Vision and Text Integration

Naeemullah Khan
naeemullah.khan@kaust.edu.sa

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

LMH
Lady Margaret Hall

July 25, 2025

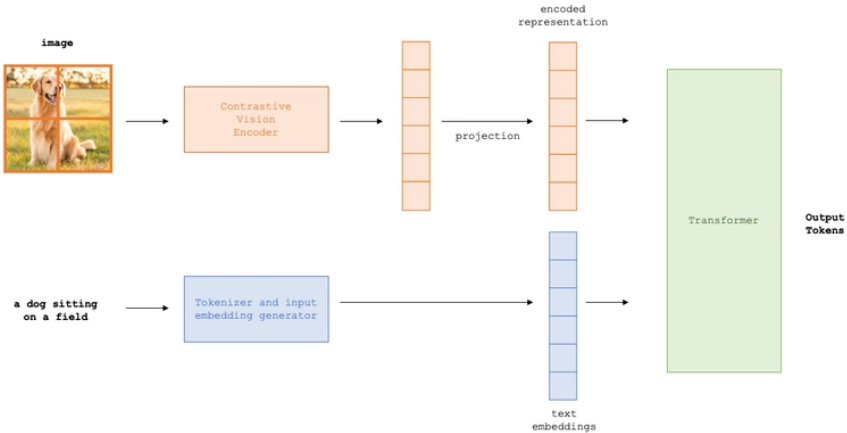


image
Contrastive Vision Encoder
projection
encoded representation
Transformer
Output Tokens
a dog sitting on a field
Tokenizer and input embedding generator
text embeddings

Table of Contents

Table of Contents (cont.)

Motivation

- The human brain naturally integrates vision and language (e.g., describing scenes, reading captions).
- This inspires artificial systems to process and relate visual and textual information.
- Recent applications:
  - Image captioning
  - Visual question answering (VQA)
  - Generative AI

Motivation (cont.)

- Joint vision-language understanding is crucial for intelligent systems.
- Foundation for:
  - Autonomous agents
  - Assistive technologies
  - Content creation tools

Learning Outcomes

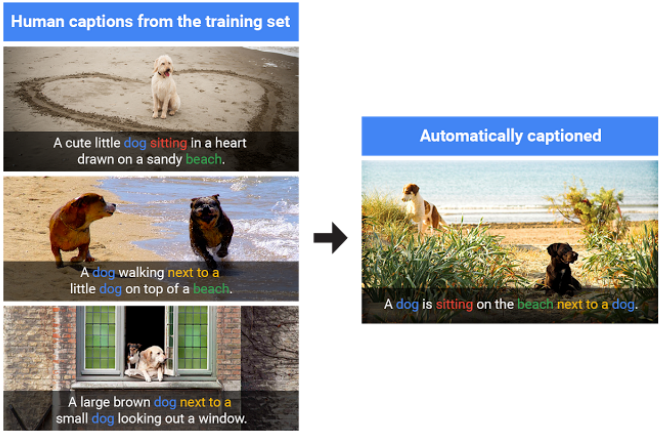
By the end of this session, learners will be able to:

- Understand how vision and text can be jointly modeled.
- Explore state-of-the-art models such as CLIP and diffusion-based text-to-image (T2I) methods.
- Implement and evaluate image and video captioning systems.
- Grasp classifier-free guidance and text-conditioned generation techniques.
- Analyze the limitations of current approaches and propose future research directions.

Vision and Text Integration: **Image and Video Captioning**

What is Captioning?

- **Definition:** Generating a descriptive text from an image or a video.

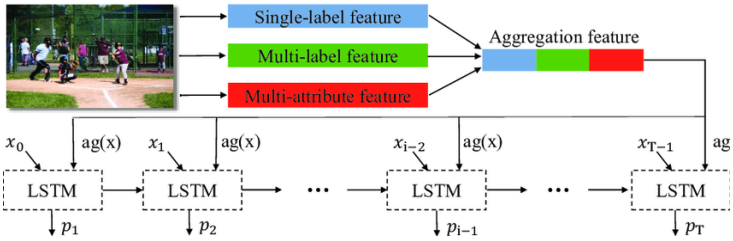


What is Captioning? (cont.)

- **Challenges:**
  - Semantics
  - Scene understanding
  - Temporal context
- **Applications:**
  - Accessibility
  - Content indexing
  - Social media
  - Surveillance

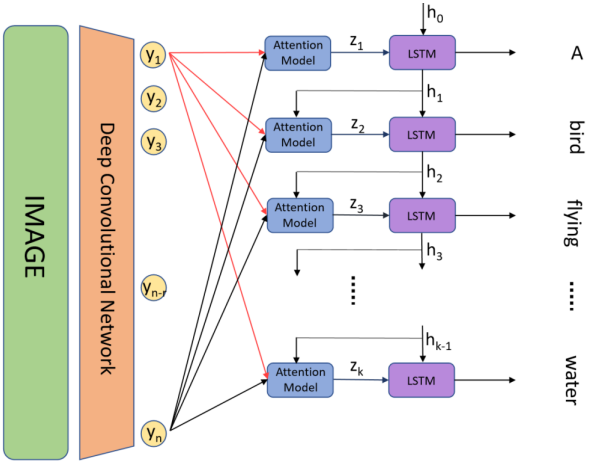
Image Captioning Pipeline

- **Feature Extraction:** Use a CNN (e.g., ResNet) to extract visual features from the image.
- **Language Modeling:** Use an RNN, LSTM, or Transformer to generate a textual description based on the extracted features.
- **Encoder-Decoder Architecture:** The CNN acts as the encoder and the RNN/LSTM/Transformer as the decoder.



Typical image captioning pipeline.

Image Captioning Pipeline (cont.)

