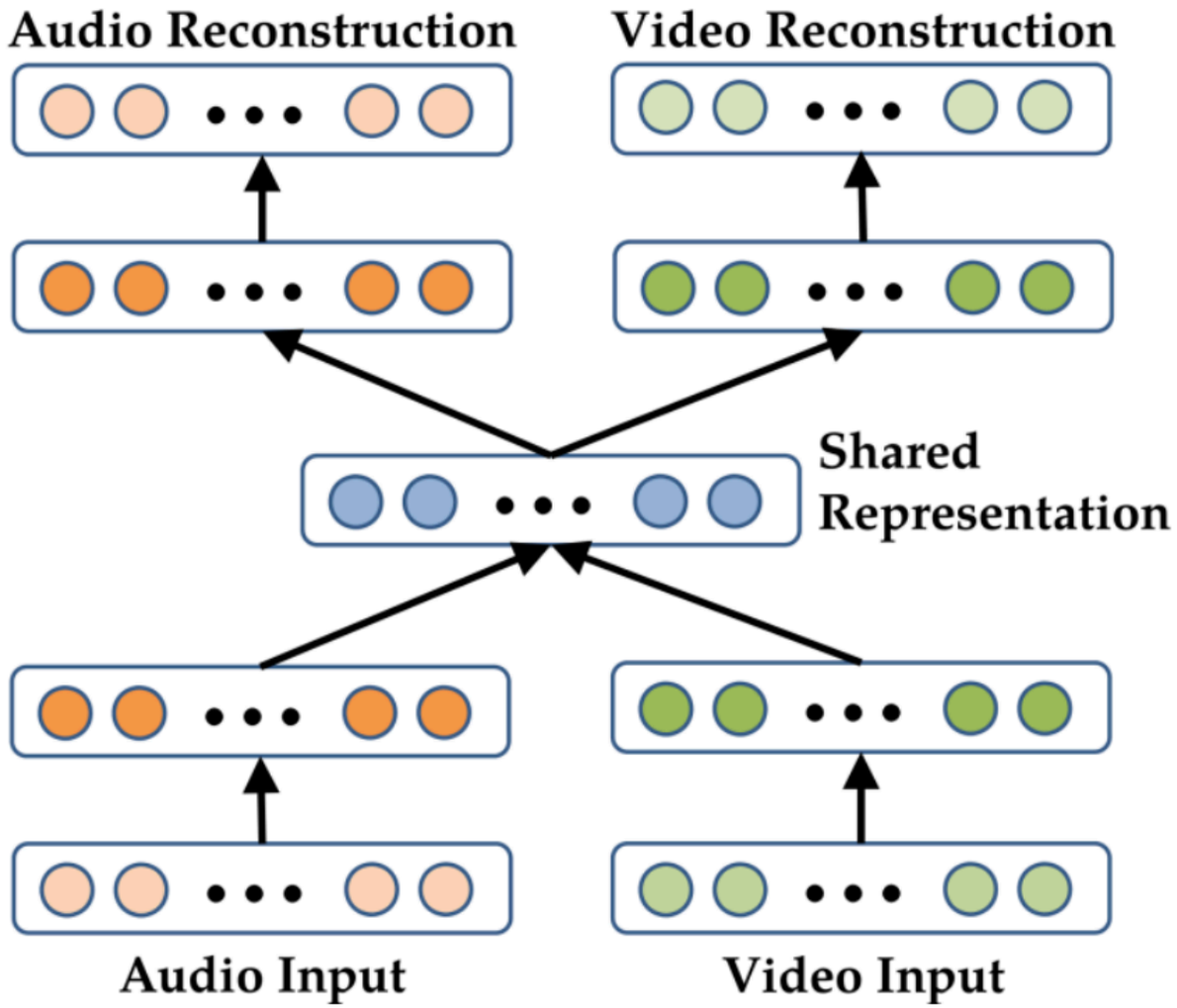
Joint Multimodal Representations

- 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

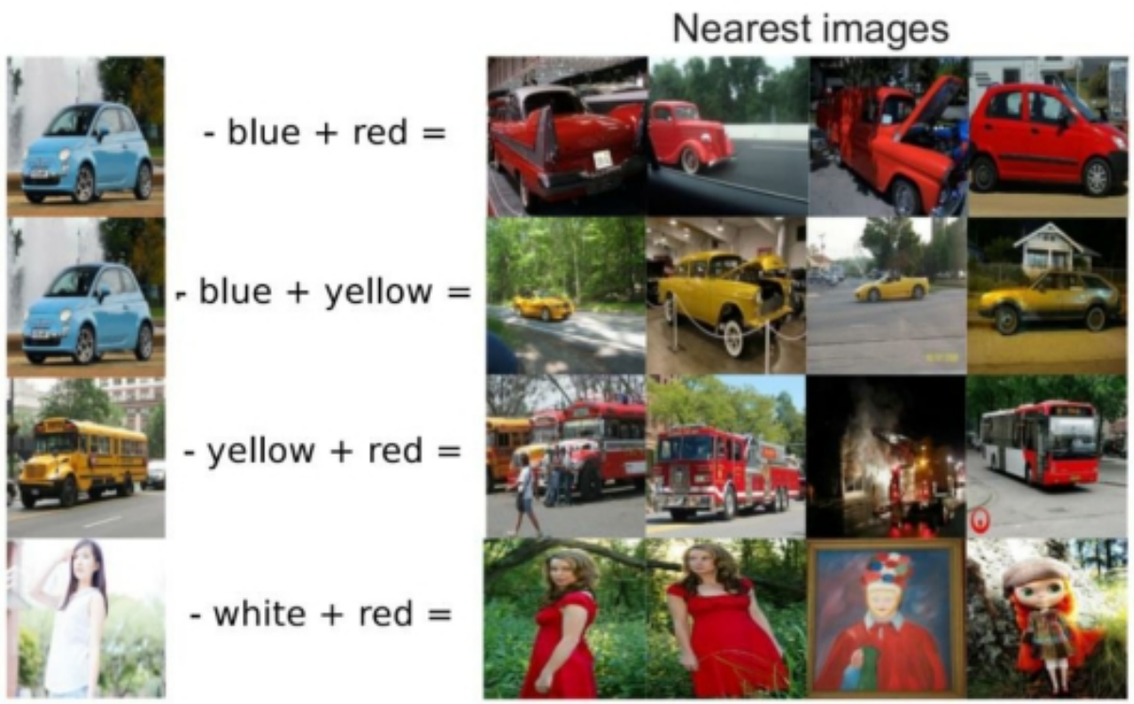
- Bimodal auto-encoder
  - Used for audio-visual speech recognition
- Individual modalities can be pretrained
  - RBMs
  - Denoising Autoencoders
- Train the model to reconstruct the other modality
  - Use both
  - Remove audio
  - Remove video

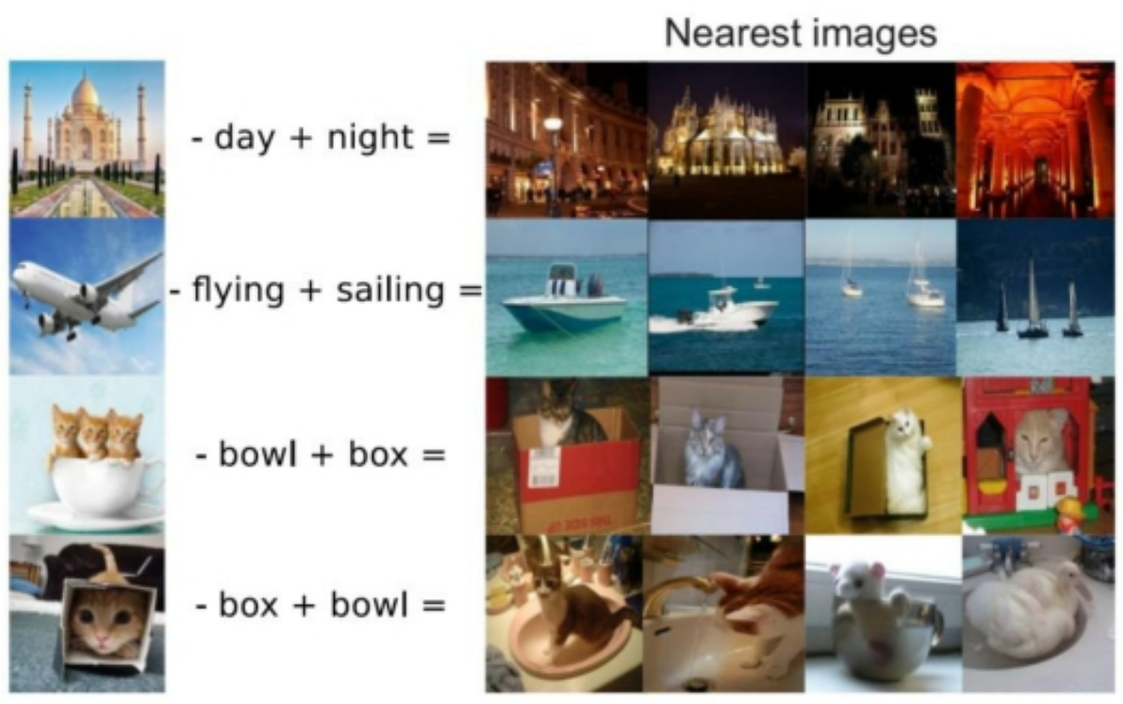


Ngiam et al., Multimodal Deep Learning, 2011.

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



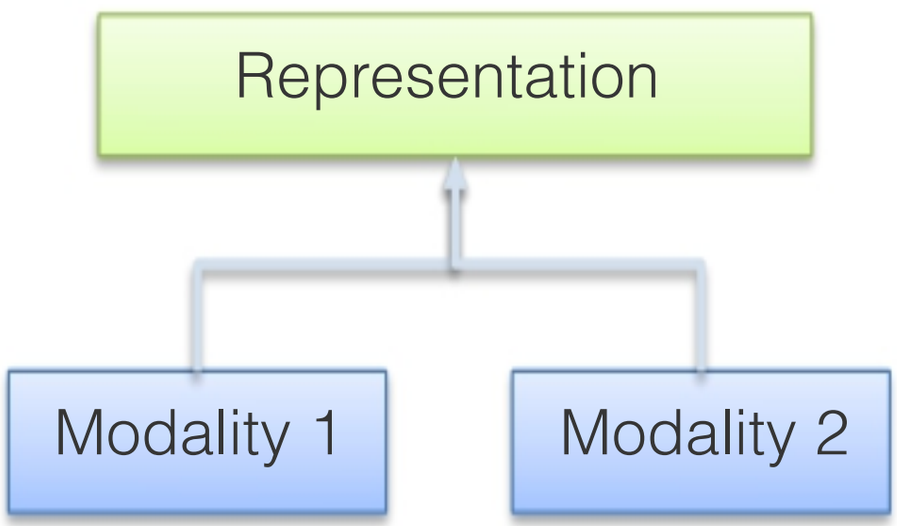


Kiros et al. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language models.

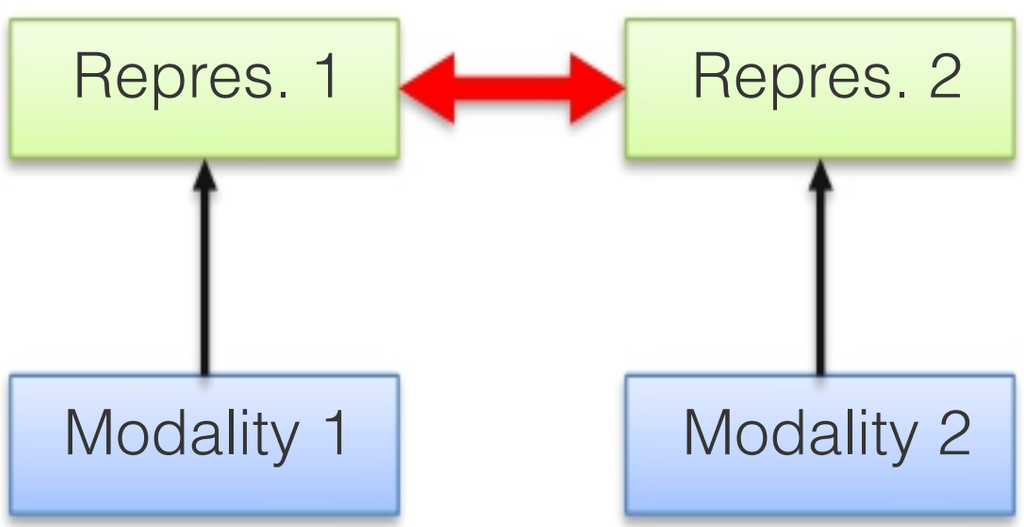
Core Challenge 1: Representation

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

A Joint representations:



B Coordinated representations:

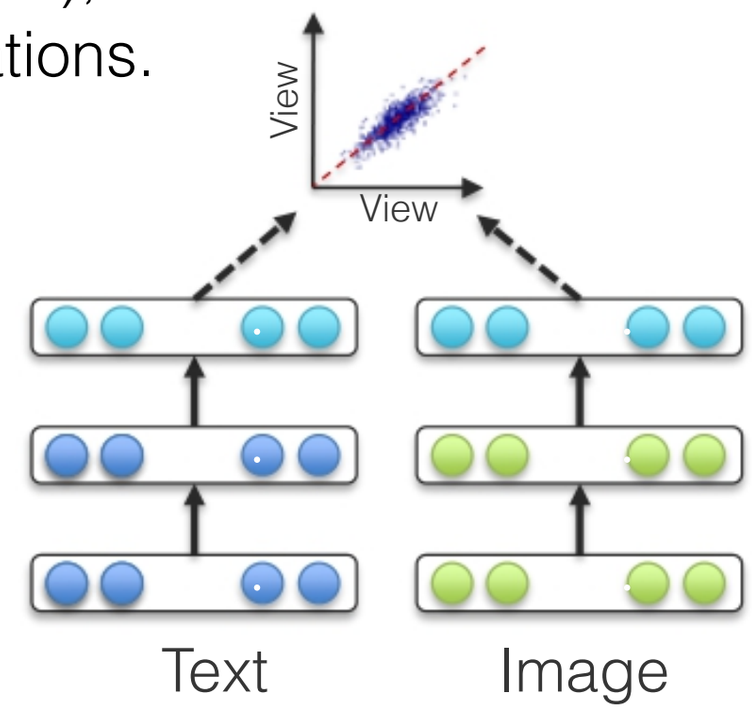


Coordinated Representation: Deep CCA

- Learn linear projections that are maximally correlated:

