



DATA 8005 Advanced Natural Language Processing

Multi-modal Language Models

Chengqi Duan, Tianshuo Yang, Mengzhao Chen

Fall 2024

Papers

- Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution
- GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models
- Chameleon: Mixed-Modal Early-Fusion Foundation Models



DATA 8005 Advanced Natural Language Processing

Qwen2-VL: Enhancing Vision-Language Model's Perception
of the World at Any Resolution

Chengqi Duan

Fall 2024

Discussions

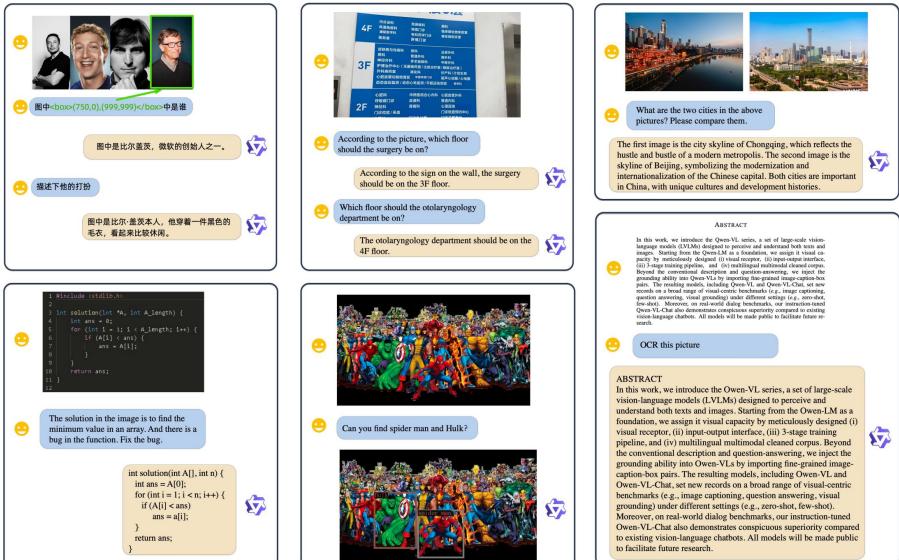
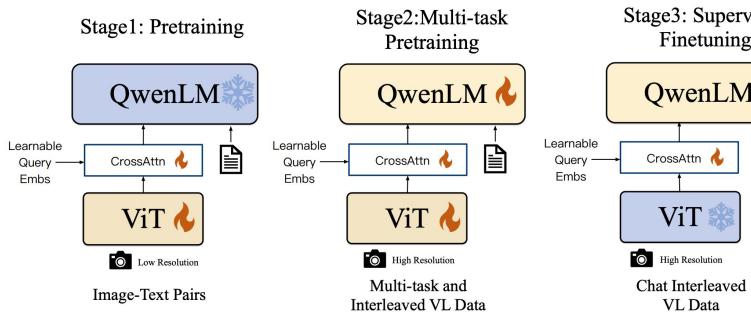
- Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at **Any Resolution**
- Normally, MLLMs take images at fixed resolution, how to train the model to take images at different resolutions?

Catalogue

- Recap on Qwen-VL
- Motivation
- A brief on Qwen2-VL
- Methodology
- Training
- Experiments
- Ablation study

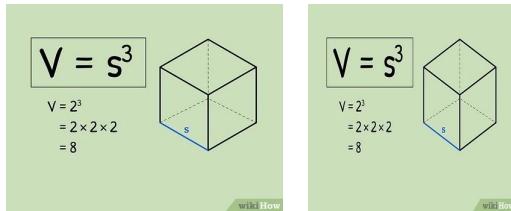
About Qwen-VL

- A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond
- Based on Qwen 7B LM
- 1.9B ViT + 0.08B VL Adapter



Motivation for Qwen2-VL

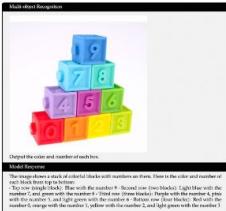
- Current LVLMs are constrained by fixed image input size
- Detailed info lost in high resolution image + information sensitive to scale



- The inadequate information extracted by encoders (CLIP etc.)
- The low performance to include video understanding in LVLMs

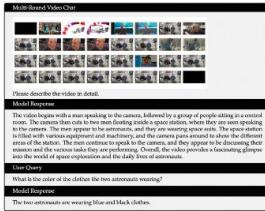
What Qwen2-VL proposes

SOTA understanding



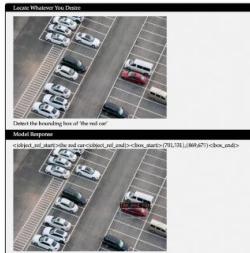
General Chat

Long video (20 min +)



Video Understanding

Multilingual



Grounding



Multilingual OCR

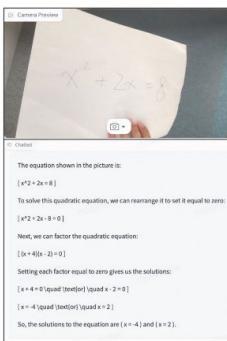


Long Document Understanding

Math & Code



Live Chat



Formula Recognition

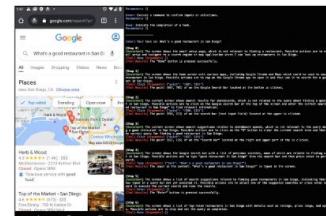


Function Calling



Beyond

UI Interaction



Capable for
Device operation

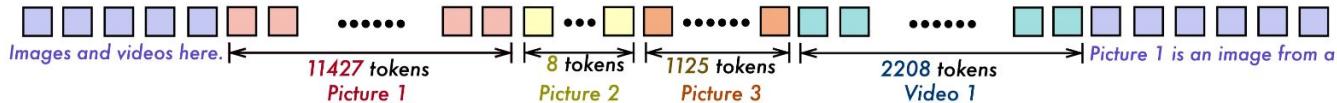
How are they able to do this?

and videos here.

Picture 1 is an image from a blog



QwenLM Decoder

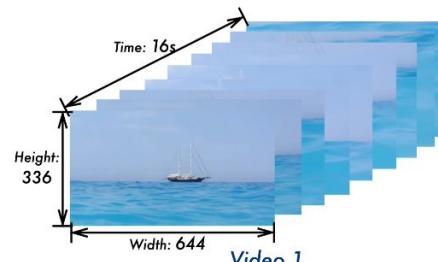
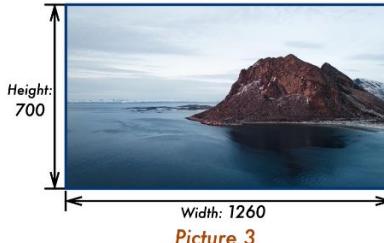


Vision Encoder

Native Resolution Input



Height: 28
Width: 224
Picture 2

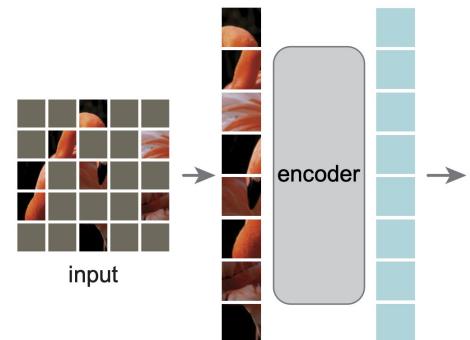


ViT encode images of different scales

- Replace absolute pos embed with 2D-RoPE for 2D info extraction

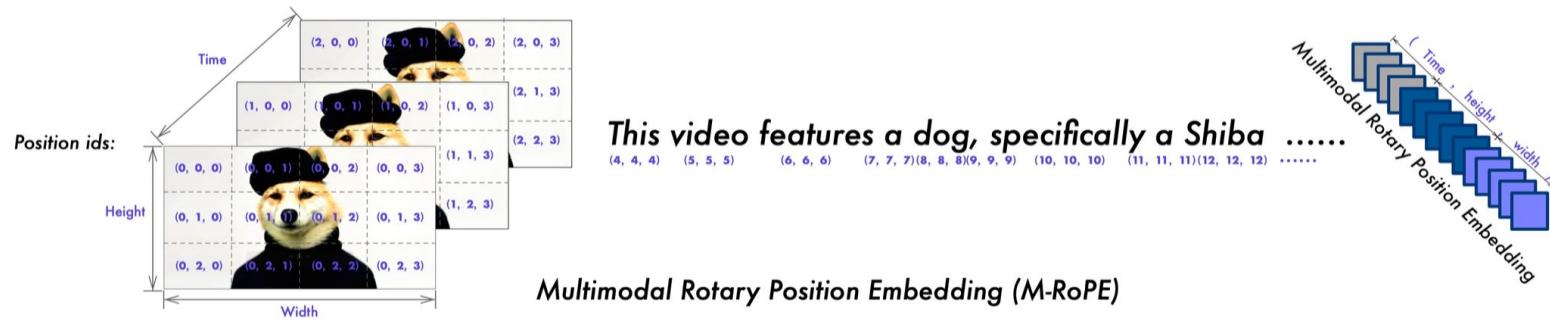
$$\vec{p}_t^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k \cdot t), & \text{if } i = 2k \\ \cos(\omega_k \cdot t), & \text{if } i = 2k + 1 \end{cases} \quad \omega_k = \frac{1}{10000^{2k/d}}$$

- Reduction of token numbers: compress adjacent 2x2 tokens with MLP
- ViT of patch 14



Multimodal Rotary Position Embedding for LLM

- Replace 1D-RoPE in LLM to M-RoPE (Time, Height and Width)



- Mixed training with image and video data (2fps + adjust resolution + 3D conv)

Training

- 1st Stage: Train ViT (DFN) with image-text pairs for semantic info extraction
- 2nd Stage: Finetune all parameters to train Qwen2 with 600B data
- Image text relationship, OCR, Image classification etc.
- 3rd Stage: Only finetune LLM on 800B data

Model Name	Vision Encoder	LLM	Model Description
Qwen2-VL-2B	675M	1.5B	The most efficient model, designed to run on-device. It delivers adequate performance for most scenarios with limited resources.
Qwen2-VL-7B	675M	7.6B	The performance-optimized model in terms of cost, significantly upgraded for text recognition and video understanding capabilities. It delivers significant performance across a broad range of visual tasks.
Qwen2-VL-72B	675M	72B	The most capable model, further improvements in visual reasoning, instruction-following, decision-making, and agent capabilities. It delivers optimal performance on most complex tasks.