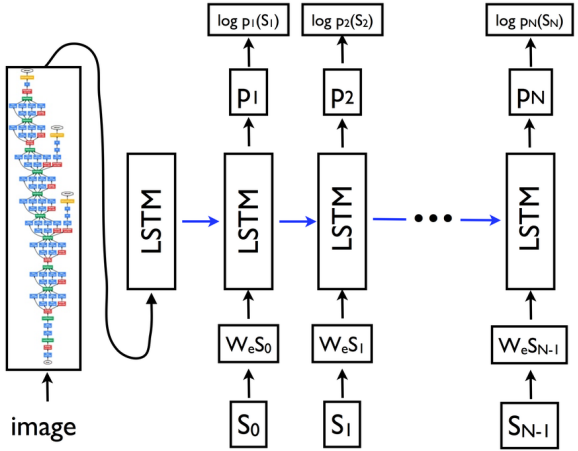


Active Learning in Image Captioning

Example: Show and Tell (Vinyals et al., 2015)

- One of the first successful end-to-end image captioning models.
- Uses a CNN (Inception) as encoder and an LSTM as decoder.
- Trained to maximize the likelihood of the correct caption given the image.

Example: Show and Tell (Vinyals et al., 2015) (cont.)



Show and Tell model architecture (Vinyals et al., 2015).

Example: Show and Tell (Vinyals et al., 2015) (cont.)

Helping visually impaired people better understand the content of images on the web.

– Pascal dataset
– Flickr30k dataset
– SBU dataset
– COCO dataset

$I \rightarrow$ input image
$S = \{S_1, S_2, \ldots\} \rightarrow$ target sequence of words
$p(S|I) \rightarrow$ likelihood
S_t comes from a given dictionary
NIC: Neural Image Caption

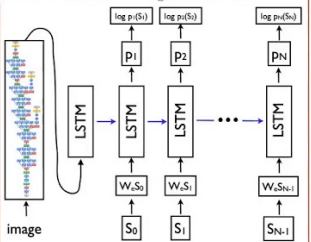
$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

$$\log p(S|I) = \sum_{t=0}^{N} \underbrace{\log p(S_t|I, S_0, \ldots, S_{t-1})}_{\text{model with an RNN}}$$

$$\text{LSTM} \quad h_{t+1} = f(h_t, x_t)$$

$$\text{CNN} \underbrace{\quad \text{input (image \& words)}}$$

$$\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
m_t &= o_t \odot c_t \\
p_{t+1} &= \text{Softmax}(m_t)
\end{aligned}$$



$$\begin{aligned}
x_{-1} &= \text{CNN}(I) \\
x_t &= W_e S_t, \quad t \in \{0 \ldots N-1\} \\
p_{t+1} &= \text{LSTM}(x_t), \quad t \in \{0 \ldots N-1\}
\end{aligned}$$

$S_0 \rightarrow$ start symbol
$S_N \rightarrow$ end symbol
$S_t \rightarrow$ one-hot-vector
$W_e \rightarrow$ word embeddings

$$L(I,S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

BeamSearch: Iteratively consider the set of the k best sentences up to time t as candidates to generate sentences of size $t+1$, and keep only the resulting best k of them. (beam size: k = 20)

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

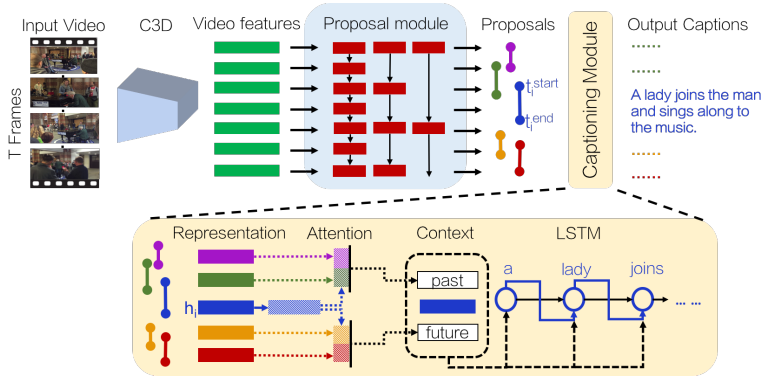
Example: Show and Tell (Vinyals et al., 2015) (cont.)

Figure 5. A selection of evaluation results, grouped by human rating.

Video Captioning

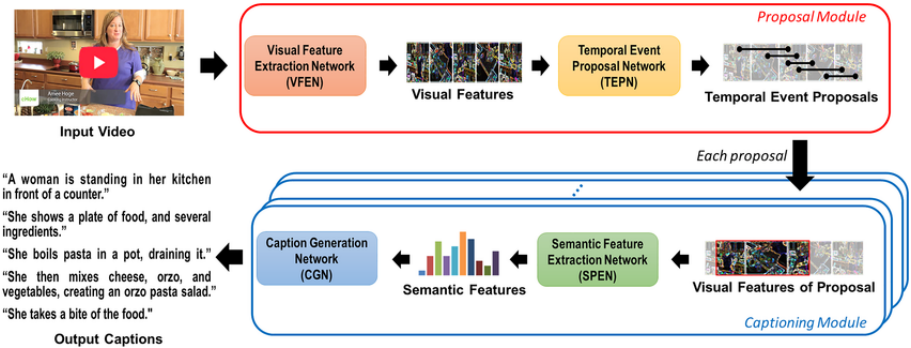
- **Video Captioning:** Generating descriptive text for a sequence of video frames.
- **Temporal Sequence Modeling:** Adds modeling of temporal dependencies between frames, in addition to spatial features.
- **Common Architecture:**
  - CNN for frame-level feature extraction.
  - RNN (e.g., LSTM/GRU) for modeling temporal sequence.
  - Attention mechanisms to focus on relevant frames or regions.
- **Recent Advances:**
  - Transformer-based models for parallel sequence modeling.
  - Memory networks for capturing long-term dependencies.

Video Captioning (cont.)