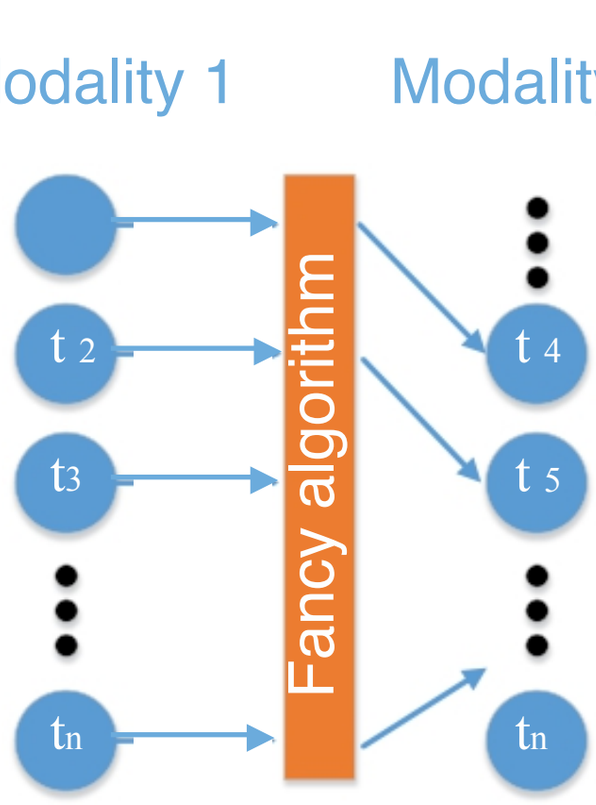
$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1, \theta_2)}{\operatorname{argmax}} \operatorname{corr}(f_1(X_1; \theta_1), f_2(X_2; \theta_2)).$$

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

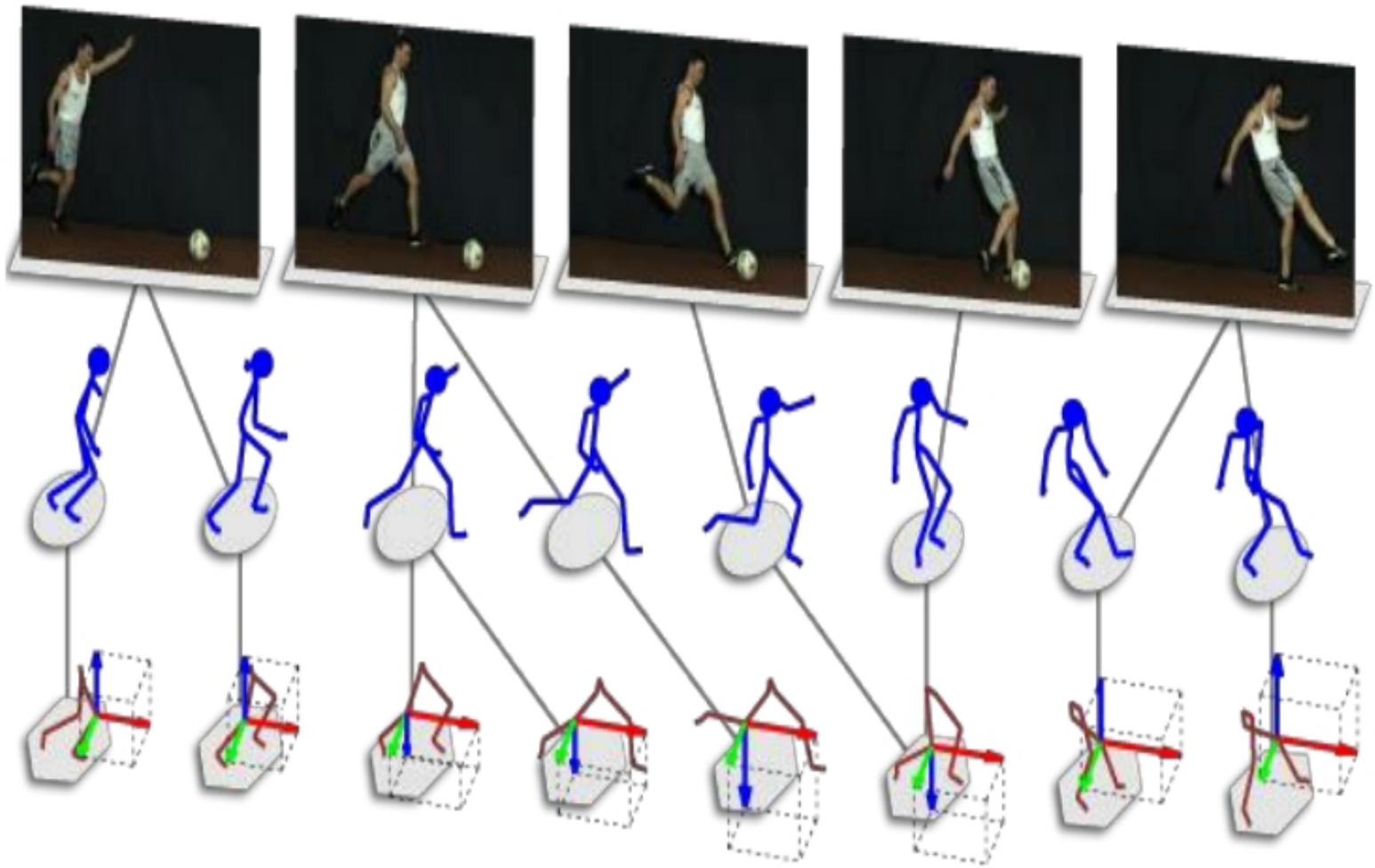
The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

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

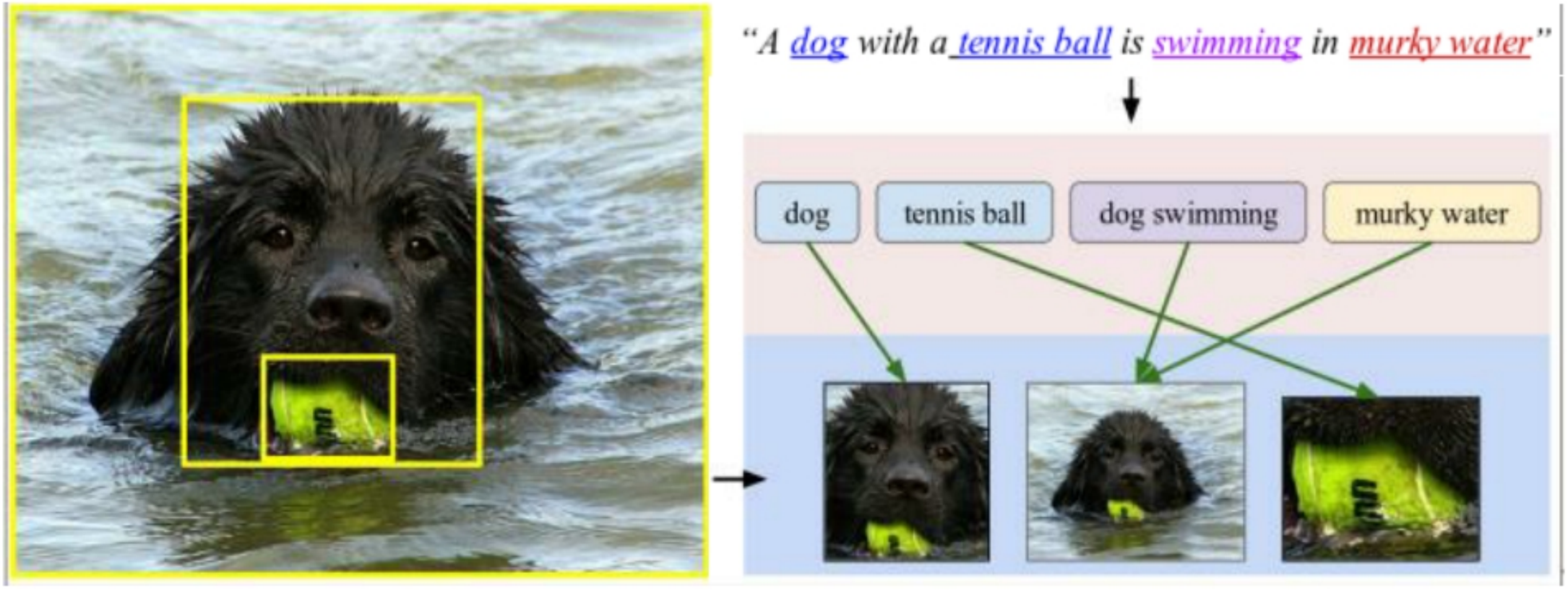
Example: Temporal Sequence Alignment

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



Implicit Alignment

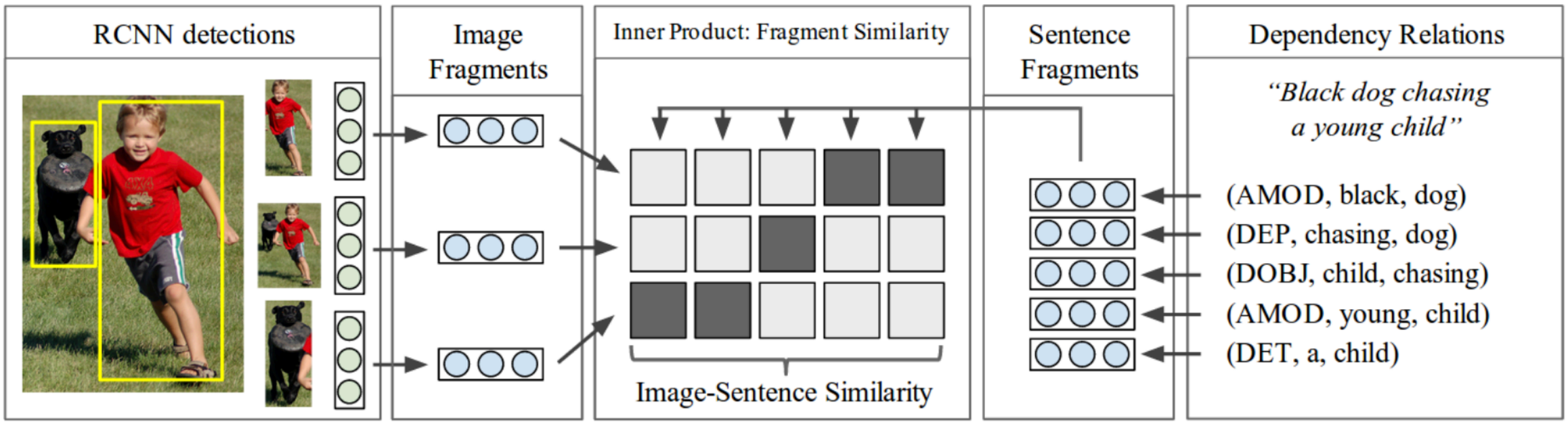
- Vision-language alignment, a.k.a. visual grounding.



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.



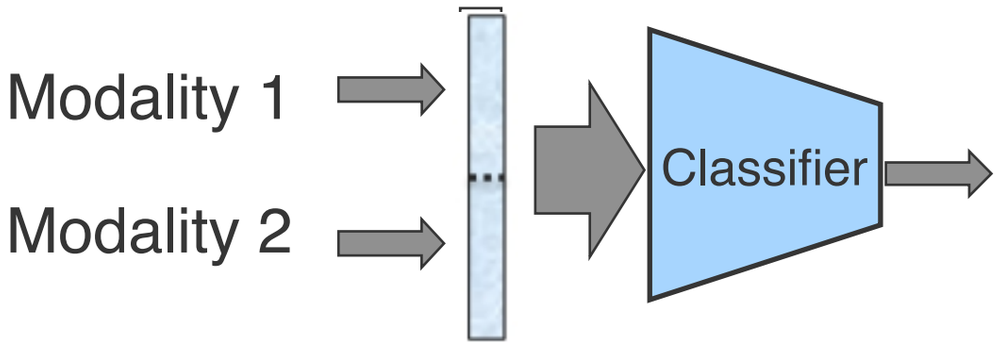
Karpathy et al. 2014 Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Core Challenge 3: Fusion

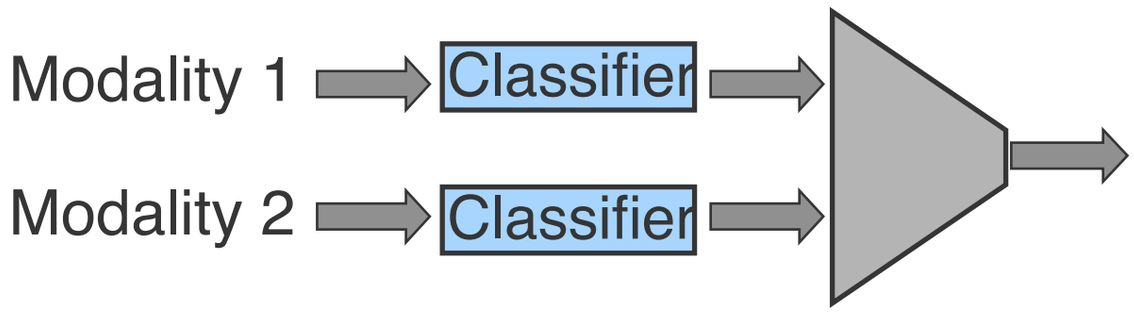
- **Definition**: To join information from two or more modalities to perform a prediction task.