Data format

The Dataset Format Example of ChatML

```
<|im_start|>user  
<|vision_start|>Picture1.jpg<|vision_end|><|vision_start|>Picture2.jpg<|vision_end|>What do the  
two pictures have in common?<|im_end|>  
<|im_start|>assistant  
Both pictures are of SpongeBob SquarePants. <|im_end|>  
<|im_start|>user  
What is happening in the video?<|vision_start|>video.mp4<|vision_end|><|im_end|>  
<|im_start|>assistant  
The protagonist in the video is frying an egg.<|im_end|>
```

Referring Grounding

```
<|vision_start|>Picture1.jpg<|vision_end|>  
<|object_ref_start|>the eyes on a giraffe<|object_ref_end|><|box_start|>(176,106),(232,160)  
<|box_end|>
```

Visual Agent

```
<|im_start|>system  
You are a helpful assistant.  
# Actions  
## You have the following actions.  
### Tap  
Tap: A gentle tap that commands, chooses, or navigates through a smartphone's user interface.  
Parameters: [{"name": "point", "description": "The specific spot of interest on the monitor, denoted by the coordinates (x, y) where x and y range from 0 to 1000.", "required": True}]  
### Home  
Home: Go to phone's home screen. Parameters: []  
### Other Actions ...  
## Continuously take action until the task is completed.  
*FUNCTION*: The action to take, should be one of {Actions}.  
*ARGS*: The input of the action.  
*RESULT*: Action results.  
*RETURN*: Reply based on action results. <|im_end|>  
<|im_start|>user  
Find a pizza restaurant nearby in Map. <|vision_start|>Screenshot_1.jpg<|vision_end|> <|im_end|>  
<|im_start|>assistant  
Several e-mails are displaying on the phone screen. To open Map, I need go back to the home screen  
and find the corresponding app icon.  
*FUNCTION*: Home *ARGS*: {}  
*RESULT*: <|vision_start|>Screenshot_2.jpg<|vision_end|>  
*RETURN*: I return to the home screen. Next, I need to find the icon of Map and tap on it.  
*FUNCTION*: Tap *ARGS*: {"point": (348,291)}  
*RESULT*: <|vision_start|>Screenshot_3.jpg<|vision_end|>  
*RETURN*: [Thinking for the next action.]  
[Other subsequent actions.] ....  
I have found the pizza restaurant nearby in Map. <|im_end|>
```

Experiments

Table 2: Performance Comparison of Qwen2-VL Models and State-of-the-art.

Benchmark	Previous SoTA	Claude-3.5 Sonnet	GPT-4o	Qwen2-VL-72B	Qwen2-VL-7B	Qwen2-VL-2B
MMMU _{val} (Yue et al., 2023)	66.1 (X.AI, 2024b)	68.3	69.1	64.5	54.1	41.1
DocVQA _{test} (Mathew et al., 2021)	94.1 (Chen et al., 2024c)	95.2	92.8	96.5	94.5	90.1
InfoVQA _{test} (Mathew et al., 2021)	82.0 (Chen et al., 2024c)	-	-	84.5	76.5	65.5
AI2D (Kembhavi et al., 2016)	87.6 (Chen et al., 2024c)	80.2(94.7)	84.6(94.2)	88.1	83.0	74.7
ChartQA _{test} (Masry et al., 2022)	88.4 (Chen et al., 2024c)	90.8	85.7	88.3	83.0	73.5
TextVQA _{val} (Singh et al., 2019)	84.4 (Chen et al., 2024c)	-	-	85.5	84.3	79.7
OCRBench (Liu et al., 2023e)	852 (Yao et al., 2024)	788	736	877	866	809
MTVQA (Tang et al., 2024)	23.2 (Team et al., 2023)	25.7	27.8	30.9	25.6	18.1
VCR _{en} easy (Zhang et al., 2024c)	84.7 (Chen et al., 2024c)	63.9	91.6	91.9	89.7	81.5
VCR _{zh} easy (Zhang et al., 2024c)	22.1 (Chen et al., 2024c)	1.0	14.9	65.4	59.9	46.2
RealWorldQA (X.AI, 2024a)	72.2 (Chen et al., 2024c)	60.1	75.4	77.8	70.1	62.9
MME _{sum} (Fu et al., 2023)	2414.7 (Chen et al., 2024c)	1920.0	2328.7	2482.7	2326.8	1872.0
MMBench-EN _{test} (Liu et al., 2023d)	86.5 (Chen et al., 2024c)	79.7	83.4	86.5	83.0	74.9
MMBench-CN _{test} (Liu et al., 2023d)	86.3 (Chen et al., 2024c)	80.7	82.1	86.6	80.5	73.5
MMBench-VI.1 _{test} (Liu et al., 2023d)	85.5 (Chen et al., 2024c)	78.5	82.2	85.9	80.7	72.2
MMT-Bench _{test} (Ying et al., 2024)	63.4 (Chen et al., 2024b)	-	65.5	71.7	63.7	54.5
MMStar (Chen et al., 2024a)	67.1 (Chen et al., 2024c)	62.2	63.9	68.3	60.7	48.0
MMVet _{GPT-4-Turbo} (Yu et al., 2024)	67.5 (OpenAI, 2023)	66.0	69.1	74.0	62.0	49.5
HallBench _{avg} (Guan et al., 2023)	55.2 (Chen et al., 2024c)	49.9	55.0	58.1	50.6	41.7
MathVista _{testmini} (Lu et al., 2024a)	69.0 (X.AI, 2024b)	67.7	63.8	70.5	58.2	43.0
MathVision (Wang et al., 2024)	30.3 (OpenAI, 2023)	-	30.4	25.9	16.3	12.4
MMMU-Pro (Yue et al., 2024)	46.9 (Team et al., 2023)	51.5	51.9	46.2	43.5	37.6

Table 3: Performance of Qwen2-VL and GPT-4o on internal multilingual OCR benchmarks.

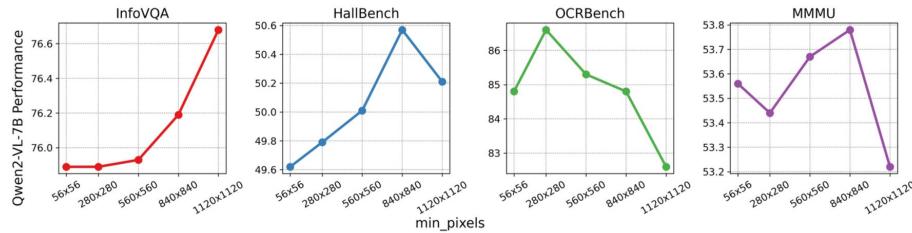
Language	Korean	Japanese	French	German	Italian	Russian	Vietnamese	Arabic
GPT-4o	87.8	88.3	89.7	88.3	74.1	96.8	72.0	75.9
Qwen2-VL-72B	94.5	93.4	94.1	91.5	89.8	97.2	73.0	70.7

- Function calling
- UI operations
- Robotic Control
- Card games
- Vision-Language Navigation

Ablation study

- Dynamic Resolution vs Fixed resolution of different image size

Strategy	Average Image Tokens	InfoVQA _{val}	RealWorldQA	OCR Bench	MMMU
Fixed Image Tokens	64	28.85	56.47	572	53.33
	576	65.72	65.88	828	52.78
	1600	74.99	69.54	824	52.89
	3136	77.27	70.59	786	53.44
Dynamic Image Tokens	1924	75.89	70.07	866	53.44

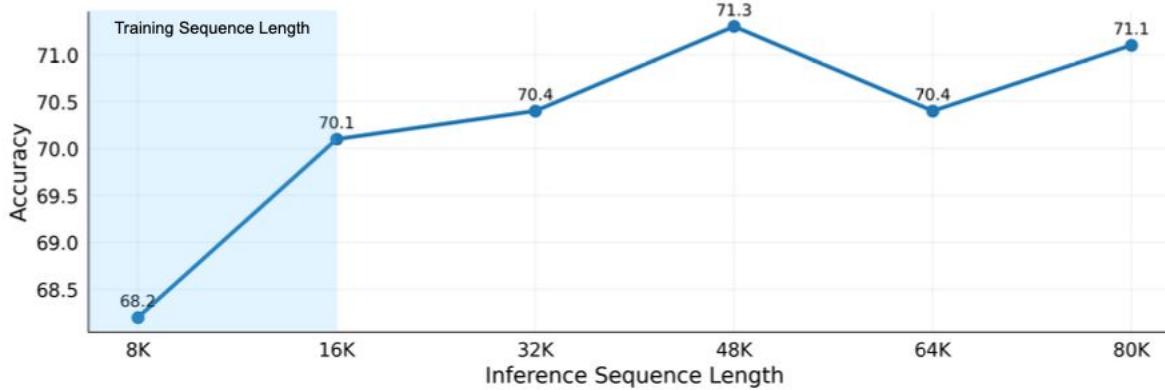


- No more one size fits all !!!