Dense-Captioning Events in Videos (Hendricks et al., 2017)

Video Captioning (cont.)



Video Captioning (cont.)

"A woman is standing in her kitchen in front of a counter."

"She shows a plate of food, and several ingredients."

"She boils pasta in a pot, draining it."

"She then mixes cheese, orzo, and vegetables, creating an orzo pasta salad."

"She takes a bite of the food."

Challenges in Captioning

- **Ambiguity:** Multiple valid captions for the same image or video.
- **Context Understanding:** Captions need to capture context and semantics accurately.
- **Temporal Dynamics:** In video captioning, capturing the dynamics of actions and events over time is crucial.
- **Data Scarcity:** High-quality annotated datasets for training captioning models are limited.
- **Evaluation Metrics:** Evaluating caption quality is subjective and often relies on metrics like BLEU, METEOR, or CIDEr, which may not align with human judgment.

Image Captioning Advances: OSCAR: Object-Semantics Aligned Pre-training

for Vision-and-Language Tasks.

OSCaR: Object State Captioning and State Change Representation

Nguyen Nguyen[1], Jing Bi[1], Ali Vosoughi[1], Yapeng Tian[2], Pooyan Fazli[3], Chenliang Xu[1]
[1]University of Rochester, [2]University of Texas at Dallas, [3]Arizona State University
{nguyen.nguyen, jing.bi, ali.vosoughi, chenliang.xu}@rochester.edu,
yapeng.tian@utdallas.edu, pooyan@asu.edu

Abstract

The capability of intelligent models to extrapolate and comprehend changes in object states is a crucial yet demanding aspect of AI research, particularly through the lens of human interaction in real-world settings. This task involves describing complex visual environments, identifying active objects, and interpreting their changes as conveyed through language. Traditional methods, which isolate object captioning and state change detection, offer a limited view of dynamic environments. Moreover, relying on a small set of symbolic words to represent changes has restricted the expressiveness of language. To address these challenges, in this paper, we introduce the Object State Captioning and State Change Representation (OSCaR) dataset and benchmark. OSCaR consists of 14,084 annotated video segments with nearly

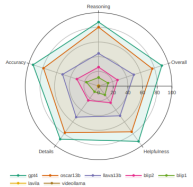


Figure 1: **Surpassing prior models in aligning with human judgements.** Our method achieves near parity with GPT-4V ratings across helpfulness, accuracy, reasoning, and other key metrics.

OSCAR: Object-Semantics Aligned Pre-training for Vision-and-Language Tasks.

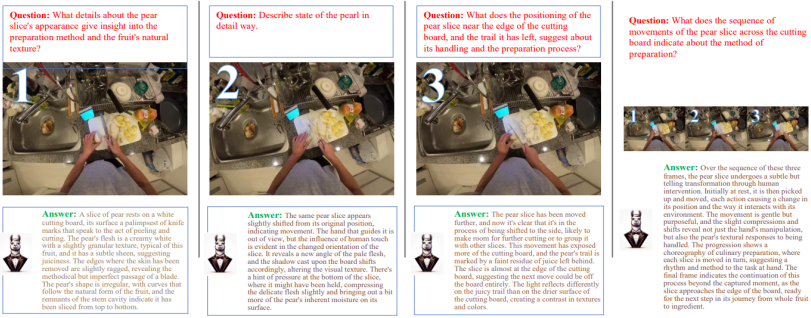


Figure 2: **OSCaR's description of state, state change, and illustration of reasoning.** State description involves the characterization of a specific region of interest within the video and the associated activity. State change entails the description of the evolution of a system over a defined temporal sequence. Furthermore, the analysis of the state of an object is centered on comprehending and elucidating the mechanisms underlying the object's evolution.

VinVL: Large-scale pre-training with improved object-visual representations.

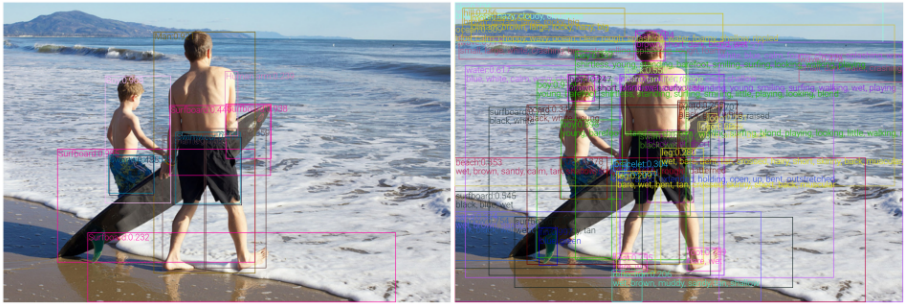
VinVL: Revisiting Visual Representations in Vision-Language Models

Pengchuan Zhang[♡†] Xiujun Li[♡♠†] Xiaowei Hu[♡] Jianwei Yang[♡] Lei Zhang[♡]
Lijuan Wang[♡] Yejin Choi[♠] Jianfeng Gao[♡]

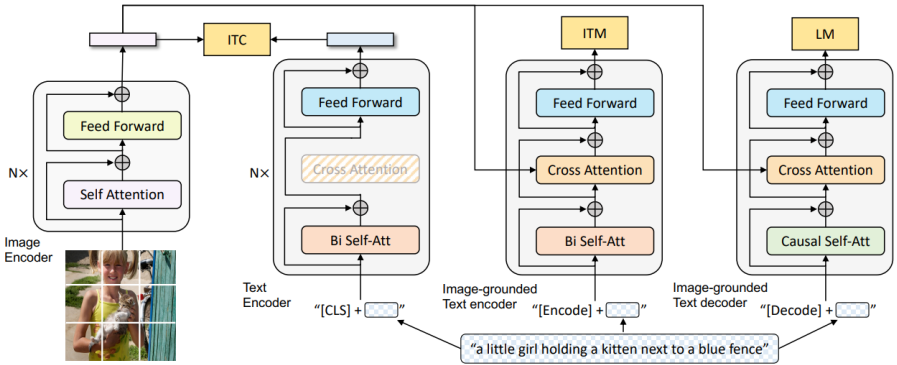
March 11, 2021

Recent Advances in Captioning (cont.)