A Model-Agnostic Approaches

1) Early Fusion



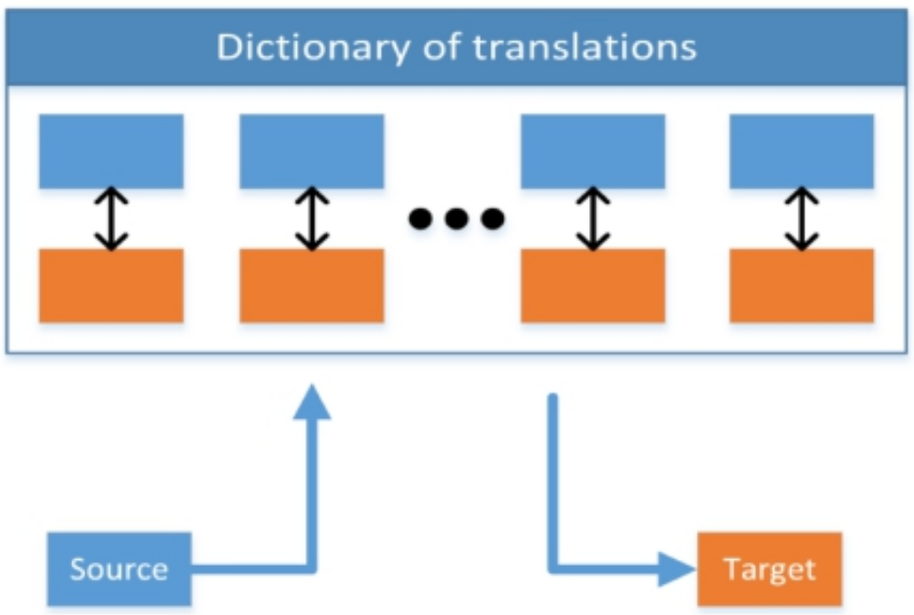
2) Late Fusion



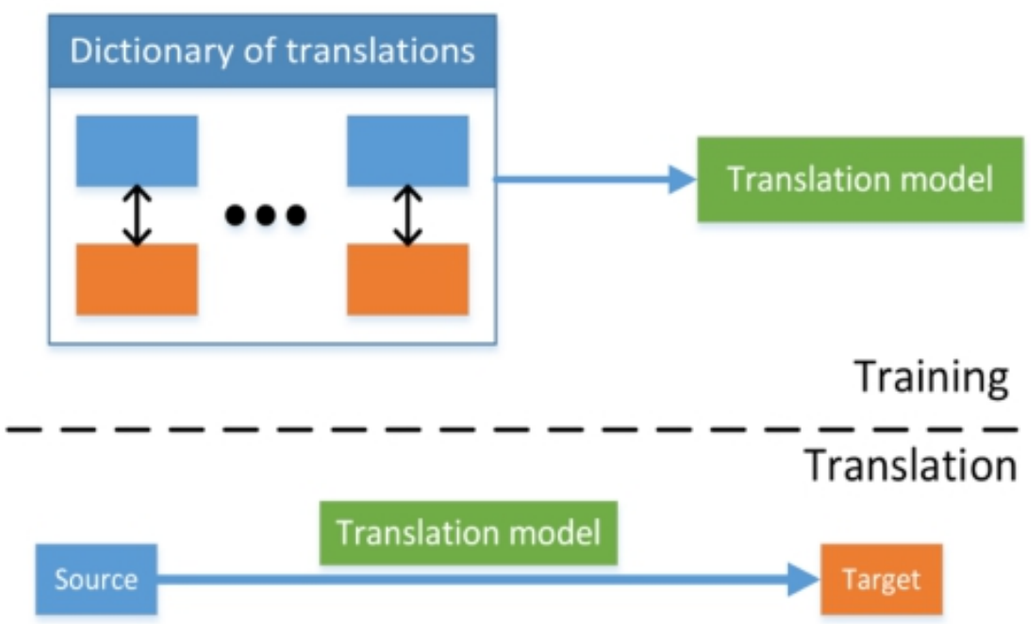
Core Challenge 4: Translation

- **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

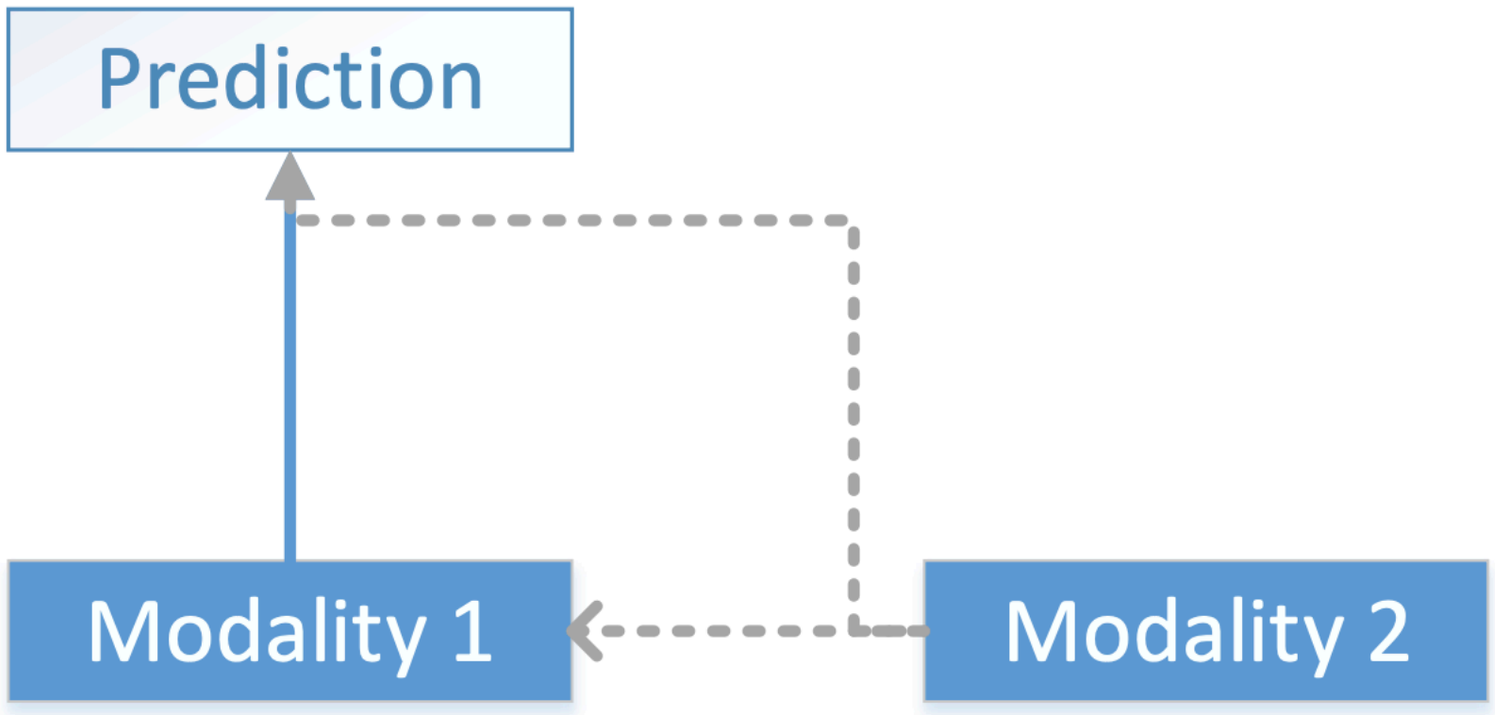
Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