Ablation study

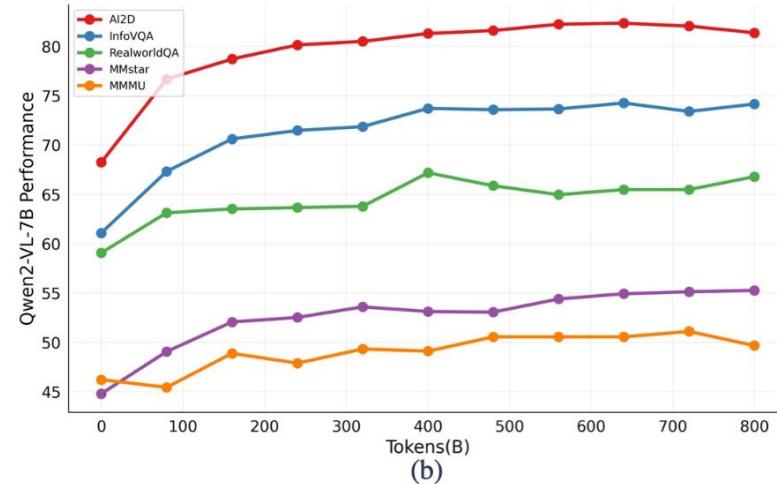
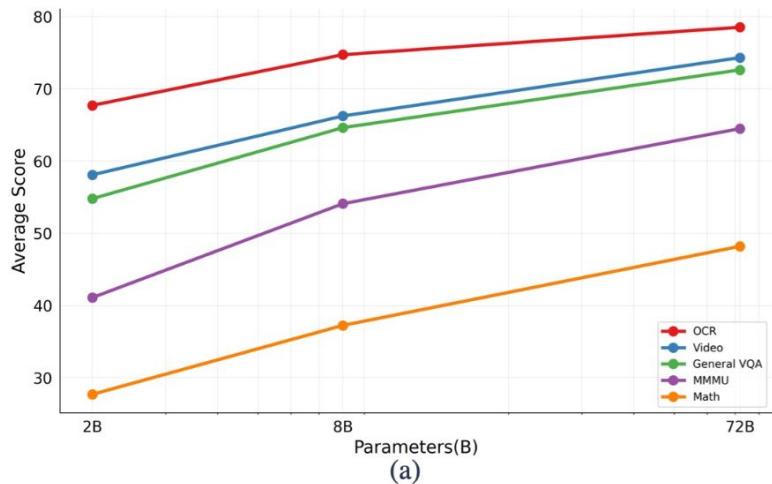
- M-RoPE

	Image Benchmarks							Video Benchmarks			
	MathVista	MMB	MMStar	RWQ	DocVQA	ChartQA	InfoVQA	TextVQA	PerceptionTest	NextQA	STAR
1D-RoPE	39.2	58.6	36.7	54.5	82.5	68.0	50.8	71.3	46.6	43.9	55.5
M-RoPE	43.4	60.6	36.7	53.7	82.8	68.4	50.3	71.8	47.4	46.0	57.9



Ablation study

- Scaling Law of Parameters and Num of Training Tokens





DATA 8005 Advanced Natural Language Processing

GLIDE: Towards Photorealistic Image Generation and Editing with
Text-Guided Diffusion Models

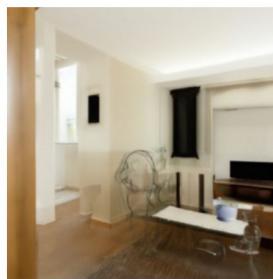
Tianshuo Yang
Fall 2024



DATA 8005 Advanced Natural Language Processing

GLIDE: Towards Photorealistic **Image Generation** and Editing
with **Text-Guided** Diffusion Models

Tianshuo Yang
Fall 2024



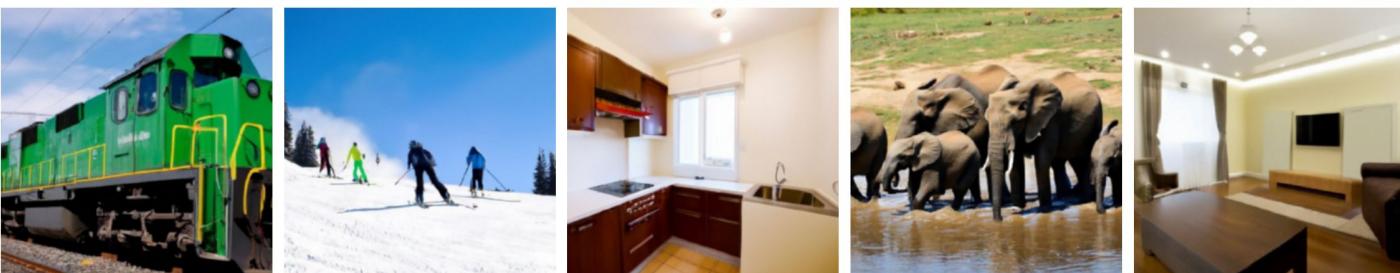
DALL-E



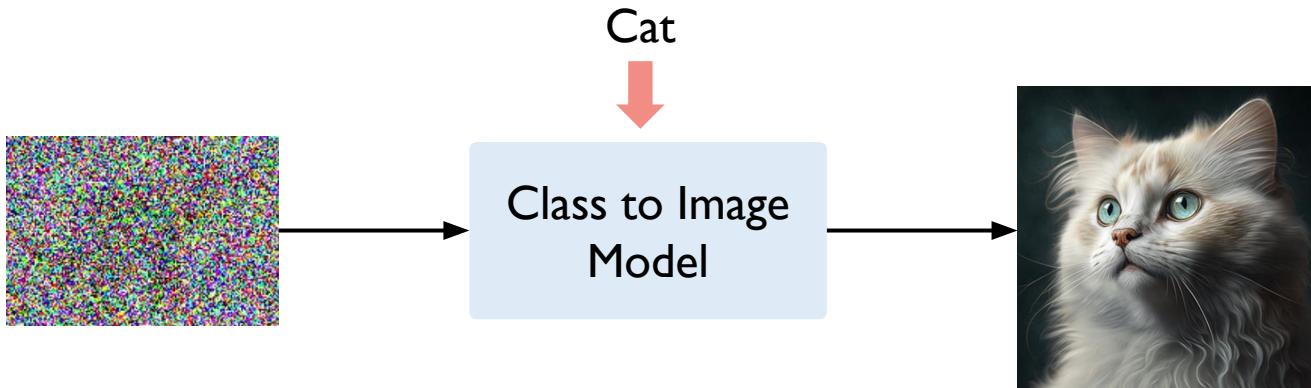
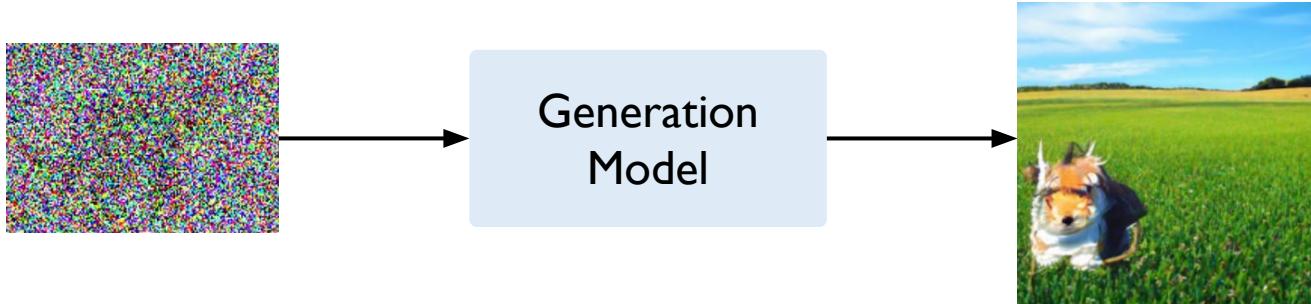
GLIDE
(CLIP
guidance)