VinVL: Large-scale pre-training with improved object-visual representations.



BLIP: Bootstrapped Language-Image Pretraining for unified vision-language understanding and generation.



Recent Advances in Captioning (cont.)

Video Captioning Advances:

- **MARN:** Multi-modal Attentive Recurrent Network for video captioning.
- **STG-KD:** Spatio-Temporal Graph Knowledge Distillation for video captioning.
- **ClipCap:** Zero-shot image-to-text generation using CLIP and language models.

Key Datasets

- **MSCOCO:** Large-scale dataset for image captioning.
- **ActivityNet Captions:** Benchmark for dense video captioning.
- **YouCook2:** Large-scale dataset of instructional cooking videos with captions.

Image and Video Captioning: **Multimodal Models**

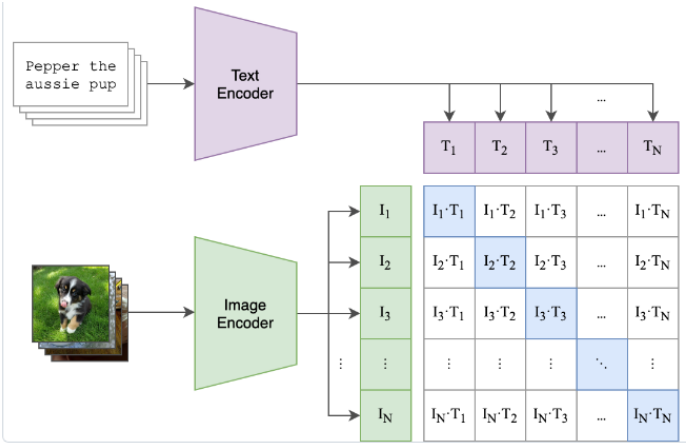
Multimodal Models

- **Definition:** Multimodal models integrate visual and textual information to generate descriptive captions for images and videos.
- **Challenges:**
  - Ambiguity in visual content
  - Context understanding
  - Temporal dynamics in videos
- **Applications:**
  - Image captioning
  - Video captioning
  - Assistive technologies (e.g., for visually impaired)

Multimodal Representation Learning

- **Goal:** Learn shared representations for different modalities (e.g., images and text).
- **Approach:** Map both images and text into a joint embedding space.
- **Benefit:** Enables comparison and retrieval across modalities.

Multimodal Representation Learning (cont.)

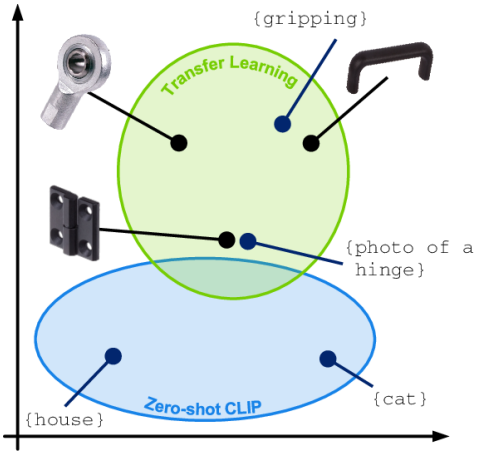


Multimodal representation learning: mapping images and text into a joint embedding space.

Joint Embedding Spaces for Image-Text Pairs

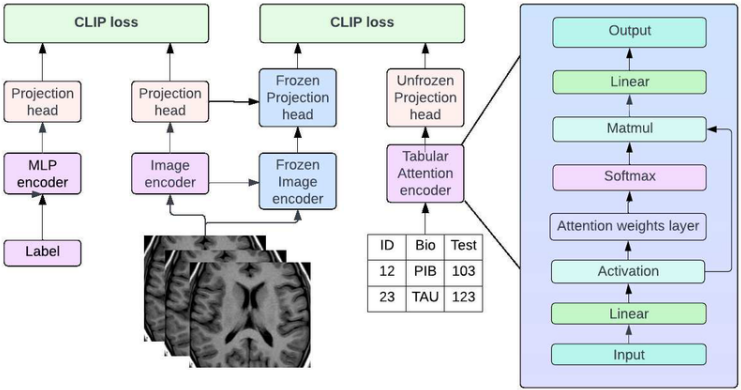
- Images and their corresponding captions are projected into a common vector space.
- Matching image-text pairs are close together; non-matching pairs are far apart.
- Used in tasks like cross-modal retrieval and caption generation.

Joint Embedding Spaces for Image-Text Pairs (cont.)



Joint embedding space for image-text pairs: matching pairs are close, non-matching pairs are distant.

- **Contrastive Objective:** Pull matching image-text pairs together, push non-matching pairs apart.
- **Popular Methods:** CLIP, ALIGN, etc.
- **Loss Function:** Often uses InfoNCE or similar contrastive losses.



Contrastive learning for multimodal models: aligning image-text pairs in a joint embedding space.

CLIP (Contrastive Language–Image Pretraining)

- **Overview:** CLIP is a multimodal model trained to connect images and text using contrastive learning.
- **Training Data:** 400 million image-text pairs collected from the internet.
- **Architecture:**
  - Separate Transformer encoders for images and text.
  - Both modalities are mapped into a shared embedding space.
- **Contrastive Loss:** Uses dot-product similarity to align matching image-text pairs and separate non-matching pairs.
- **Zero-shot Classification:** Enables classification of images without task-specific training by comparing image embeddings to text prompt embeddings.