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

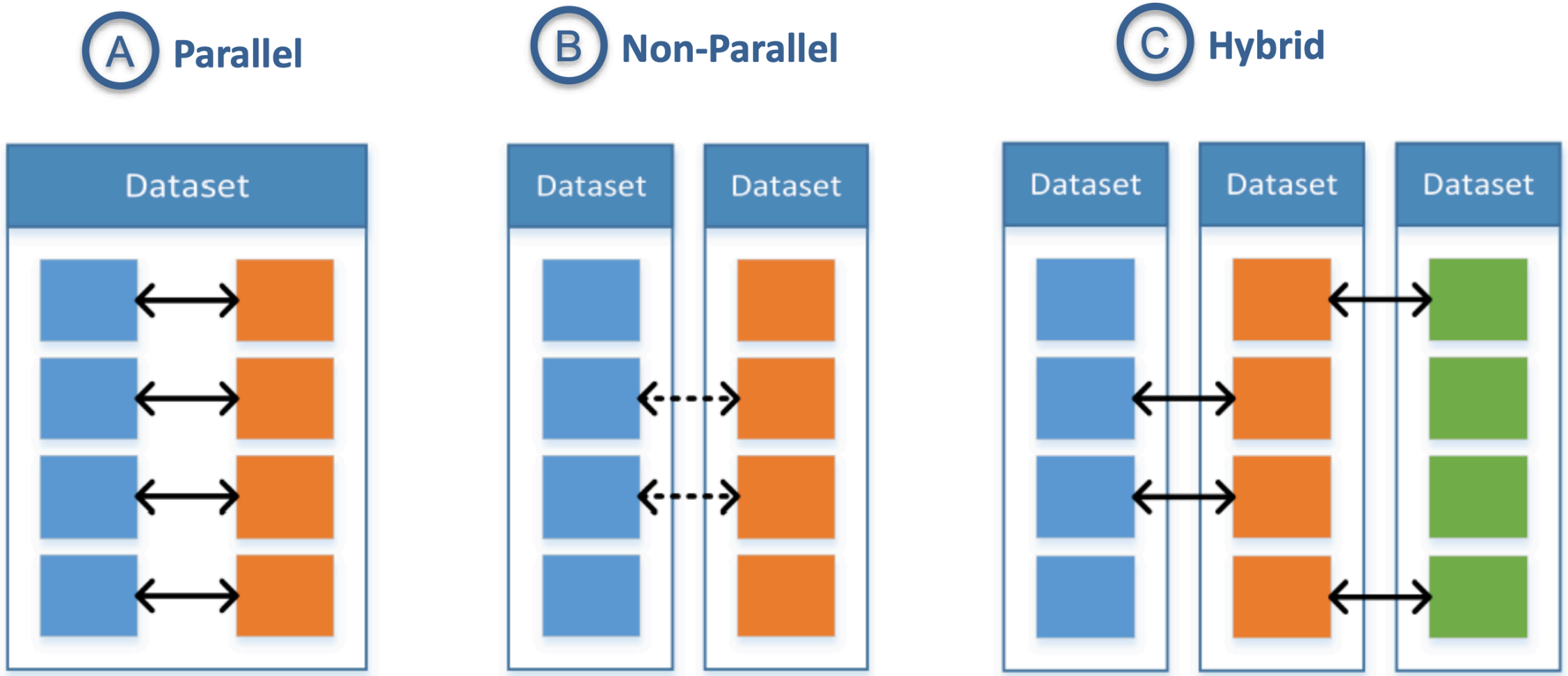
Core Challenge 5: Co-Learning

- **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

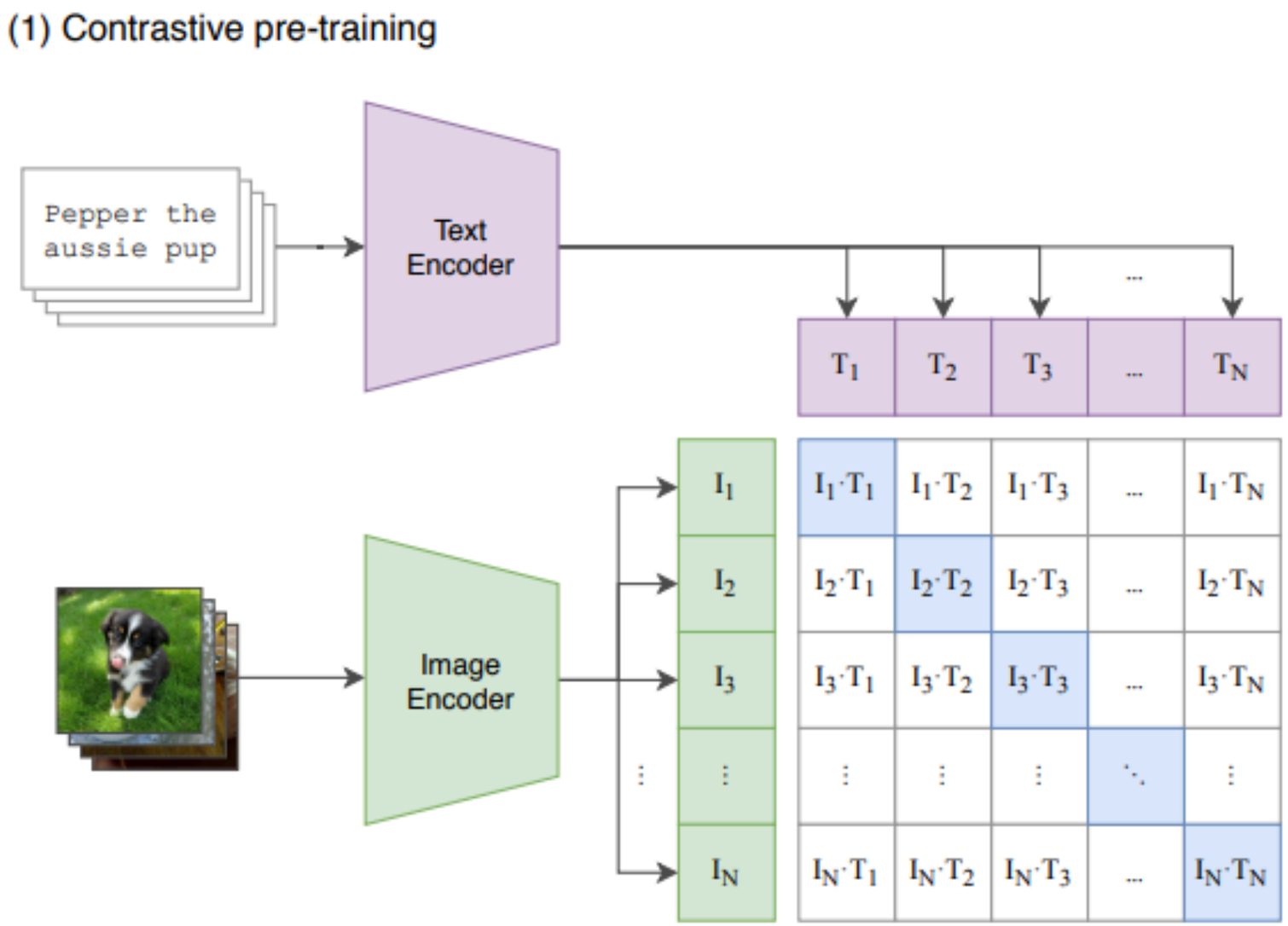
Multimodal Applications

	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

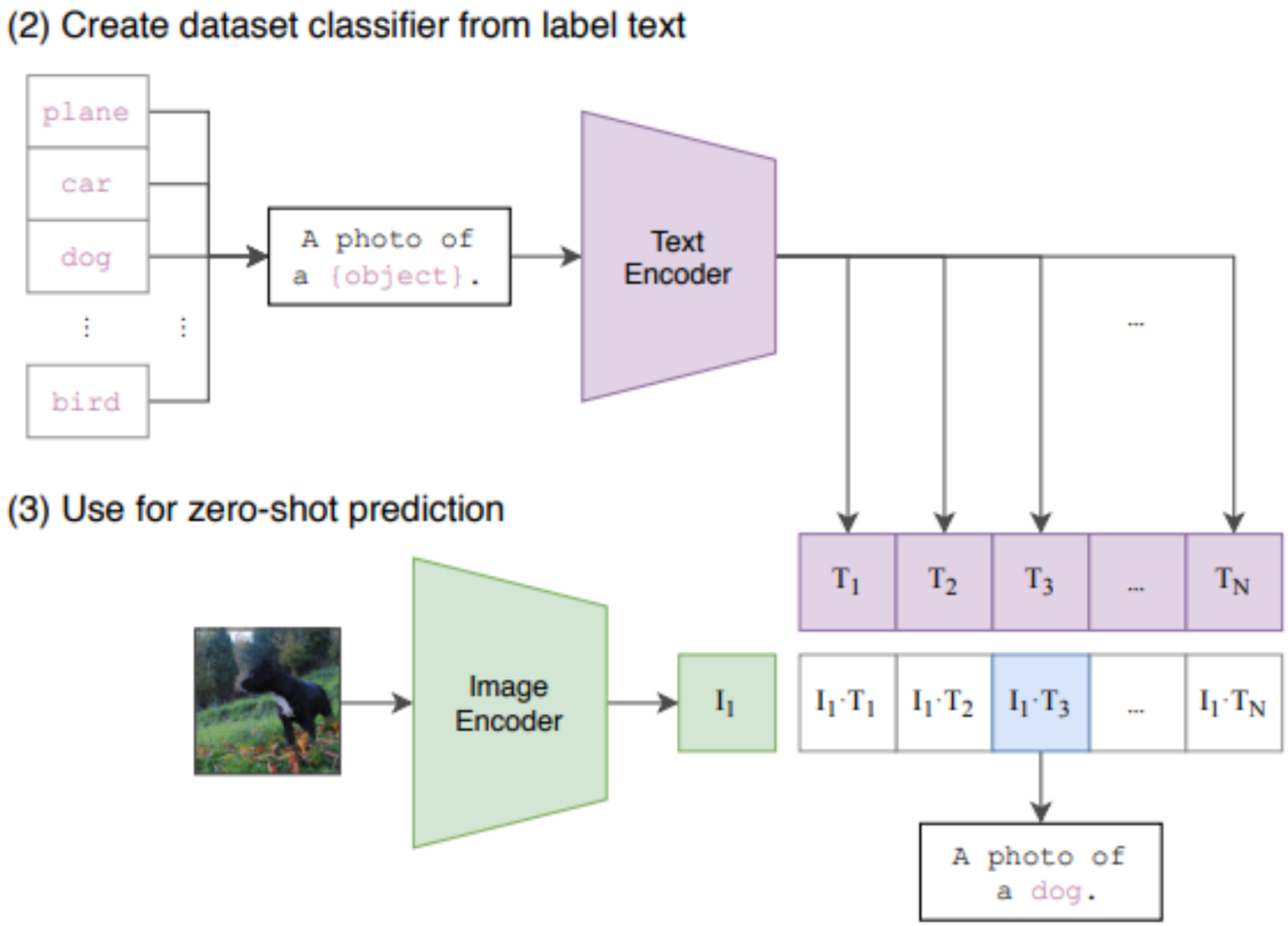
- Pre-train V+L models using image captioning data (i.e., image-text pairs) by contrastive loss



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

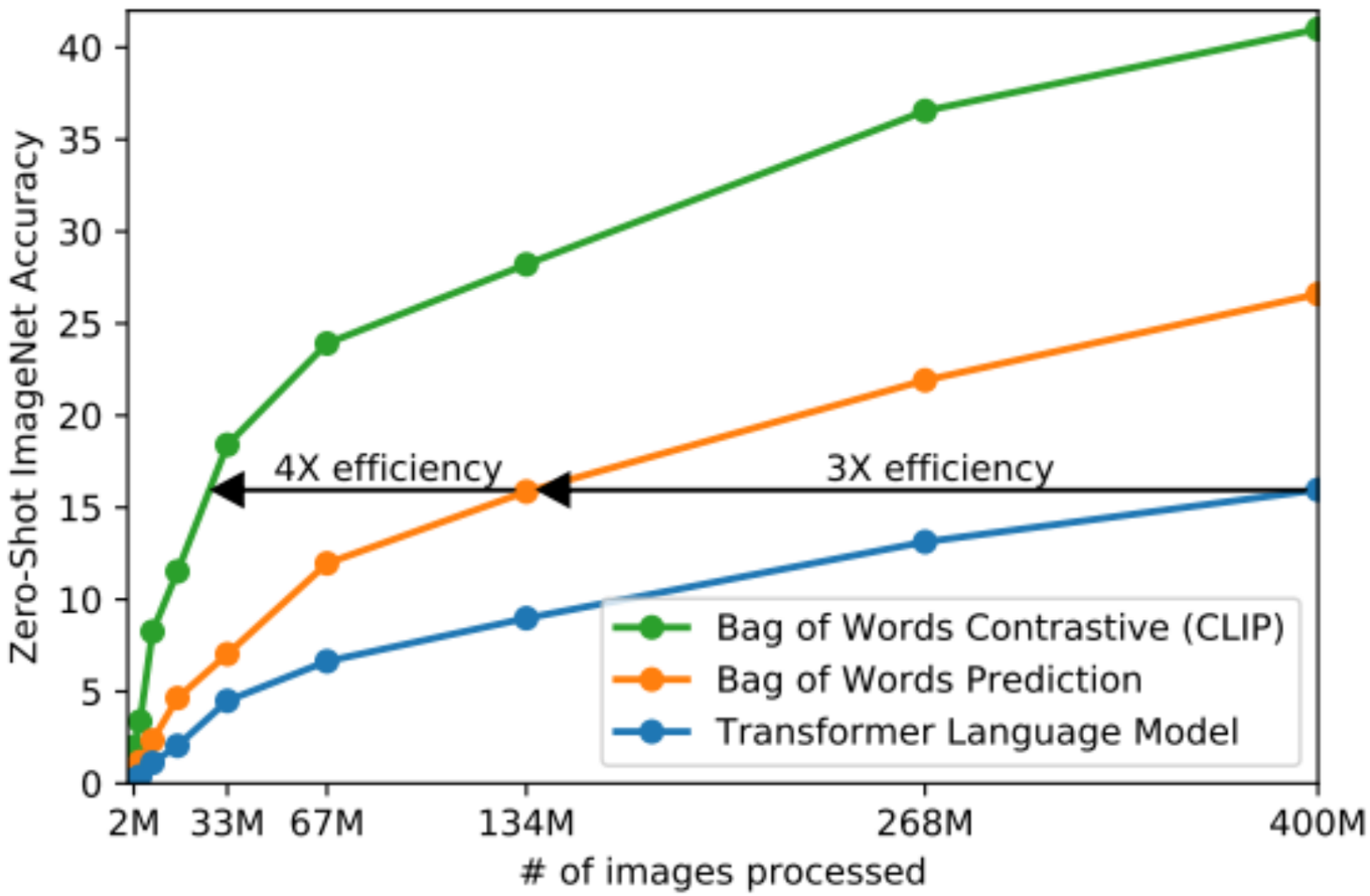


CLIP: pseudocode

```
# 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
```