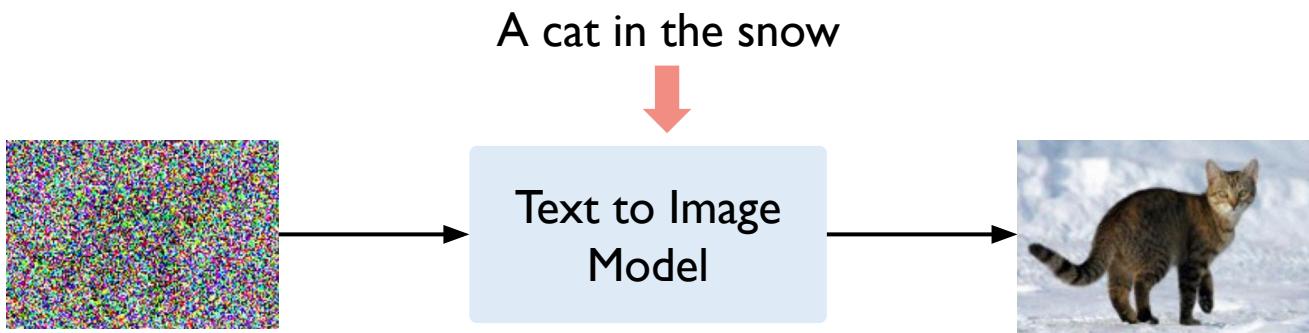
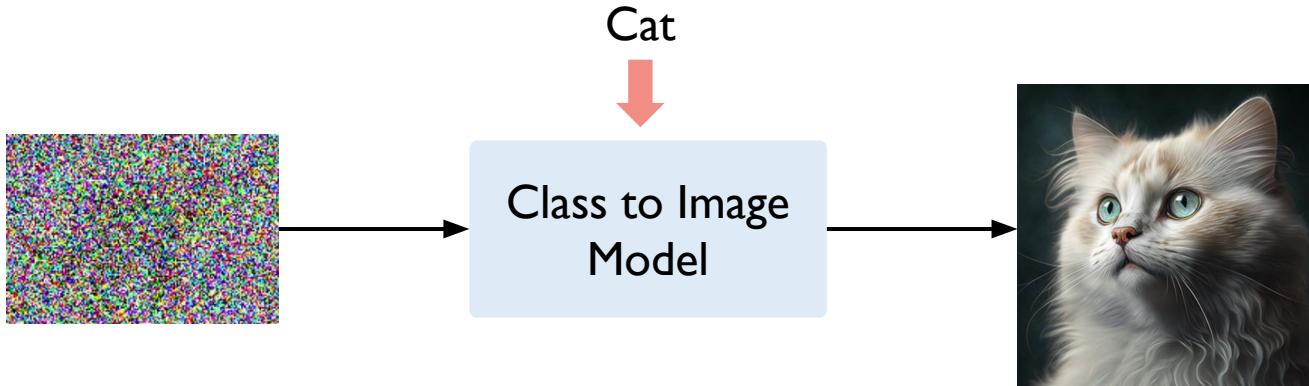
GLIDE
(Classifier-
free
guidance)



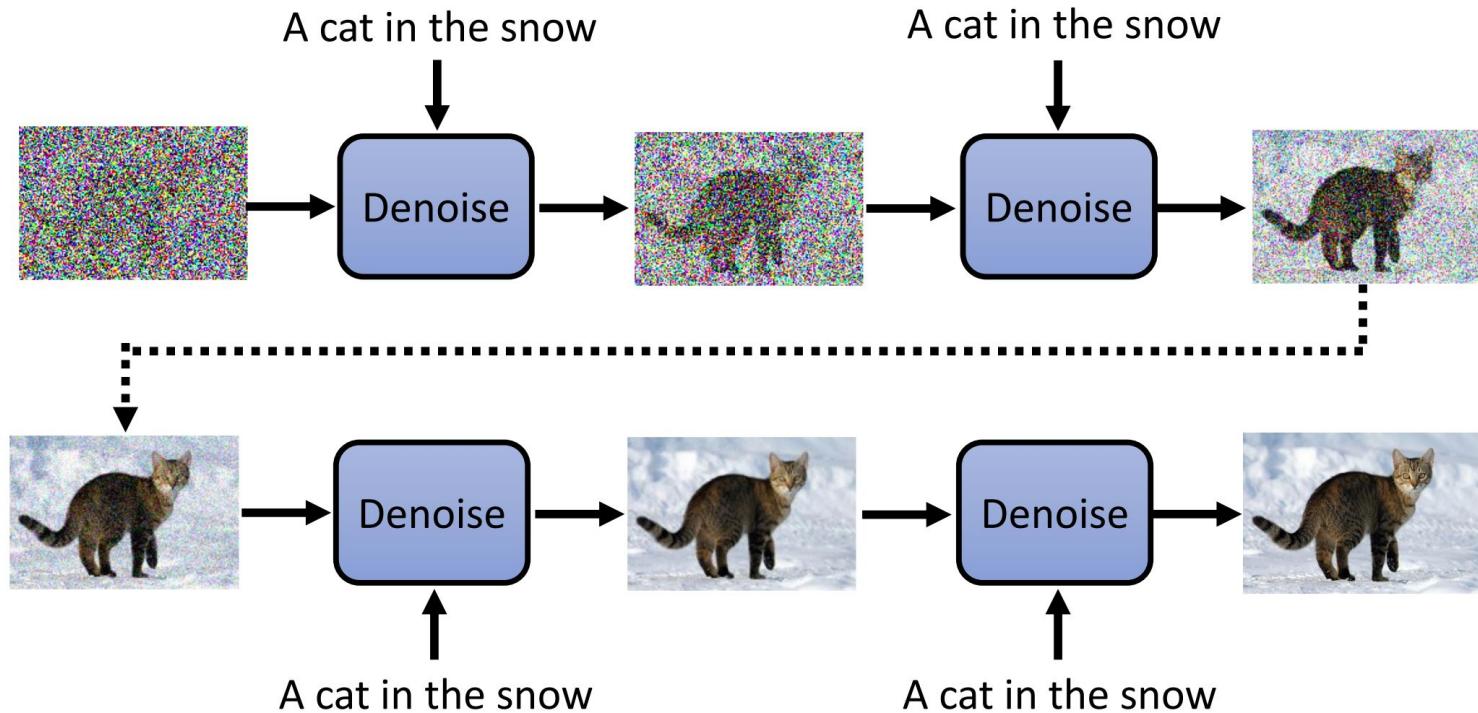
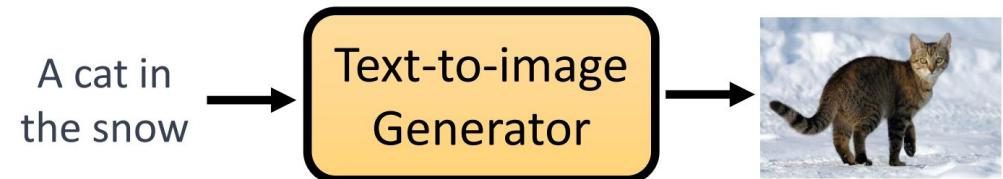
Guidance