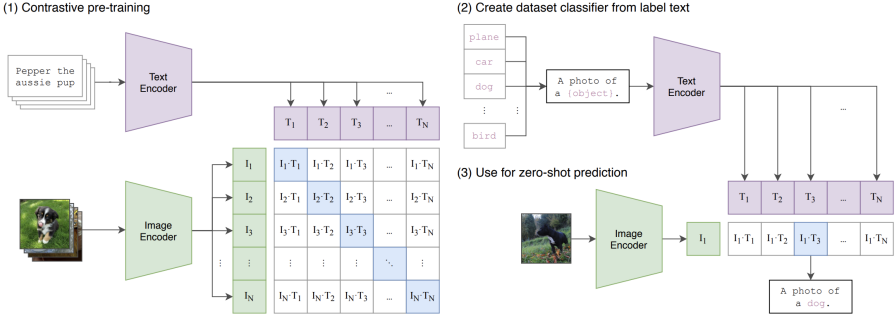


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

CLIP architecture: separate encoders for images and text, mapping both into a shared embedding space.

Why CLIP Matters:

- **Open-vocabulary vision models:** CLIP can recognize and describe a wide range of concepts beyond fixed label sets.
- **Robust to distribution shift:** Performs well on images from domains not seen during training.
- **No task-specific training required:** Enables zero-shot learning for new tasks by leveraging text prompts.

Limitations of CLIP:

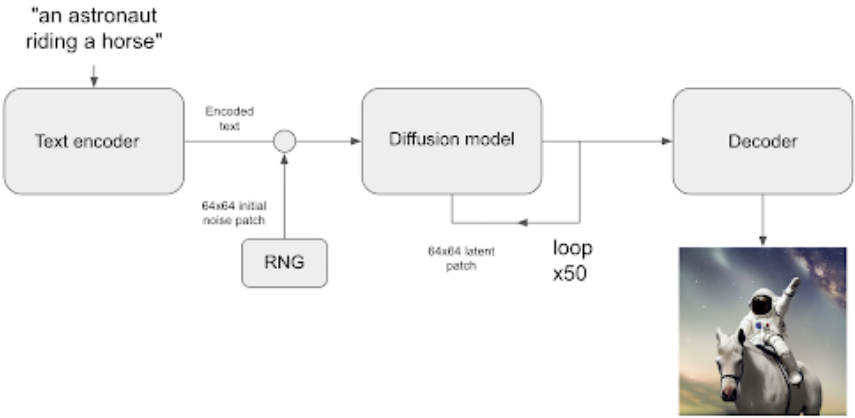
- **Struggles with fine-grained details:** May fail to distinguish subtle differences between visually similar objects.
- **Lacks true reasoning abilities:** Cannot perform complex reasoning or understand relationships beyond surface-level associations.
- **Prone to dataset biases:** Inherits biases present in the large-scale web data used for training.

Text-to-Image Generation Techniques

Text-to-Image Generation

- Text-to-image generation involves creating images from textual descriptions.
- Key techniques include:
  - Diffusion models
  - Generative adversarial networks (GANs)
  - Variational autoencoders (VAEs)
- Applications:
  - Art generation
  - Content creation
  - Virtual environments

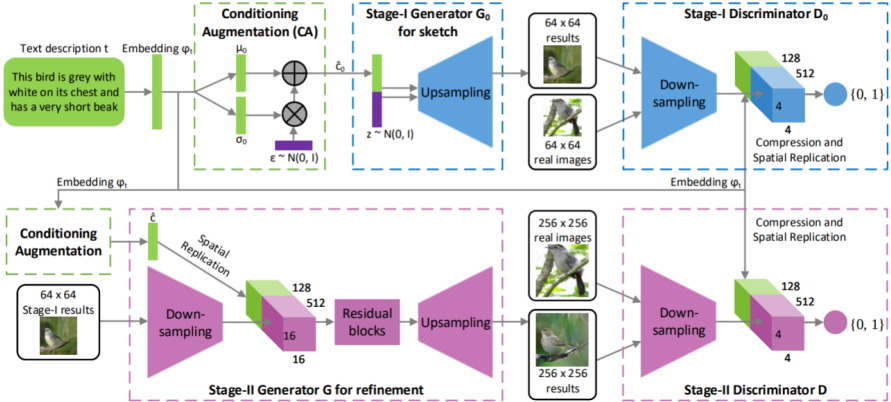
Text-to-Image Generation (cont.)



Text-to-Image Generation Overview

- **Goal:** Generate realistic images conditioned on text prompts.
- Requires alignment between vision and language features.
- **Applications:**
  - Design
  - Entertainment
  - Accessibility

StackGAN: Generates images in multiple stages, from coarse to fine details.



GAN-based Approaches (cont.)

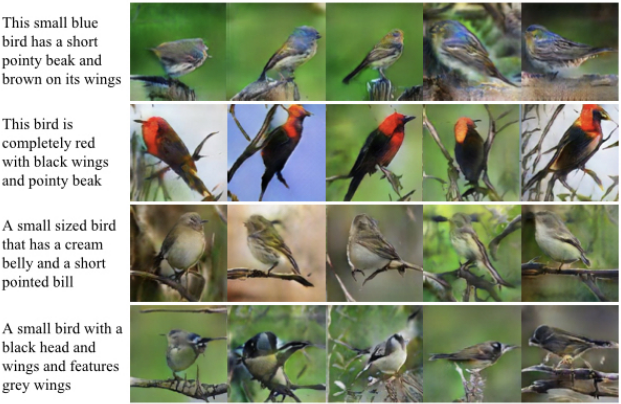
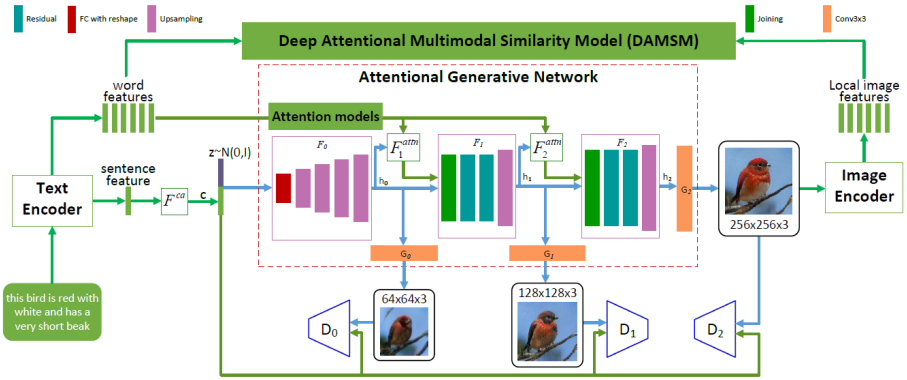