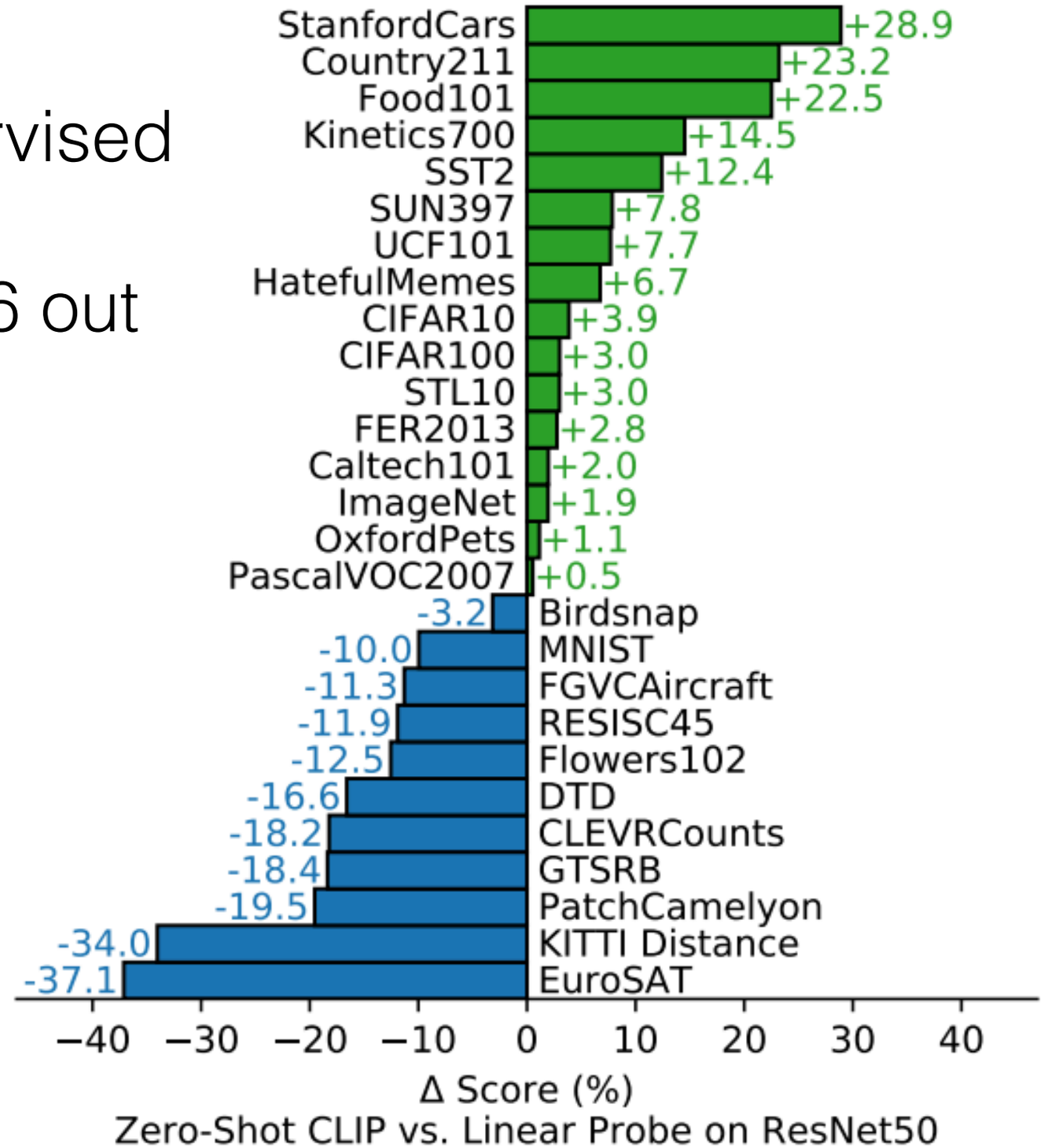
Efficiency of BoW Representations

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



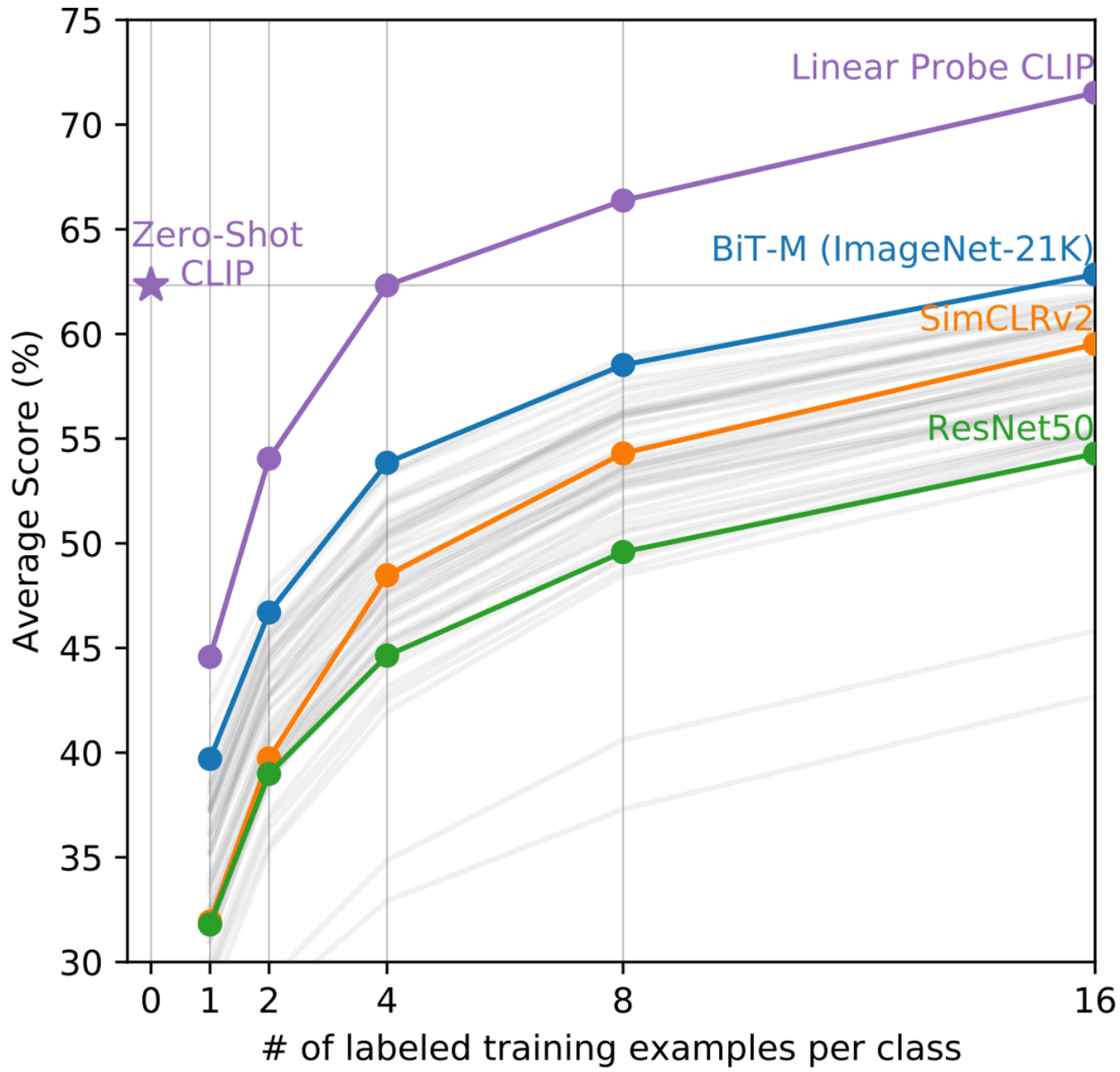
Zero-shot Image Classification

- 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

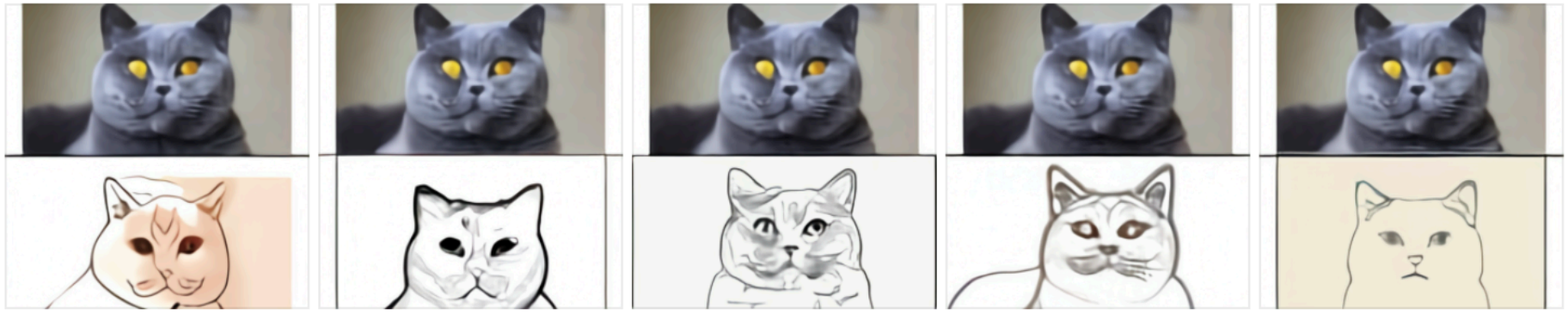
AI-GENERATED IMAGES

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



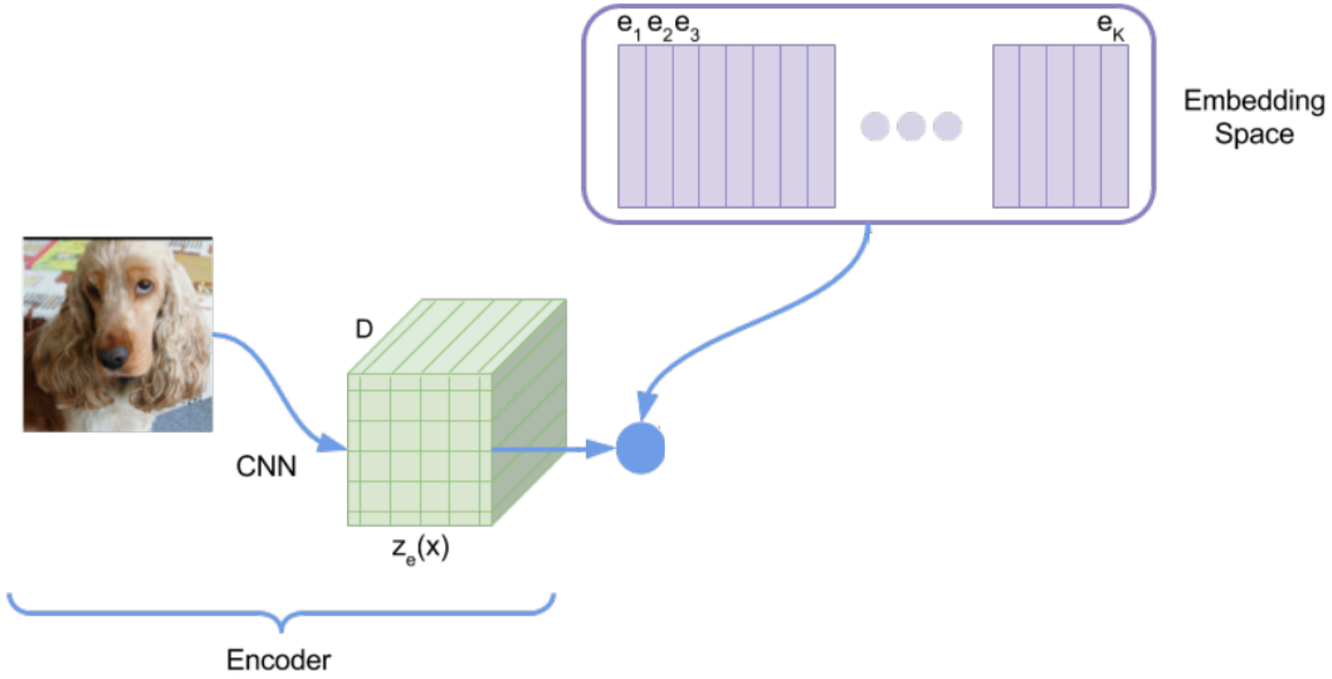
DALL-E

- **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)

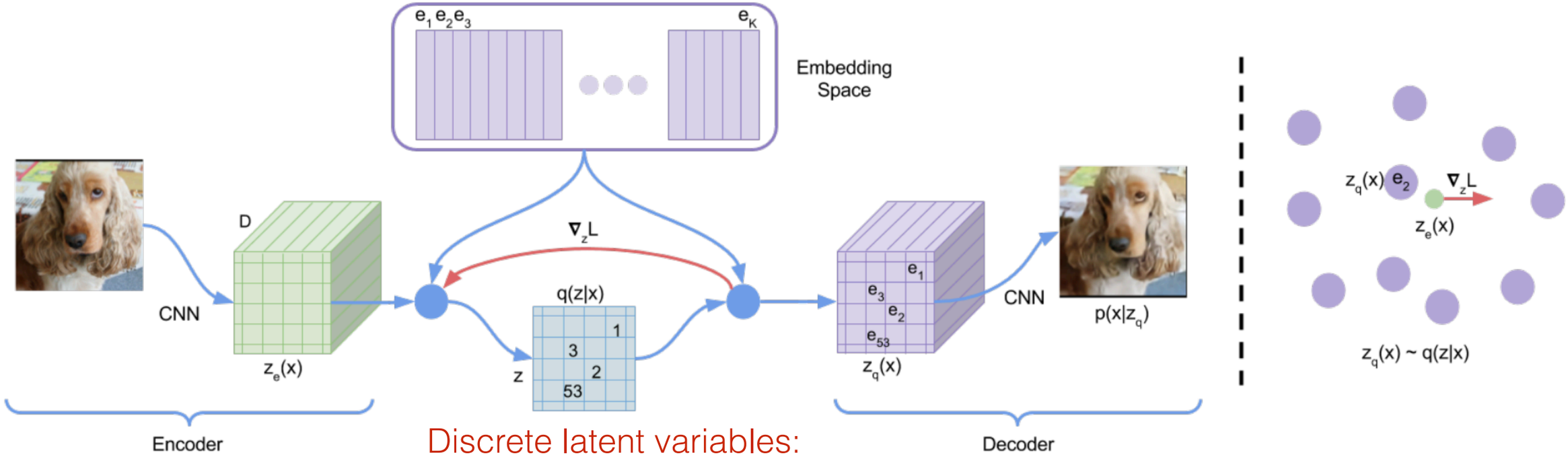


Ramesh et al. 2021. Zero-shot Text-to-Image Generation
Oord et al. 2018. Neural Discrete Representation Learning

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)

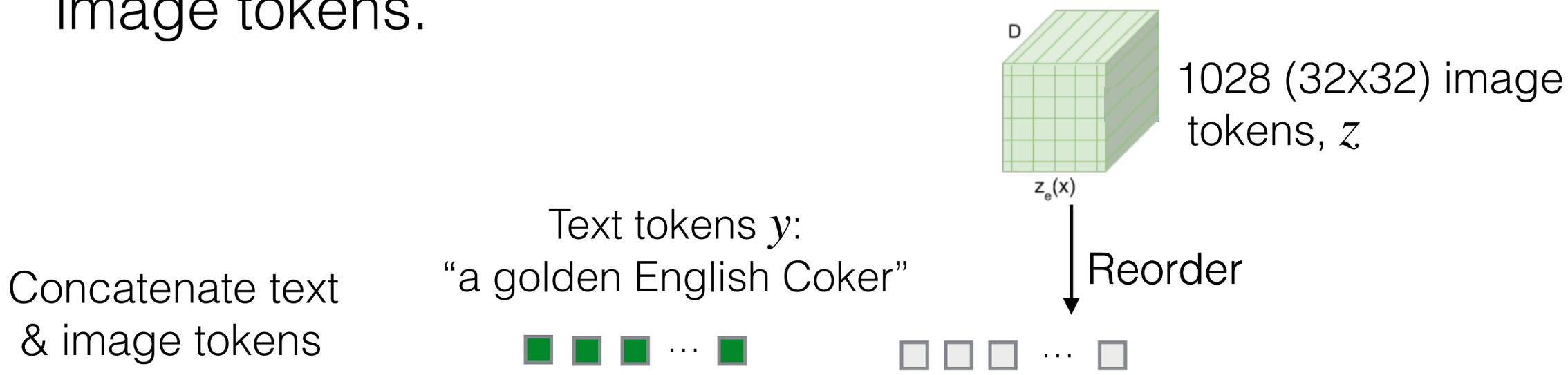


Discrete latent variables:
Gumbel-softmax — c.f. LVM lecture

VAE training:
Maximize Evidence Lower Bound

DALL-E: Language Model Training

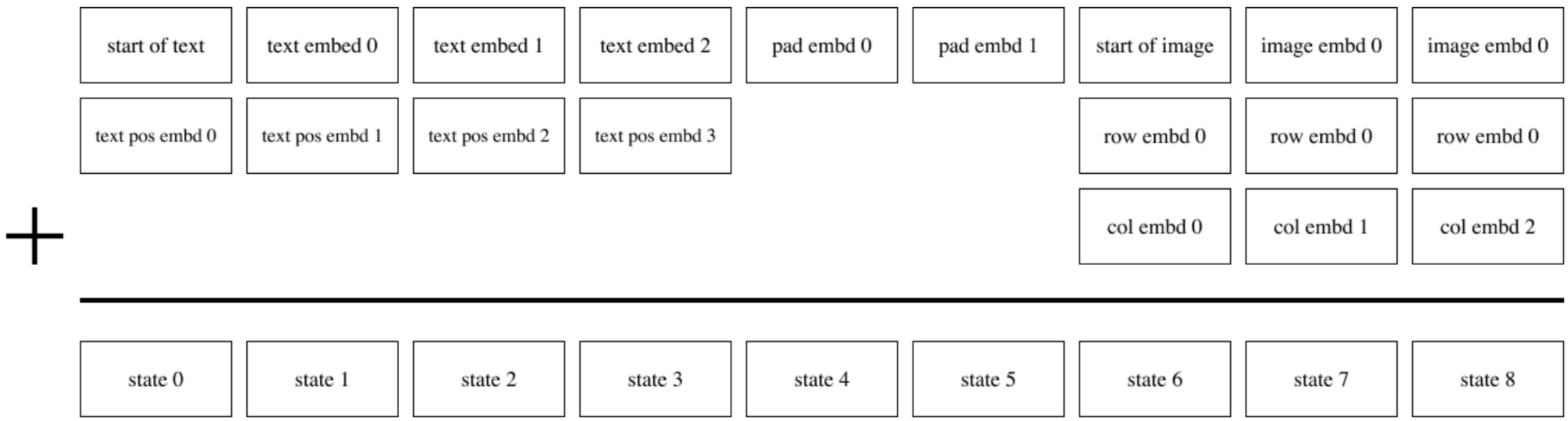
- **Stage 2**: Concatenate up to 256 text tokens with the 32x32 (=1024) image tokens, and train an autoregression transformer to model the joint distribution of the text and image tokens.