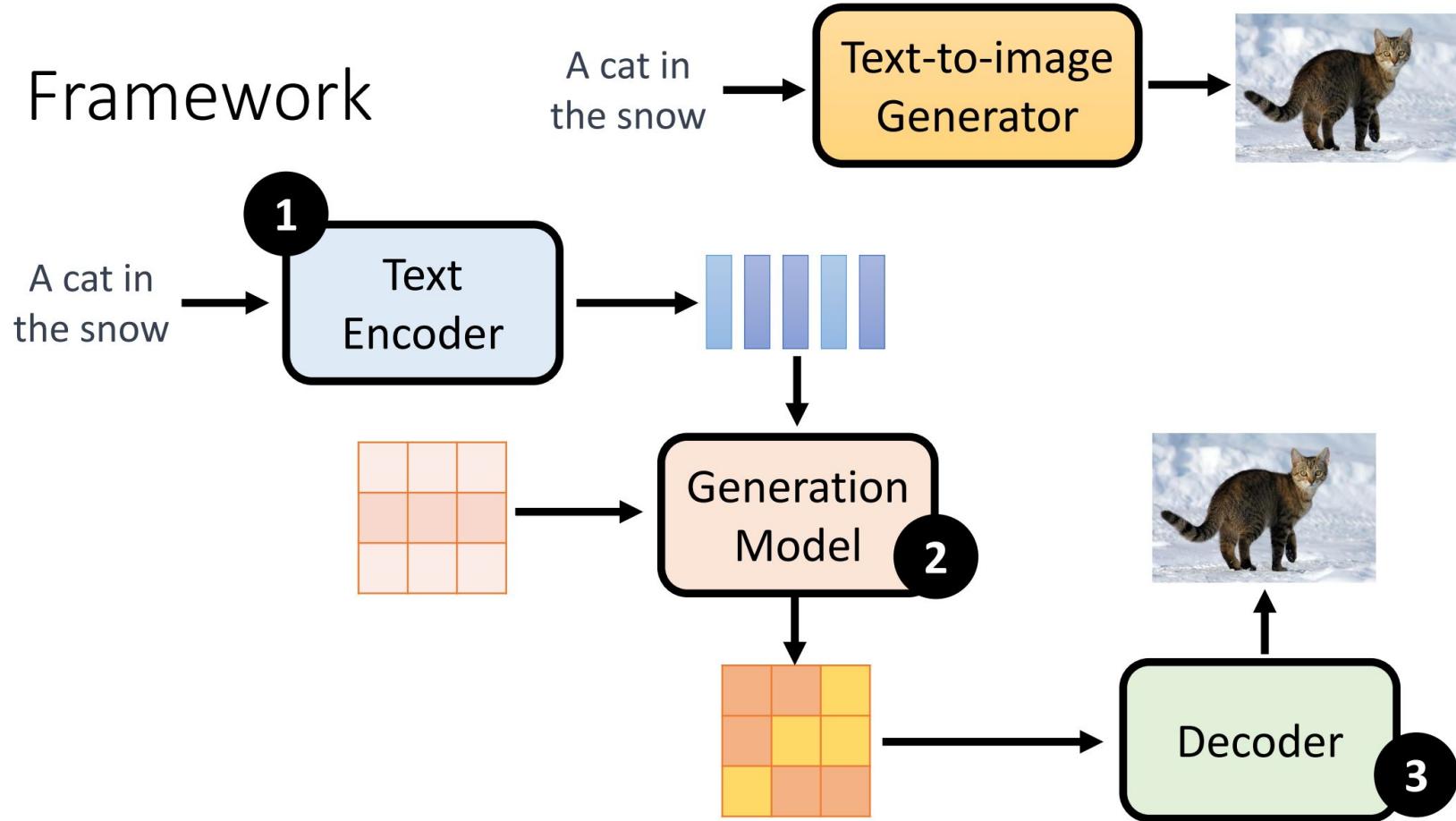
Guidance



Text-to-Image



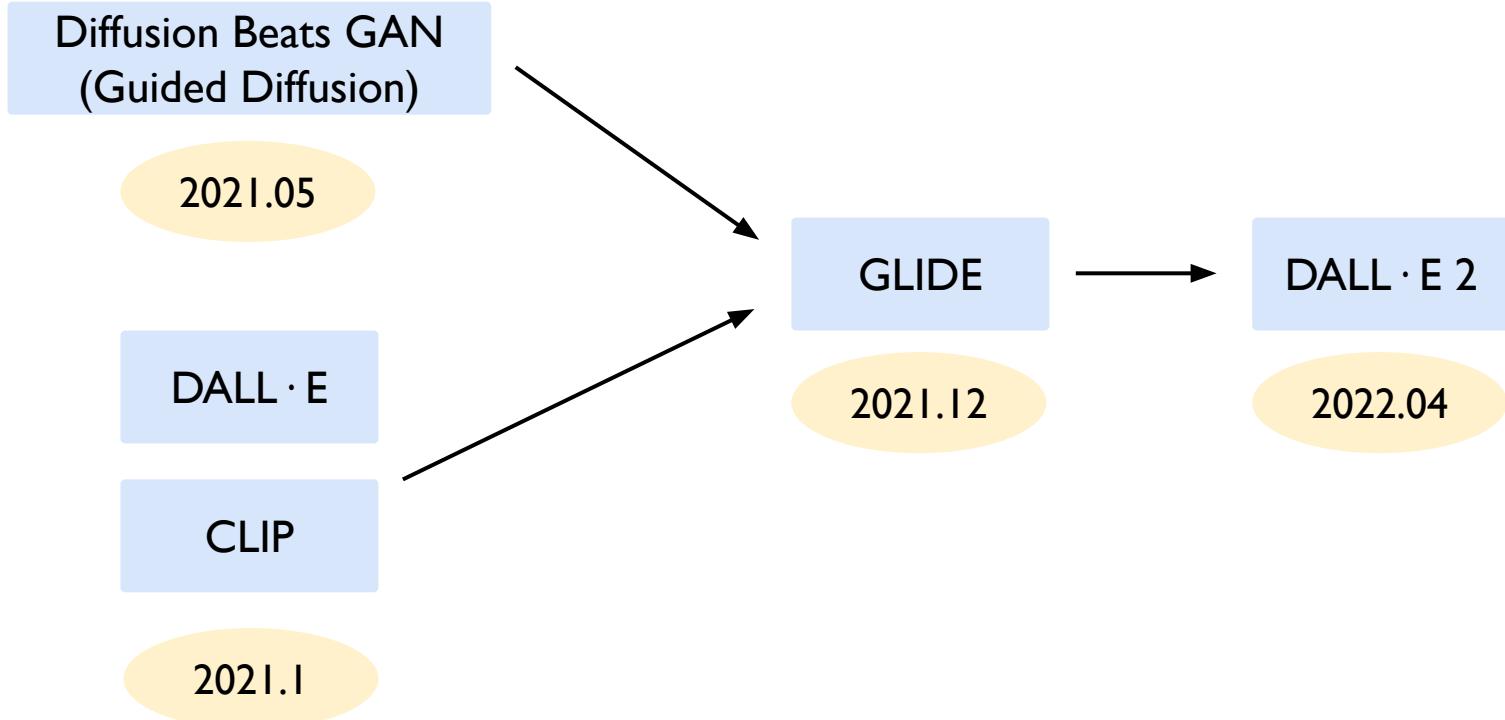
Framework



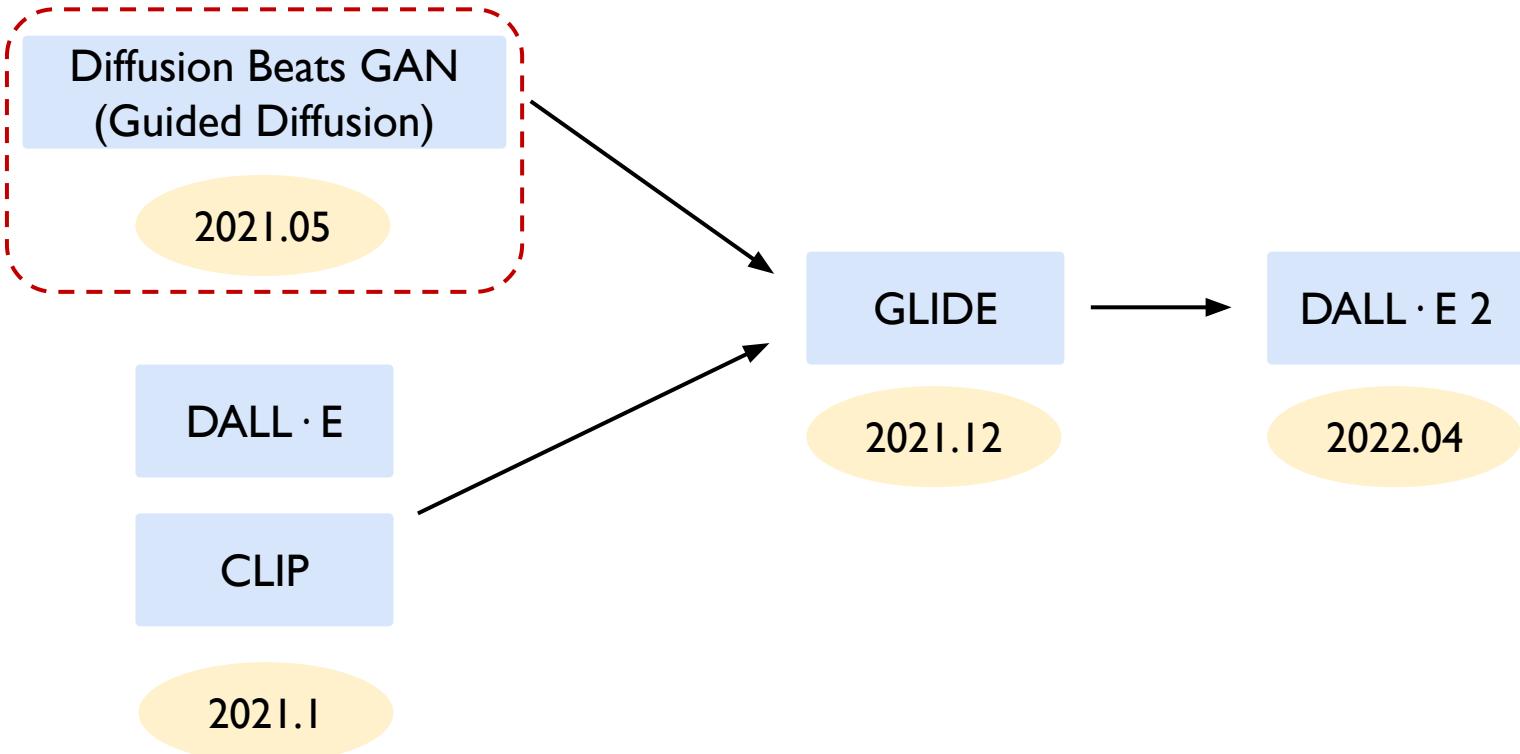
Target & Metrics

High Generation Quality (Target)	Metrics
<ul style="list-style-type: none">• Diversity• Fidelity (Photorealistic)<ul style="list-style-type: none">• Style and lighting• Shadows and reflections• Caption similarity	<ul style="list-style-type: none">• FID• Fréchet Inception Distance• IS (Inception Score)• Precision• Recall

Road Map (OpenAI)



Road Map (OpenAI)