Figure 7. Birds with different poses and viewpoints generated with the same input text embedding by our StackGAN. The noise vector z and text embedding are fixed for each row.

AttnGAN: Incorporates attention mechanisms to better align image regions with words in the text.



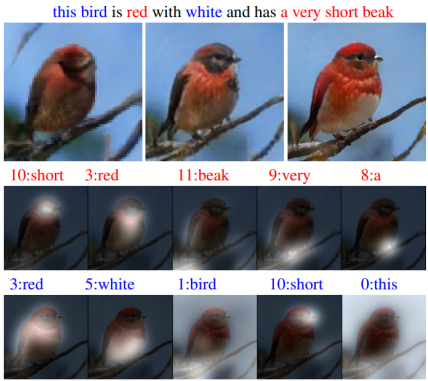


Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0, G_1 and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

GAN-based Approaches (cont.)

Stage-wise Generation:

- Images are generated progressively, refining details at each stage.

Limitations:

- Training instability
- Mode collapse (lack of diversity in generated images)

Diffusion-based Approaches

- Diffusion models generate images by iteratively denoising random noise, conditioned on text prompts.
- Notable models:
  - **DALL•E 2**
  - **Imagen**
  - **Stable Diffusion**
- **Key idea:** Learn to reverse a gradual noising process, producing high-quality images from pure noise.
- **Advantages:**
  - Superior image quality and diversity compared to GANs and VAEs
  - Better alignment with textual descriptions

DALL•E 2: A diffusion model that generates images from text prompts, achieving high fidelity and diversity.

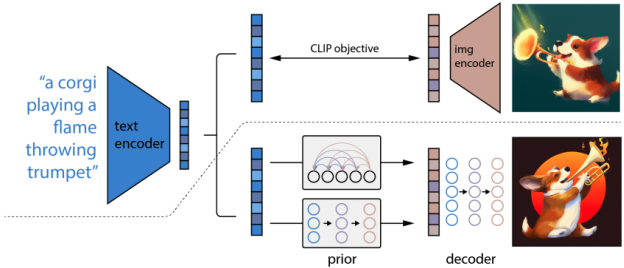


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

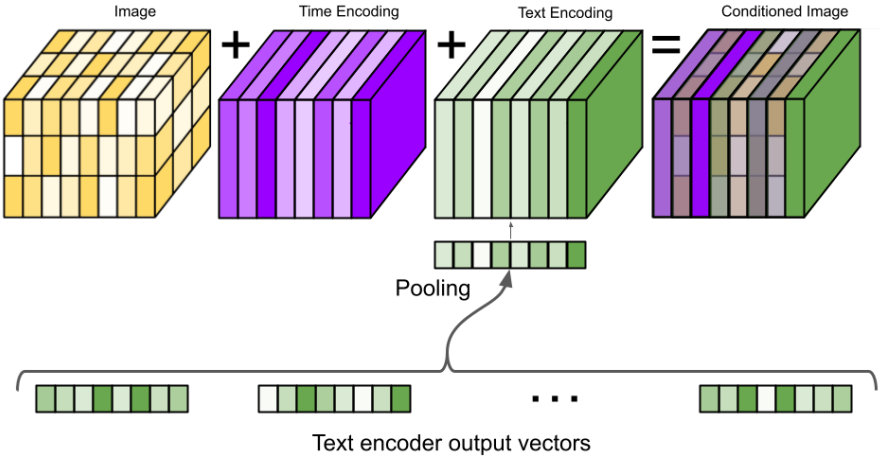
DALL•E 2 architecture (source: OpenAI)

Diffusion-based Approaches (cont.)

Diffusion-based Approaches (cont.)

Imagen: A diffusion model by Google that generates high-quality images from text, focusing on photorealism and fine details.

Diffusion-based Approaches (cont.)

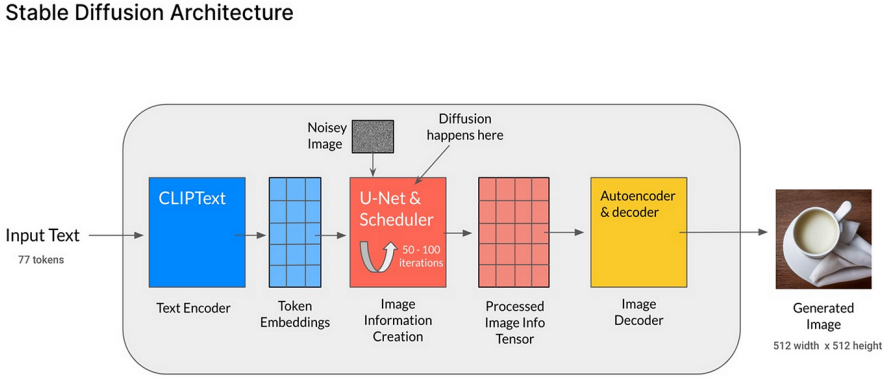


How Imagen works: from text to image generation (source: Google)

Diffusion-based Approaches (cont.)

Diffusion-based Approaches (cont.)