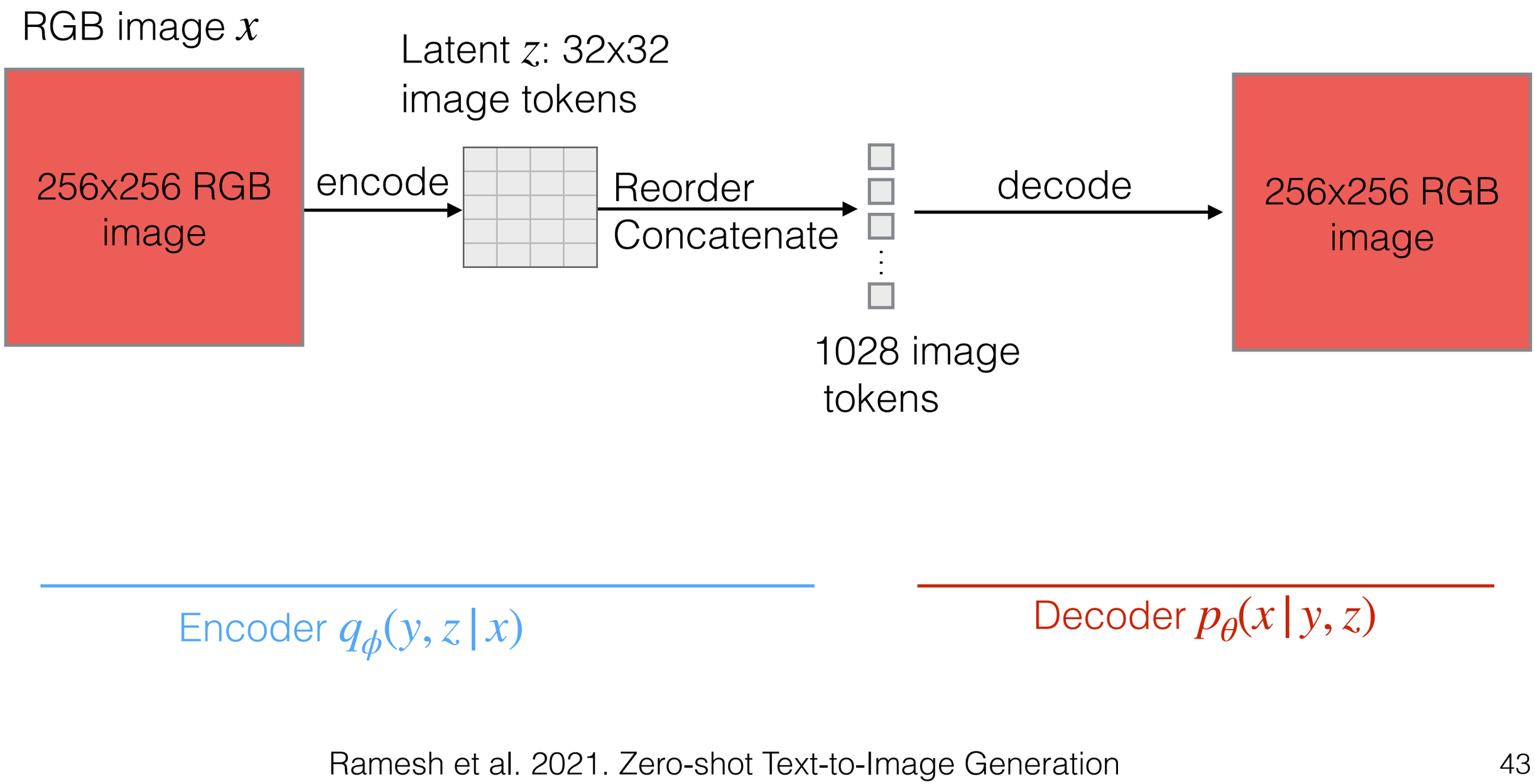
Autoregressive LM training: Maximum Likelihood Estimation $\max_{\psi} p_{\psi}(y, z)$

DALL-E: Language Model Training

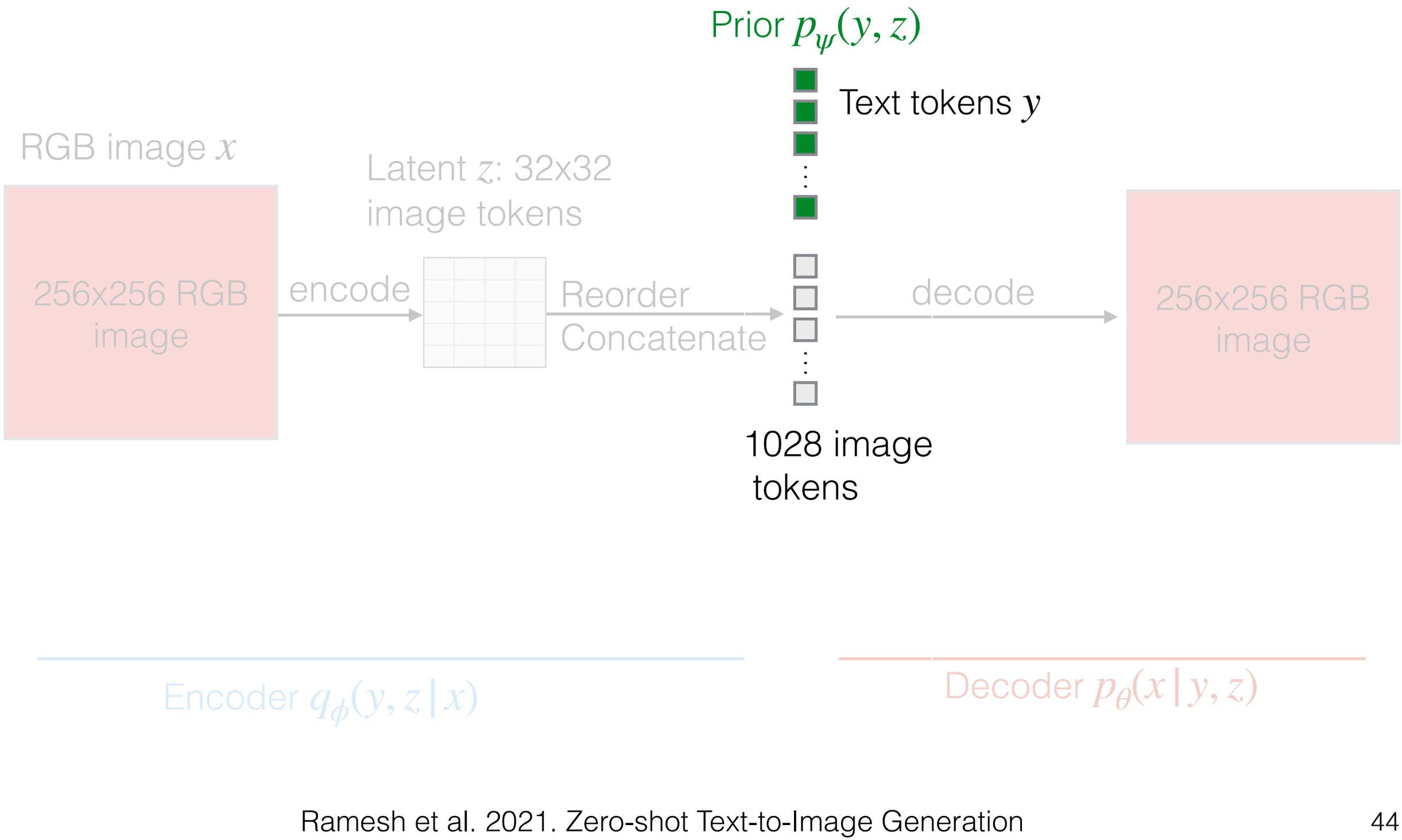
- Representation of the combined text + image token sequence



DALL-E: Stage 1



DALL-E: Stage 2



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

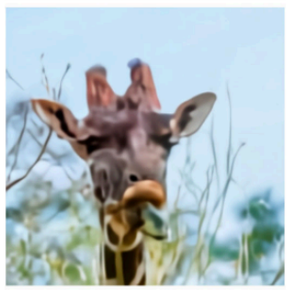
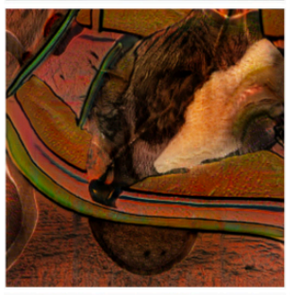
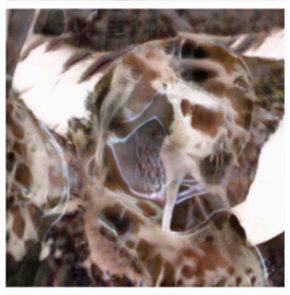
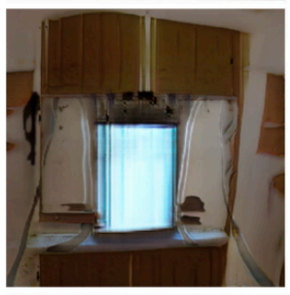
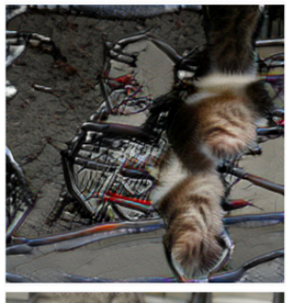
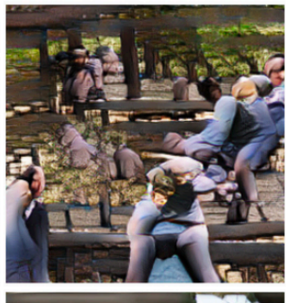
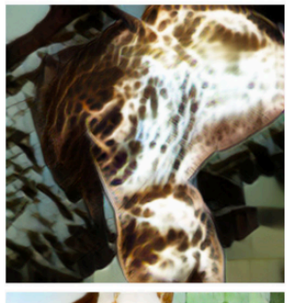
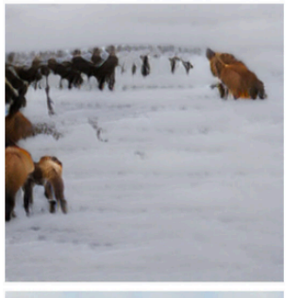
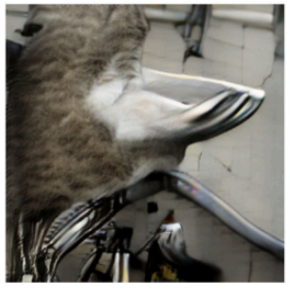
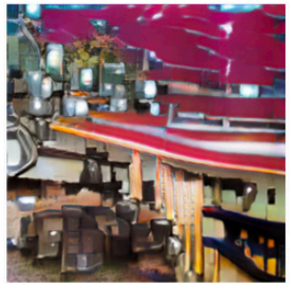
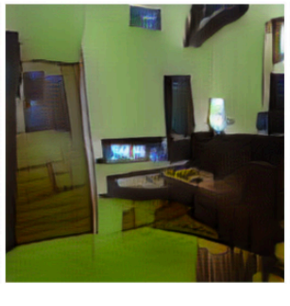
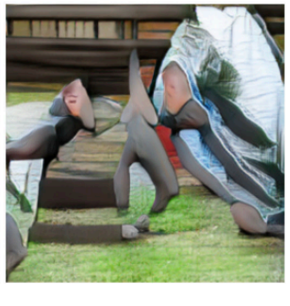
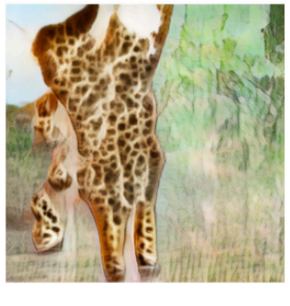
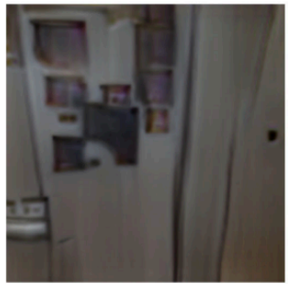
Stage 2 fixes p_θ, q_ϕ and updates p_ψ

- 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.

DALL-E: Test Time

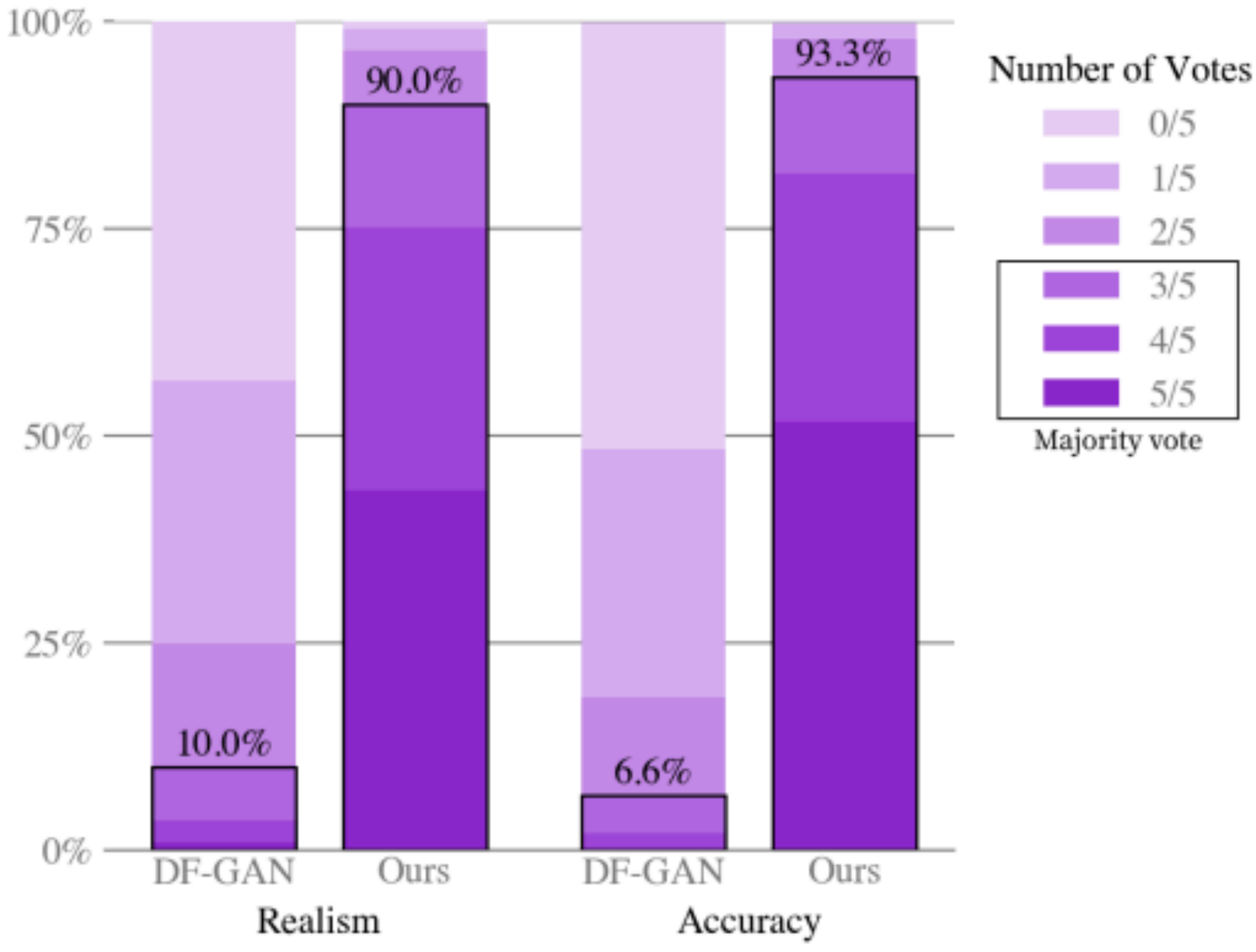
- Given a text prompt y, use the prior distribution (LM) to sample a sequence of 1028 image tokens
- Re-order 1028 image tokens to 32x32 shape
- Use dVAE's decoder to generate a RGB image from the image tokens.