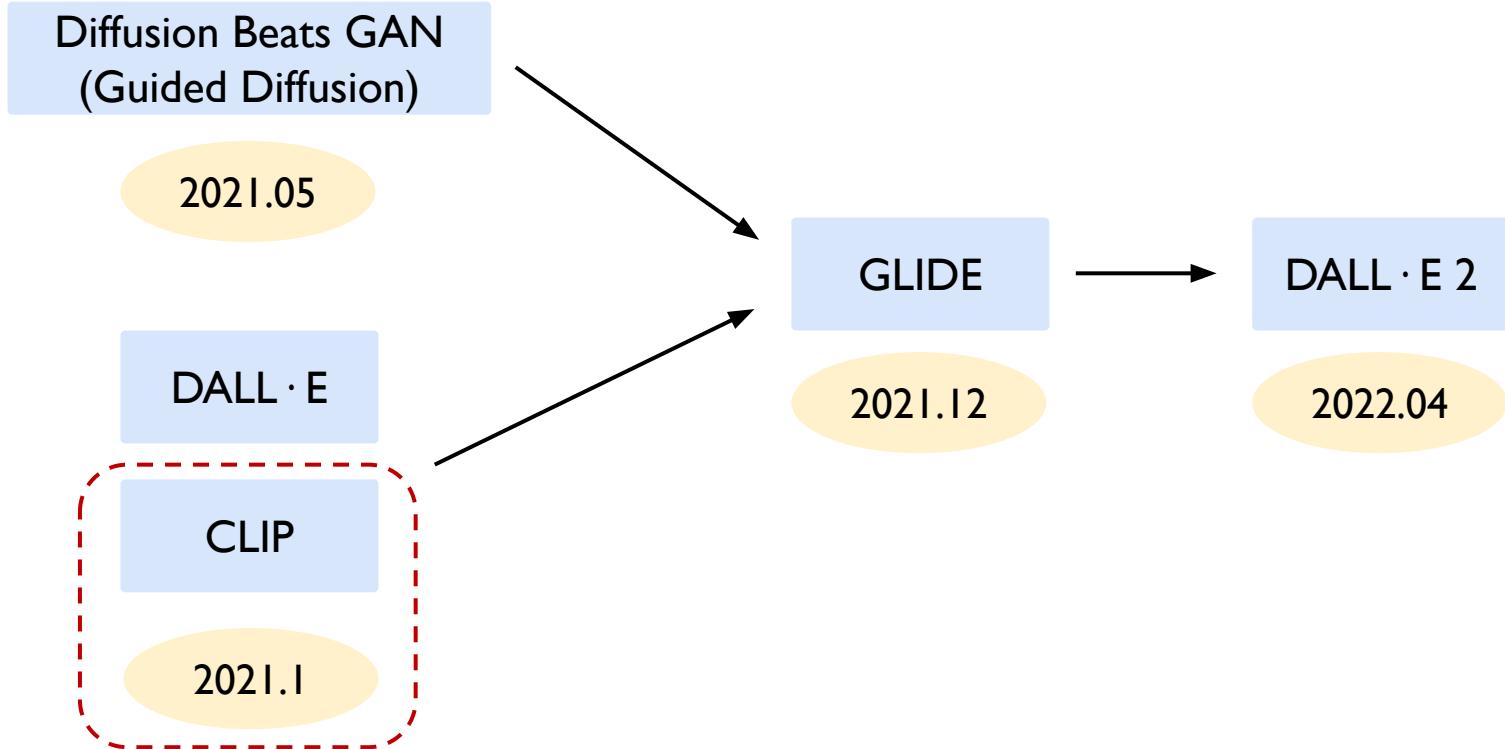


Classifier Guidance

Core Contribution:

- **Require a separate classifier model to be trained**
- Gradients from the classifier are used to **guide** the sample towards the label
 - The gradient of the log probability of a target class predicted by the classifier
- Increasing guidance scale improves sample quality at the cost of diversity

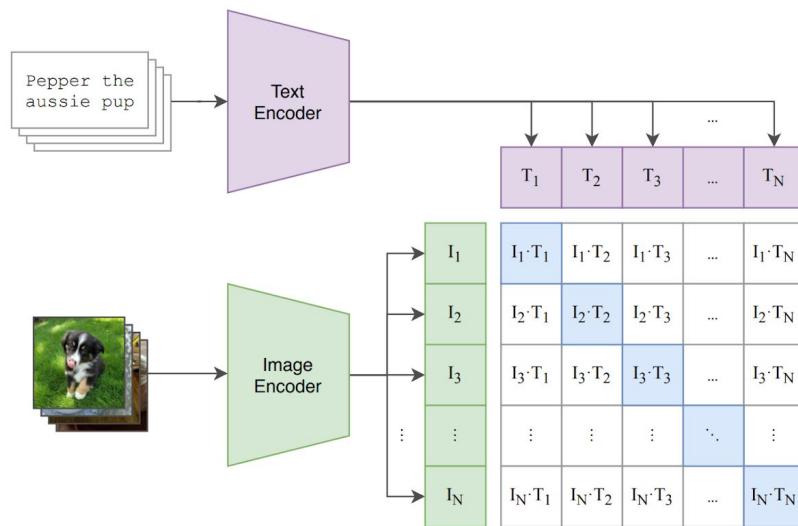
Road Map (OpenAI)



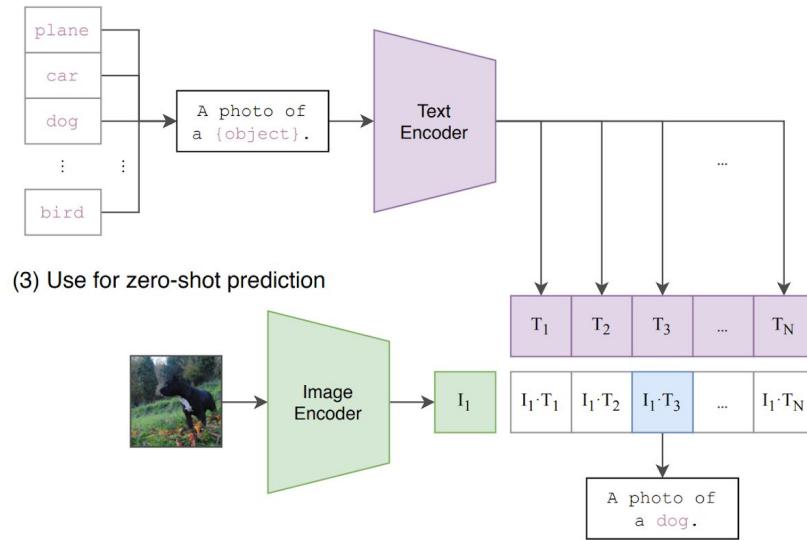
CLIP: Connecting text and images

- Build contrastive learning on 400M text-image pairs
- Unify text and image semantic spaces using cross entropy loss

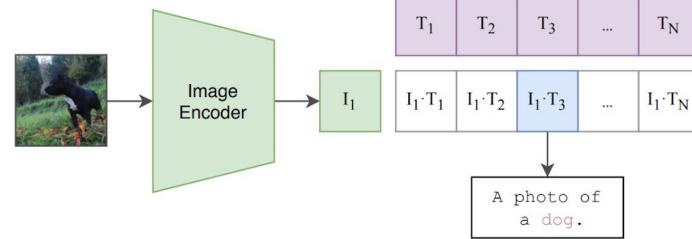
(1) Contrastive pre-training



(2) Create dataset classifier from label text



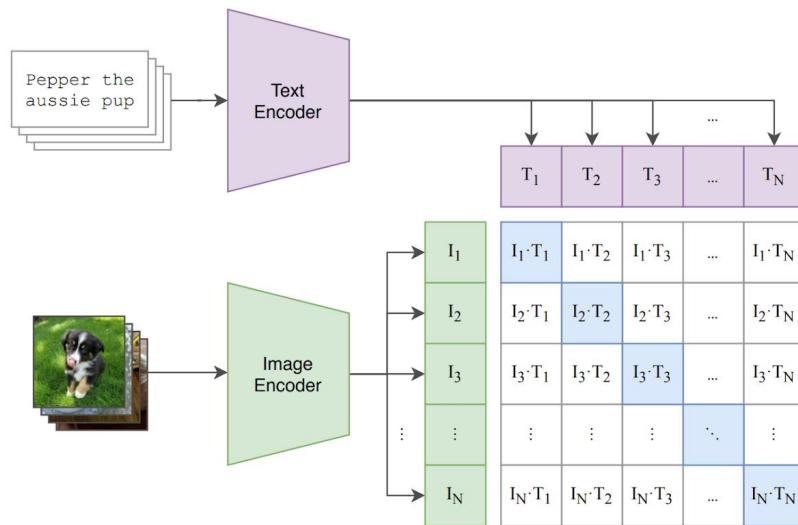
(3) Use for zero-shot prediction



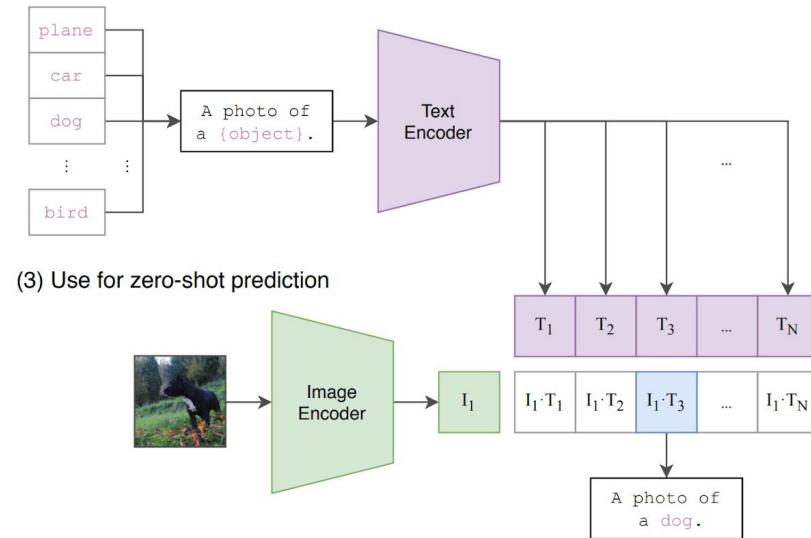
CLIP Guidance

- Replace the classifier with a CLIP model in classifier guidance
- Train a new CLIP on noised images to obtain the correct gradient