Stable Diffusion: An open-source diffusion model that generates high-quality images from text, designed for efficiency and accessibility.



Stable Diffusion architecture (source: Stability AI)

Diffusion-based Approaches (cont.)

Diffusion Model Architecture:

- **U-Net backbone:** The core neural network used for denoising, featuring an encoder-decoder structure with skip connections.
- **Text encoder:** Converts text prompts into embeddings. Common choices include T5 or CLIP Text Encoder.
- **Cross-attention:** Mechanism to inject text conditioning into the image generation process, allowing the model to align image features with textual information.

Text-Conditioned Generation

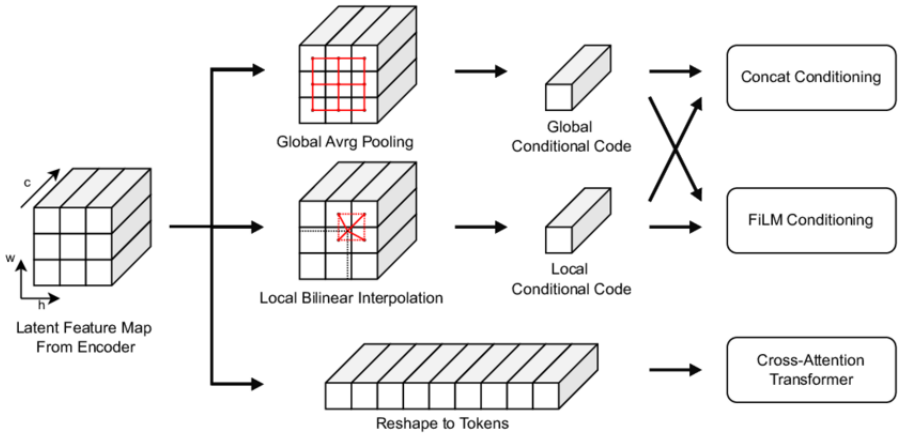
Text-to-Image Generation

- Text-to-image generation involves creating images from textual descriptions.
- Applications include art generation, content creation, and visual storytelling.
- Key challenges:
  - Ensuring high fidelity to the text description.
  - Maintaining diversity in generated images.
  - Handling complex and abstract descriptions.

Text Conditioning Strategies

- **Concatenation of Embeddings:**
  - Combine text and image embeddings before feeding them into the generative model.
  - Simple and effective for early fusion of modalities.
- **Cross-Attention in Transformer Layers:**
  - Use cross-attention mechanisms to allow the model to focus on relevant parts of the text while generating images.
  - Enables fine-grained alignment between text and image features.
- **Classifier-Free Conditioning:**
  - Train the model with and without conditioning information.
  - Allows flexible control over the influence of text during generation.

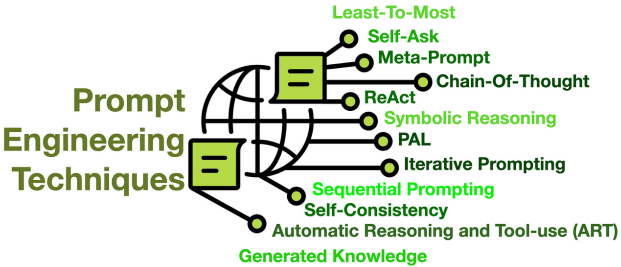
Text Conditioning Strategies (cont.)



A visualization of text conditioning strategies in text-to-image generation.

Prompt Engineering

- Crafting effective prompts is crucial for controlling the output of text-to-image models.
- Well-designed prompts can improve image quality, relevance, and adherence to user intent.
- Key aspects:
  - **Style Control:** Specify artistic styles (e.g., "in the style of Van Gogh").
  - **Object Presence:** Clearly mention desired objects and their attributes.
  - **Layout:** Indicate spatial relationships (e.g., "a cat sitting on a chair").

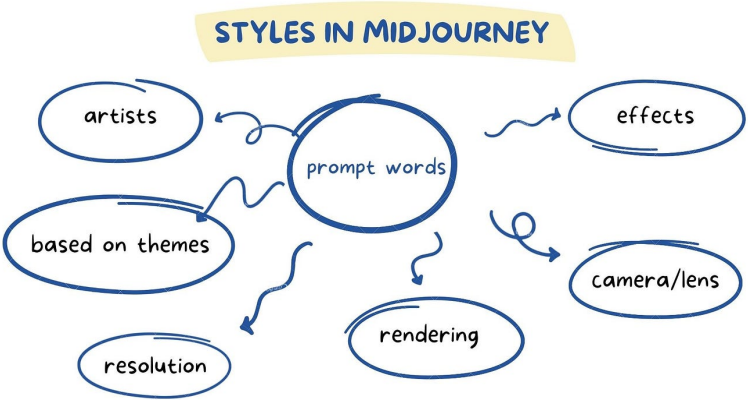

12 Prompt Engineering Techniques
Prompt Engineering Techniques
Least-To-Most
Self-Ask
Meta-Prompt
Chain-Of-Thought
ReAct
Symbolic Reasoning
PAL
Iterative Prompting
Sequential Prompting
Self-Consistency
Automatic Reasoning and Tool-use (ART)
Generated Knowledge
www.cobusgreyling.com

Prompt Engineering Examples

- **Midjourney:** ‘‘`A futuristic cityscape at sunset, vibrant colors, ultra-detailed, digital art`’’
- **DALL-E:** ‘‘`An armchair in the shape of an avocado`’’
- **SDXL:** ‘‘`A portrait of a medieval knight, oil painting, dramatic lighting`’’

- Experimenting with prompt phrasing can lead to diverse and creative outputs.

Prompt Engineering Examples (cont.)



An advanced guide to writing prompts for Midjourney.