Text-to-Image Generation

	a very cute cat laying by a big bike.	china airlines plain on the ground at an airport with baggage cars nearby.	a table that has a train model on it with other cars and things	a living room with a tv on top of a stand with a guitars sitting next to	a couple of people are sitting on a wood bench	a very cute giraffe making a funny face.	a kitchen with a fridge, stove and sink	a group of animals are standing in the snow.
Validation								
Ours								
DF-GAN								
DM-GAN								
AttnGAN								

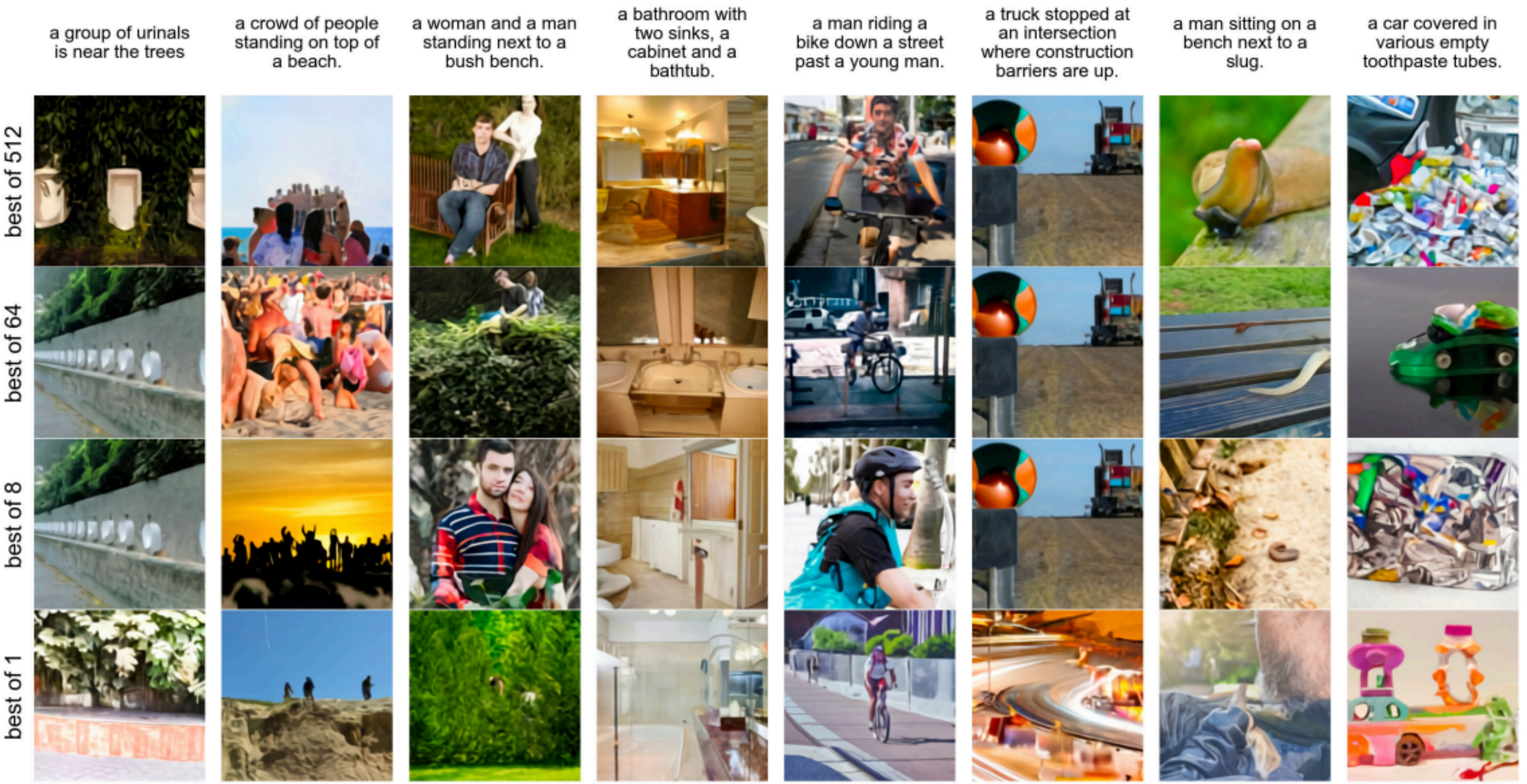
Human Eval on “Realism” and “Accuracy”

- DALL-E outperforms DF-GAN



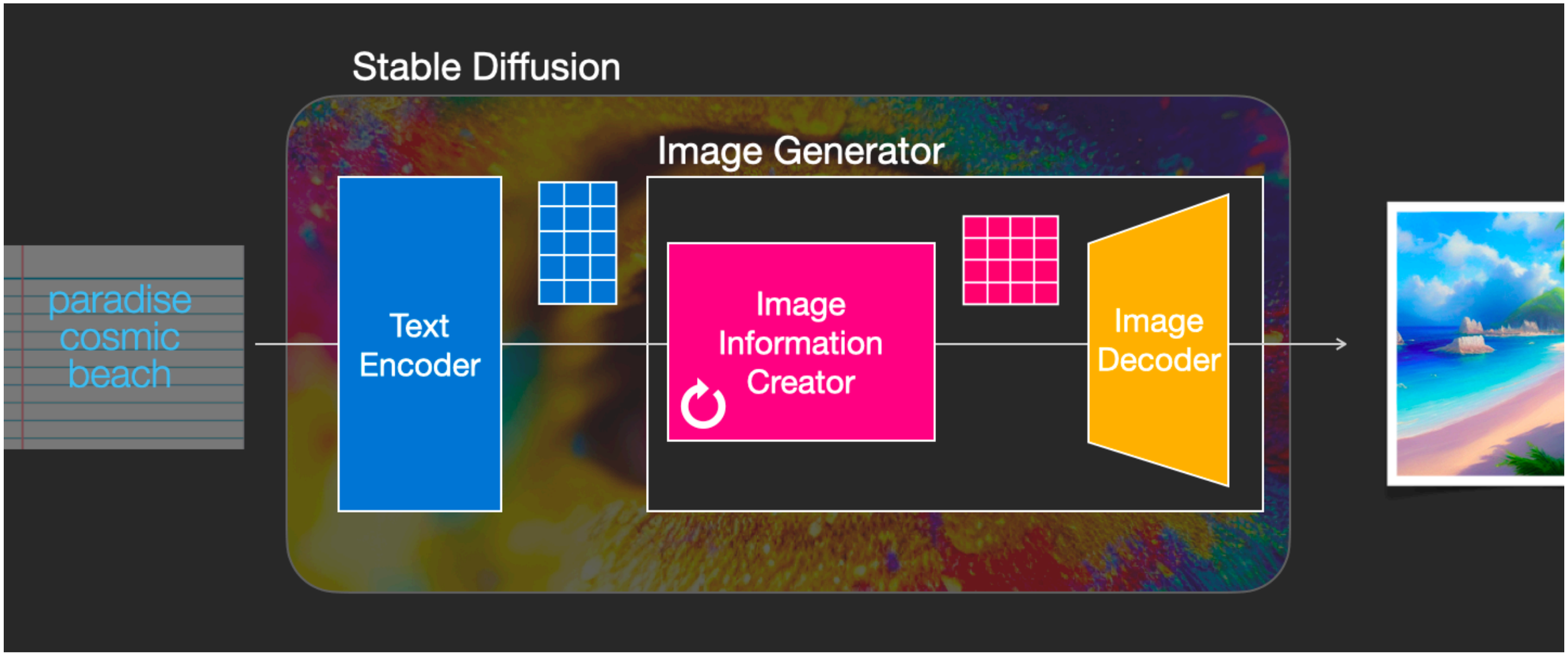
Sample, then Re-rank

- 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

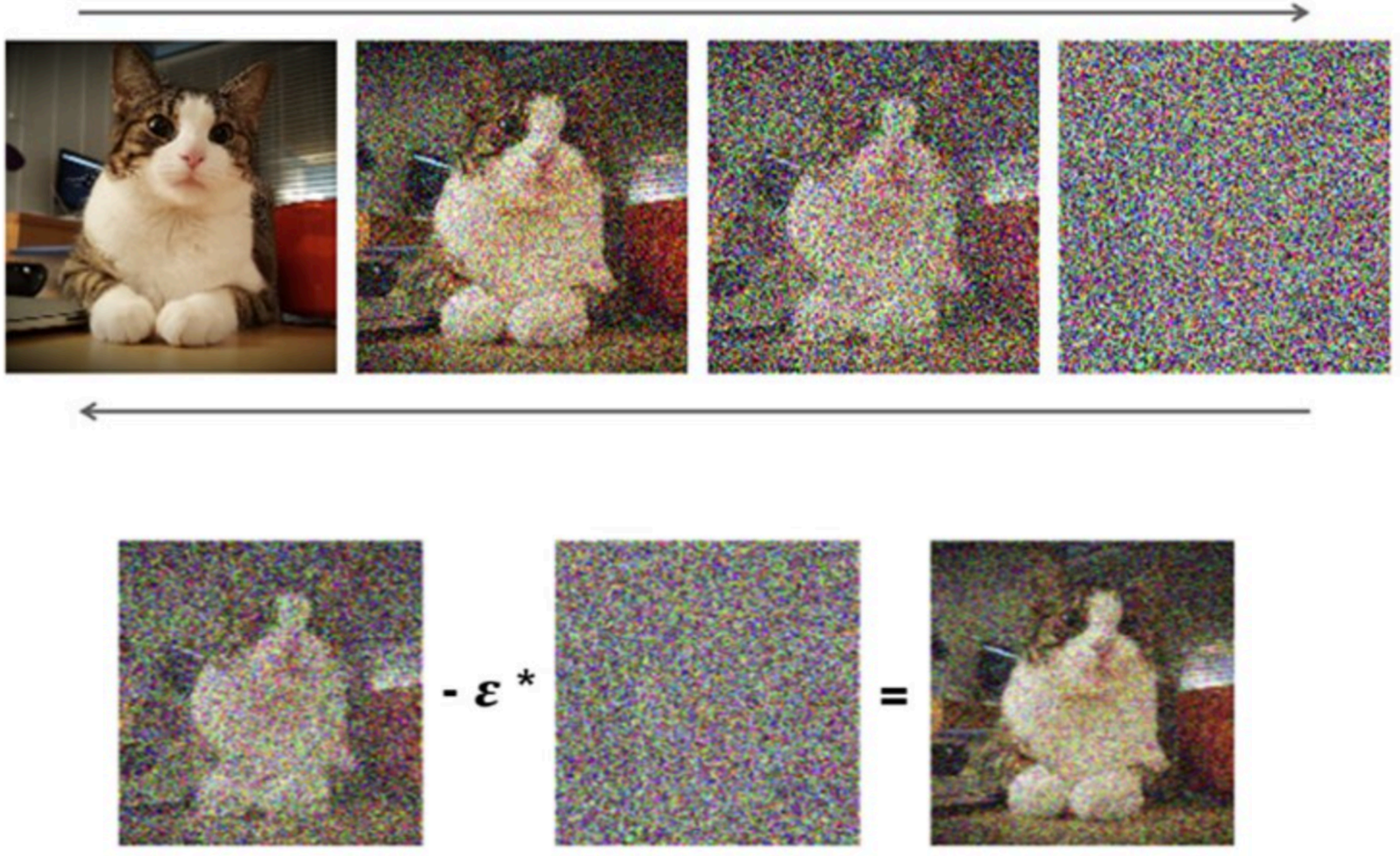
- Stable diffusion: Latent diffusion model



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

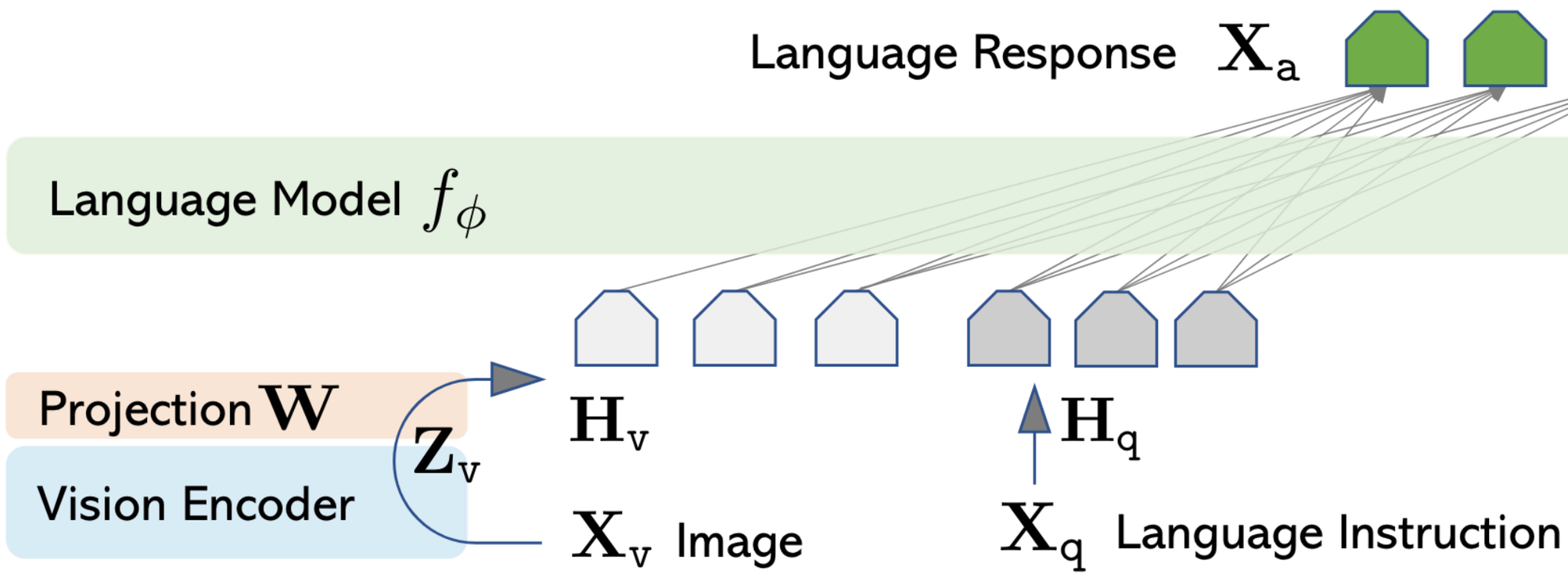
More Text-to-Image Generation Models

- Stable diffusion: Add noise & remove noise



More recent Vision-Language Models

- **LLAVA**: Concatenate the projected image patch sequence with texts

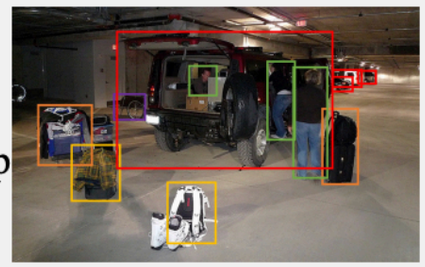


Treating images as a “foreign” language

Creating Synthetic Data for Vision Instruction Tuning

- Start with image-text pair datasets (with bounding boxes annotations), generate synthetic instruction-tuning datasets