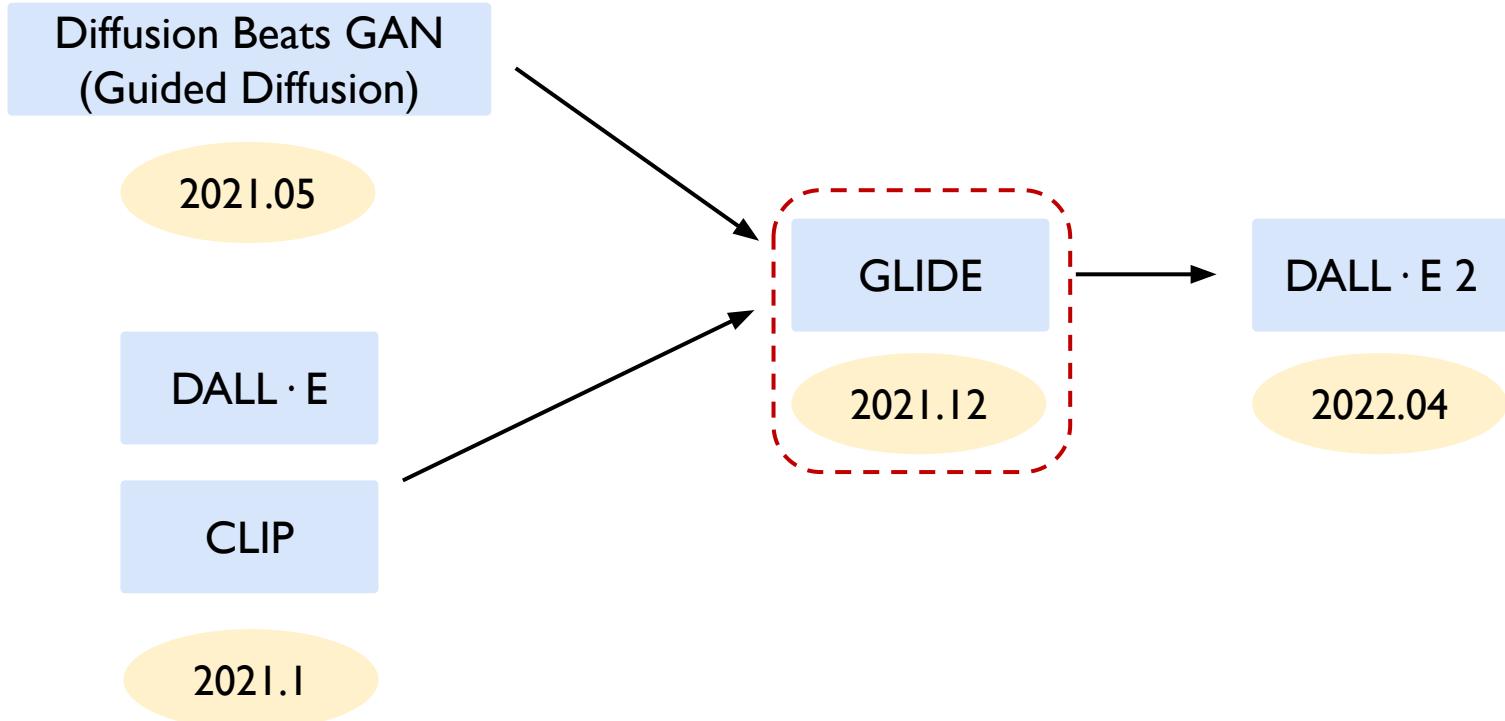
(1) Contrastive pre-training



(2) Create dataset classifier from label text



Road Map (OpenAI)



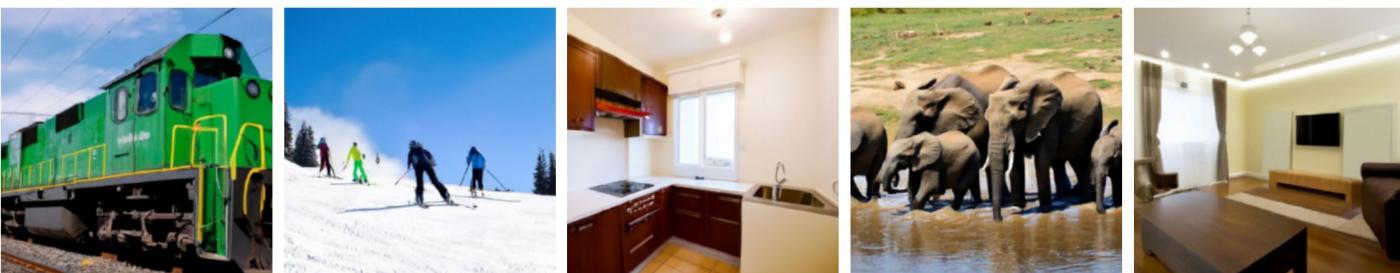
DALL-E



GLIDE
(CLIP
guidance)



GLIDE
(Classifier-
free
guidance)



Take Home Message

Classifier-free guidance is better than CLIP guidance
for T2I generation

Classifier-free Guidance (Label)

- Label y in a class-conditional diffusion model is replaced with a null label \emptyset with a fixed probability during training
- During sampling, the output of the model is extrapolated between the
 - class-conditional diffusion model
 - unconditional diffusion model

Classifier-free Guidance (Text)

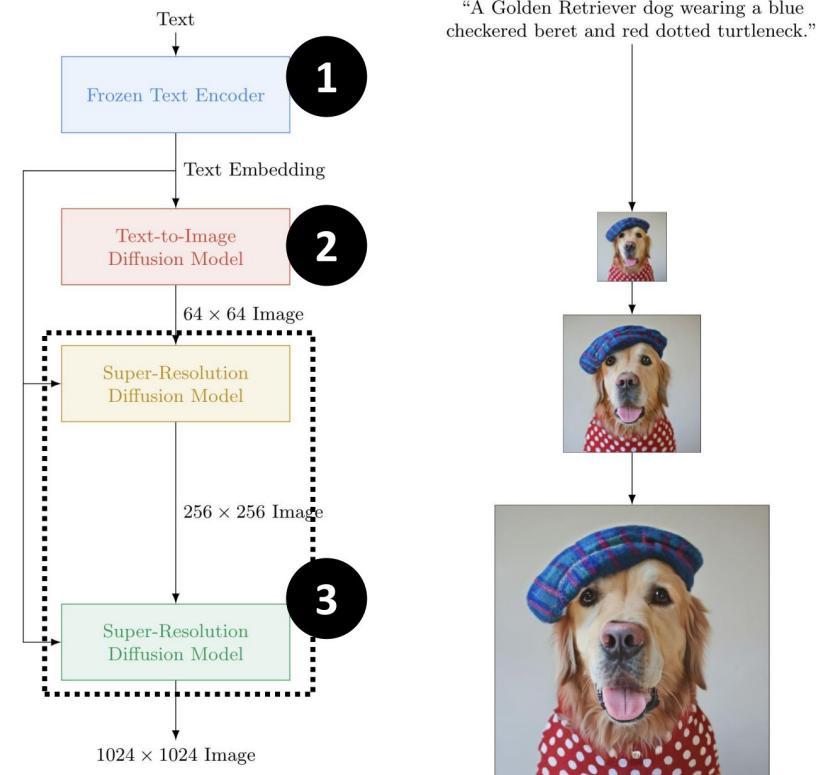
- Replace text captions with an empty sequence (which we also refer to as \emptyset) during training
 - It simplifies guidance when conditioning on information that is difficult to predict with a classifier (such as text)
 - Allows a single model to leverage its own knowledge during guidance, rather than relying on the knowledge of a separate (and sometimes smaller) classification model

Model

Cascaded Diffusion Model

- 3.5B text-conditional diffusion model at 64 × 64 resolution
- 1.5B text-conditional upsampling diffusion model to increase the resolution to 256 × 256

(Idea from Google CDM)



Training

Training (guided diffusion)

- Replace the class embedding with text token embedding

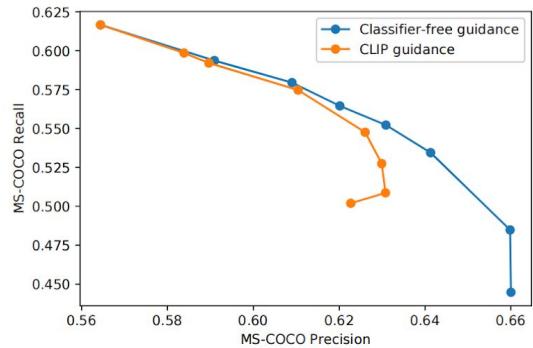
Fine-tuning (classifier-free guidance)