Challenges in Conditioning

- **Ambiguity in Natural Language:** Text descriptions can be vague or open to multiple interpretations, making it difficult for models to generate the intended image.
- **Alignment with Low-Level Image Details:** Ensuring that fine-grained details in the image match the textual description remains a significant challenge.
- **Long Prompts and Hallucinations:** Handling lengthy or complex prompts can lead to hallucinated content or loss of important details in the generated images.

Challenges in Conditioning (cont.)

Some Prompt Engineering Challenges and Their Solutions.

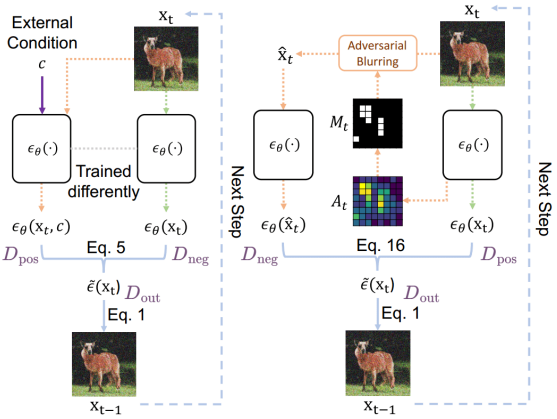
Classifier-Free Guidance (CFG)

What is Classifier-Free Guidance?

- A technique to improve the quality of text-to-image generation by controlling the influence of text prompts.
- It allows for generating images that are more aligned with the provided text while maintaining diversity.
- CFG is particularly useful in diffusion models, enhancing the coherence between text and generated images.

Key Properties of Classifier-Free Guidance

- **Balances realism and fidelity:** Adjusts the trade-off between generating realistic images and staying faithful to the text prompt.
- **No external classifier needed:** Removes the requirement for a separate classifier to guide the generation process.
- **Score interpolation:** Interpolates between conditional (with prompt) and unconditional (without prompt) model outputs to control guidance strength.



(a) Classifier-free guidance (b) Self-attention guidance

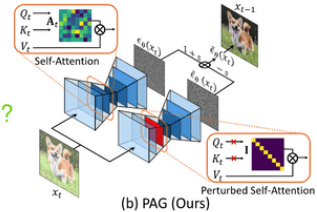
Illustration of Classifier-Free Guidance in text-to-image generation.

How does it work?

- The model generates images based on both the text prompt and random noise.
- By adjusting the guidance scale, users can control how much the text influences the final image.
- Higher guidance scales lead to images that are more closely aligned with the text, while lower scales allow for more creative freedom.

An overview of CLASSIFIER-FREE DIFFUSION GUIDANCE methods

- Training-free?
- Applicable to unconditional models?
- Single model?
- Train separate impaired model?



$$\hat{D}_{\text{out}}(x \mid \sigma) = D_{\text{neg}}(x \mid \sigma) + (1+\gamma)\left(D_{\text{pos}}(x \mid \sigma) - D_{\text{neg}}(x \mid \sigma)\right).$$

Illustration of Classifier-Free Guidance in text-to-image generation.

Implementation of Classifier-Free Guidance

- **Duplicate forward pass:** For each input, perform two forward passes through the model:
  - One with the conditioning (e.g., text prompt)
  - One without conditioning (unconditional)
- **Combine outputs:** The final score is computed as:

$$s = s_{\text{uncond}} + w \cdot (s_{\text{cond}} - s_{\text{uncond}})$$

  where:
  - s_{cond} is the model output with conditioning
  - s_{uncond} is the model output without conditioning
  - w is the guidance scale parameter
- **Effect of w:** Controls the strength of guidance. Higher w enforces stronger adherence to the prompt.

Benefits of Classifier-Free Guidance

- **Simpler and faster:** Eliminates the need for a separate classifier, reducing complexity and computational overhead.
- **Effective for text-guided diffusion:** Provides strong alignment between generated images and text prompts in diffusion models.
- **Foundation of modern T2I systems:** Forms the basis of Stable Diffusion XL (SDXL) and other state-of-the-art text-to-image generation systems.

Image and Video Captioning: **Summary**

Limitations

- Captioning lacks true understanding
- Multimodal alignment is shallow
- Hallucinations in T2I models
- Dataset bias and fairness issues
- High computational cost

Future Directions

- Vision-language reasoning (e.g., VQA with logic)
- 3D scene and video generation from text
- Interactive multimodal agents
- Multilingual and cultural alignment
- Personalized generation

Key Takeaways

- Vision and text integration unlocks powerful applications
- Techniques include captioning, CLIP, T2I, CFG
- Rapid evolution of generative multimodal models
- Ongoing challenges and exciting frontiers

Image and Video Captioning: **References**

References

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D.

Show and Tell: A Neural Image Caption Generator.

In *CVPR*, 2015.

https://arxiv.org/abs/1411.4555

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.

Learning Transferable Visual Models From Natural Language Supervision (CLIP).

In *ICML*, 2021.

https://arxiv.org/abs/2103.00020

References (cont.)

[3] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M.

Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL-E 2).

arXiv preprint, 2022.

https://arxiv.org/abs/2204.06125

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B.

High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion).

In *CVPR*, 2022.

https://arxiv.org/abs/2112.10752

References (cont.)

[5] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M.

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Imagen).

In *NeurIPS*, 2022.

https://arxiv.org/abs/2205.11487

[6] BLIP: Bootstrapped Language-Image Pretraining.

https://github.com/salesforce/BLIP

[7] OpenAI's DALL•E 2 Documentation.

https://platform.openai.com/docs/guides/images

References (cont.)

[8] Midjourney Documentation.

https://docs.midjourney.com/

[9] SDXL (Stable Diffusion XL) Documentation.

https://stability.ai/guides/sdxl

Credits

Dr. Prashant Aparajeya

Computer Vision Scientist — Director(AISimply Ltd)

p.aparajeya@aisimply.uk

This project benefited from external collaboration, and we acknowledge their contribution with gratitude.