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Two-Stage Training

- Stage 1: Pre-training for Feature Alignment.
  - Only pre-train the project matrix ***W*** that maps image tokens to the text token embedding space
  - Pre-train the model to describe the image in language
- Stage 2: Fine-tuning End-to-End.
  - Fine-tune the whole model parameters (including the projection matrix and the LLM decoder), keep the vision encoder frozen
  - Fine-tune on multi-turn conversation data

$$p(\mathbf{X}_\mathtt{a}|\mathbf{X}_\mathtt{v}, \mathbf{X}_\mathtt{instruct}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_\mathtt{v}, \mathbf{X}_{\mathtt{instruct},<i}, \mathbf{X}_{\mathtt{a},<i}),$$

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_\mathtt{a}$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_\mathtt{a}$ <STOP> $\cdots$

InternVL Family: Closing the Gap to Commercial Multimodal Models with Open-Source Suites —— A Pioneering Open-Source Alternative to GPT-5

[🆕 Blog] [🤔 FAQs] [🗨️ Chat Demo] [📖 Document] [🌐 API] [🚀 Quick Start]

[🔥 InternVL3.5 Report] [📜 InternVL3.0 Report] [📜 InternVL2.5 MPO] [📜 InternVL2.5 Report]

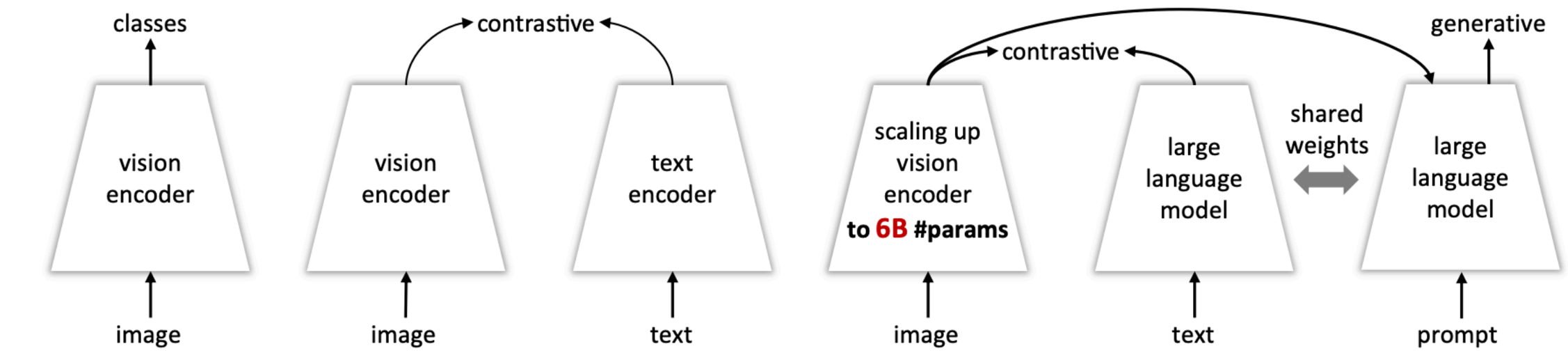
[📜 Mini-InternVL Paper] [📜 InternVL2 Blog] [📜 InternVL 1.5 Paper] [📜 InternVL 1.0 Paper]

[📖 2.0 中文解读] [📖 1.5 中文解读] [📖 1.0 中文解读]

Switch to the Chinese version (切换至中文版)

InternVL

- https://github.com/OpenGVLab/InternVL
- Scale up model size for VLMs

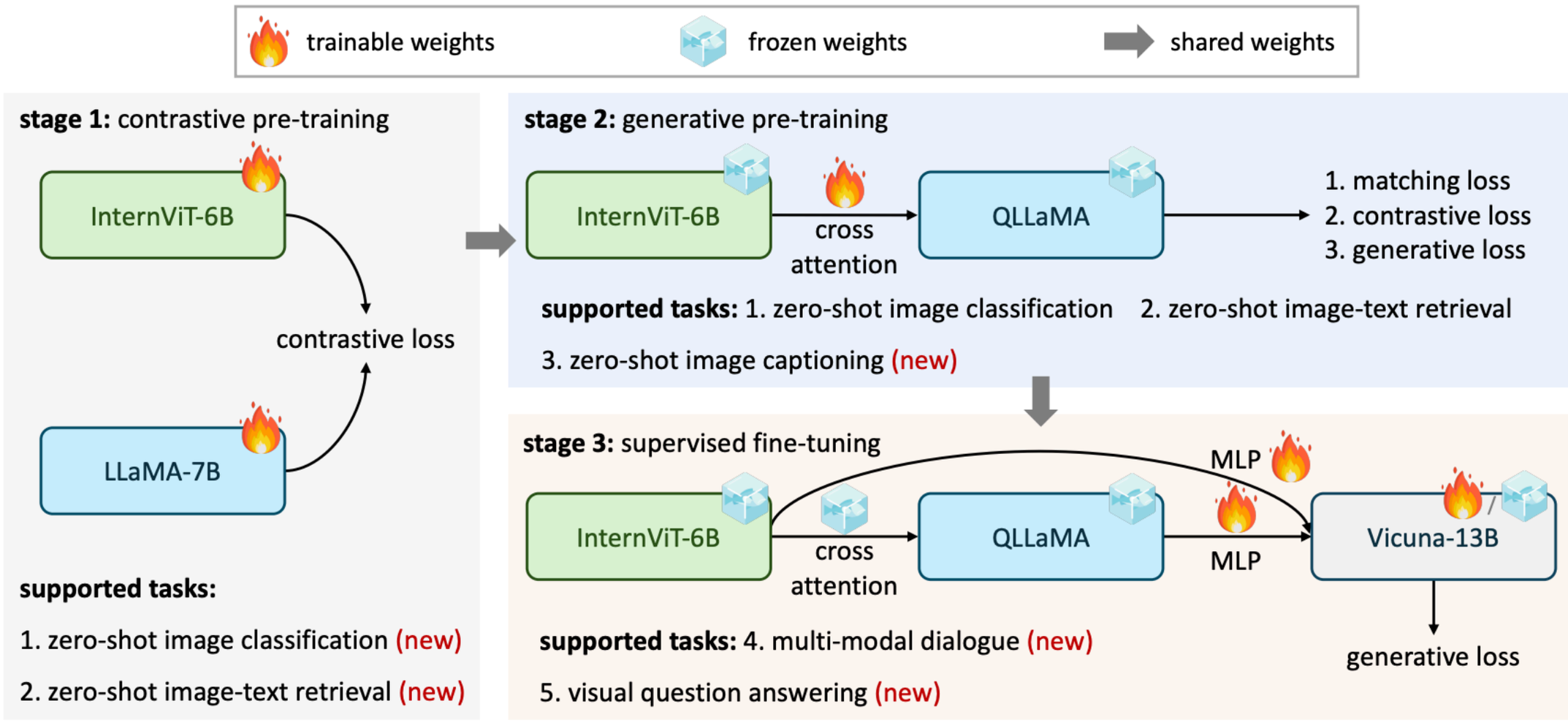


(a) Supervised pre-training (b) Contrastive pre-training (c) InternVL: Scaling up vision encoder and aligning with LLM (ours)

Figure 1. **Comparisons of different vision and vision-language foundation models.** (a) indicates the traditional vision foundation model, *e.g.* ResNet [57] pre-trained on classification tasks. (b) represents the vision-language foundation models, *e.g.* CLIP [117] pre-trained on image-text pairs. (c) is our InternVL, which presents a workable way to align the large-scale vision foundation model (*i.e.*, InternViT-6B) with the large language model and is versatile for both contrastive and generative tasks.

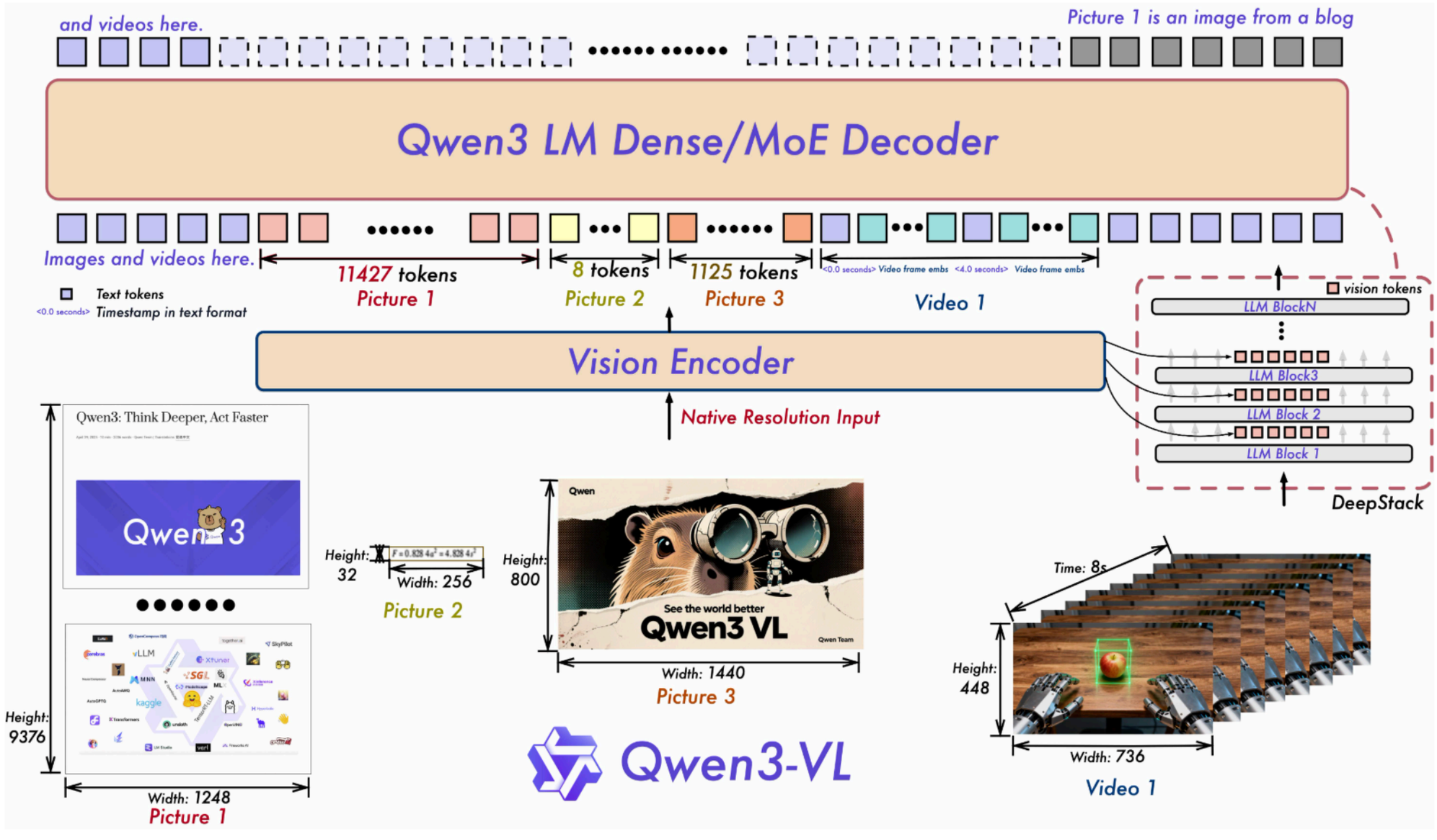
InternVL

- Demonstrate better zero-shot capabilities



Qwen3-VL

- https://github.com/QwenLM/Qwen3-VL



Qwen3-VL

- **Visual Agent**: Operates PC/mobile GUIs—recognizes elements, understands functions, invokes tools, completes tasks.
- **Visual Coding Boost**: Generates Draw.io/HTML/CSS/JS from images/videos.
- **Advanced Spatial Perception**: Judges object positions, viewpoints, and occlusions; provides stronger 2D grounding and enables 3D grounding for spatial reasoning and embodied AI.
- **Long Context & Video Understanding**: Native 256K context, expandable to 1M; handles books and hours-long video with full recall and second-level indexing.
- **Enhanced Multimodal Reasoning**: Excels in STEM/Math—causal analysis and logical, evidence-based answers.
- **Upgraded Visual Recognition**: Broader, higher-quality pretraining is able to “recognize everything”—celebrities, anime, products, landmarks, flora/fauna, etc.
- **Expanded OCR**: Supports 32 languages (up from 10); robust in low light, blur, and tilt; better with rare/ancient characters and jargon; improved long-document structure parsing.
- **Text Understanding on par with pure LLMs**: Seamless text–vision fusion for lossless, unified comprehension.

Qwen3-VL

1. **Interleaved-MRoPE**: Full-frequency allocation over time, width, and height via robust positional embeddings, enhancing long-horizon video reasoning.
2. **DeepStack**: Fuses multi-level ViT features to capture fine-grained details and sharpen image–text alignment.
3. **Text–Timestamp Alignment:** Moves beyond T-RoPE to precise, timestamp-grounded event localization for stronger video temporal modeling.

Questions?