- 20% of text token sequences are replaced with the empty sequence

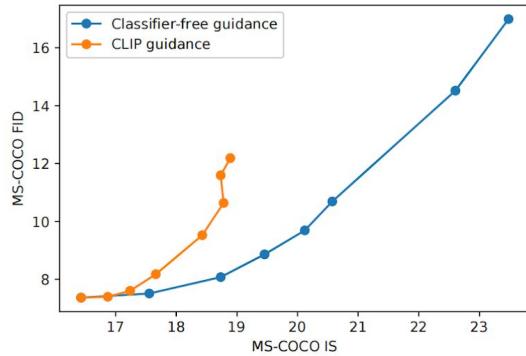
CLIP guidance

- Train a new CLIP on noised images to obtain the correct gradient

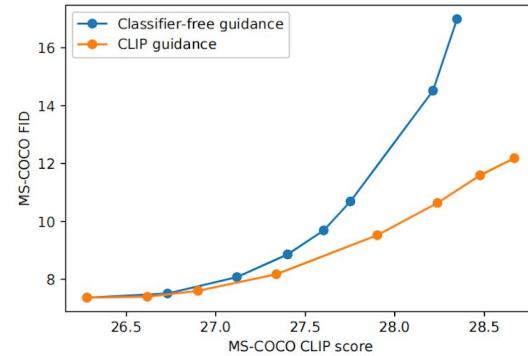
Results



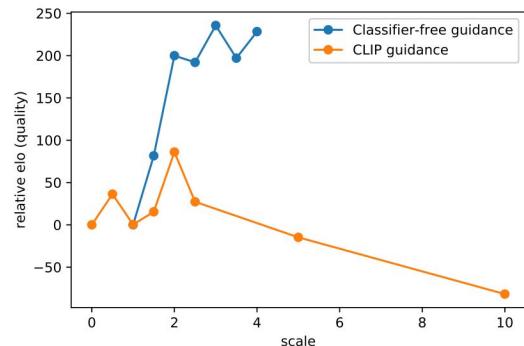
(a) Precision/Recall



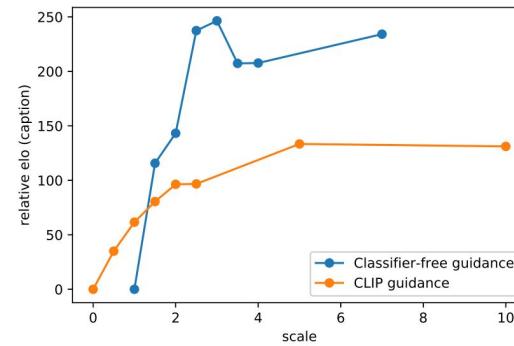
(b) IS/FID



(c) CLIP score/FID



(a) Photorealism



(b) Caption Similarity

Target & Metrics

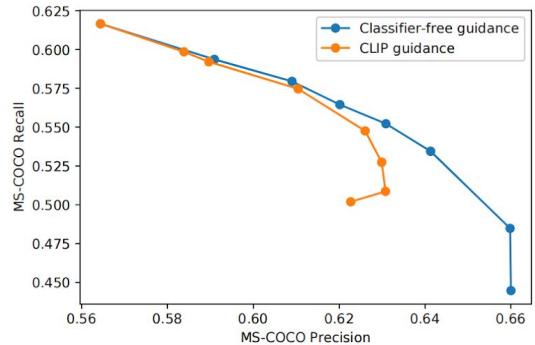
High Generation Quality (Target)	Metrics
<ul style="list-style-type: none">• Diversity• Fidelity (Photorealistic)<ul style="list-style-type: none">• Style and lighting• Shadows and reflections• Caption similarity	<ul style="list-style-type: none">• FID• Fréchet Inception Distance• IS (Inception Score)• Precision• Recall

Target & Metrics

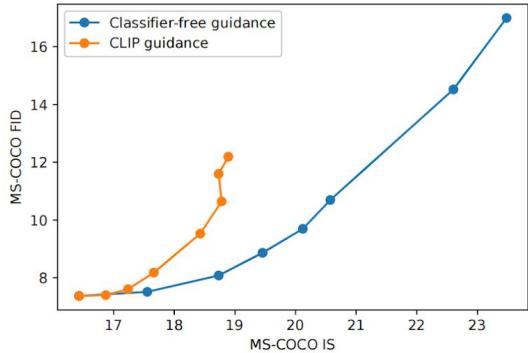
Metrics

- FID Both diversity and fidelity
 - Fréchet Inception Distance
- IS (Inception Score) Fidelity
- Precision Fidelity
- Recall Diversity

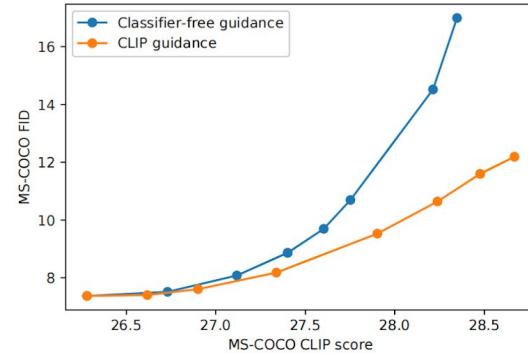
Results



(a) Precision/Recall

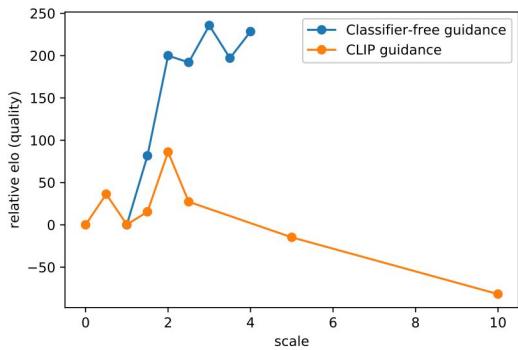


(b) IS/FID

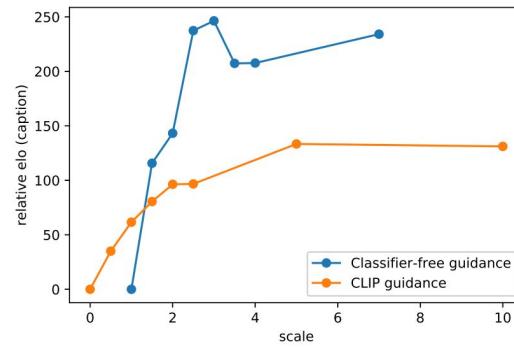


(c) CLIP score/FID

Results



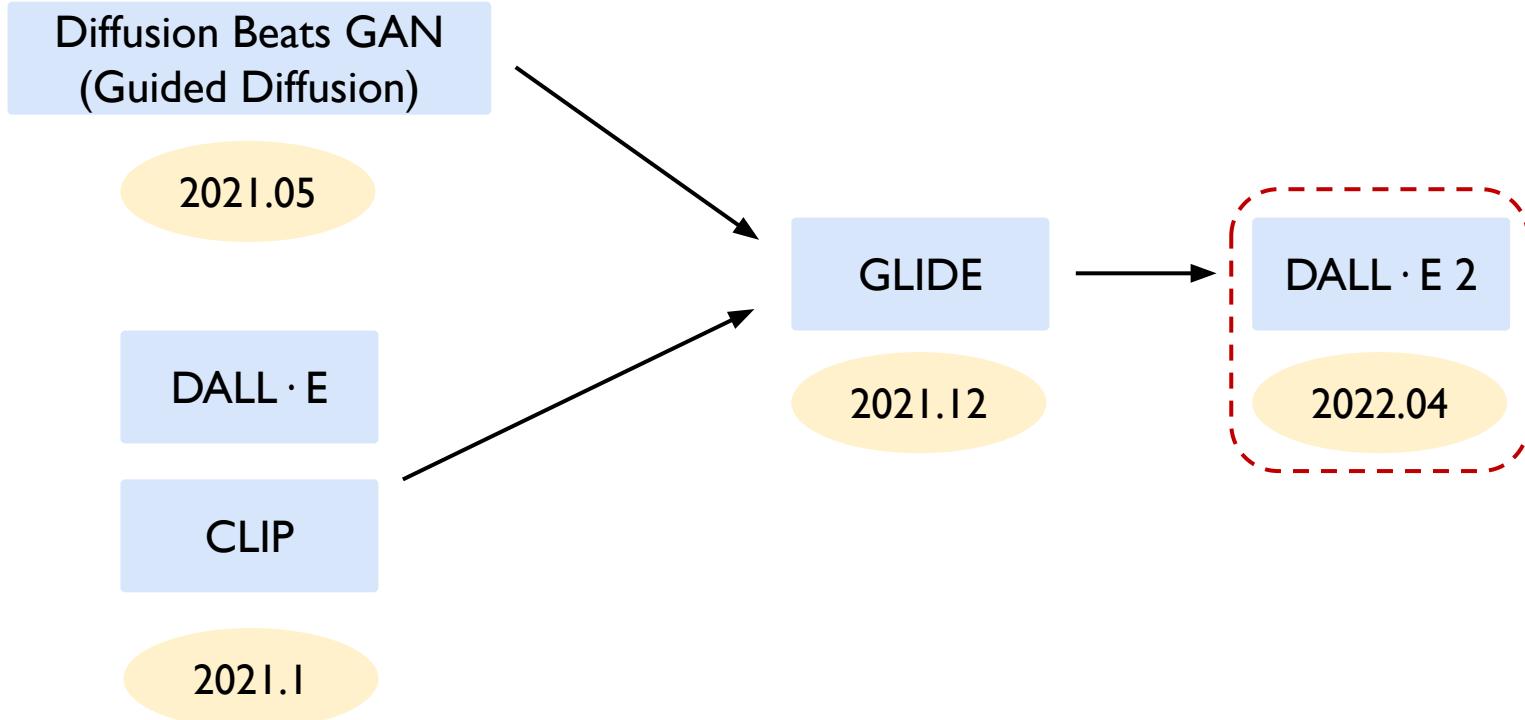
(a) Photorealism



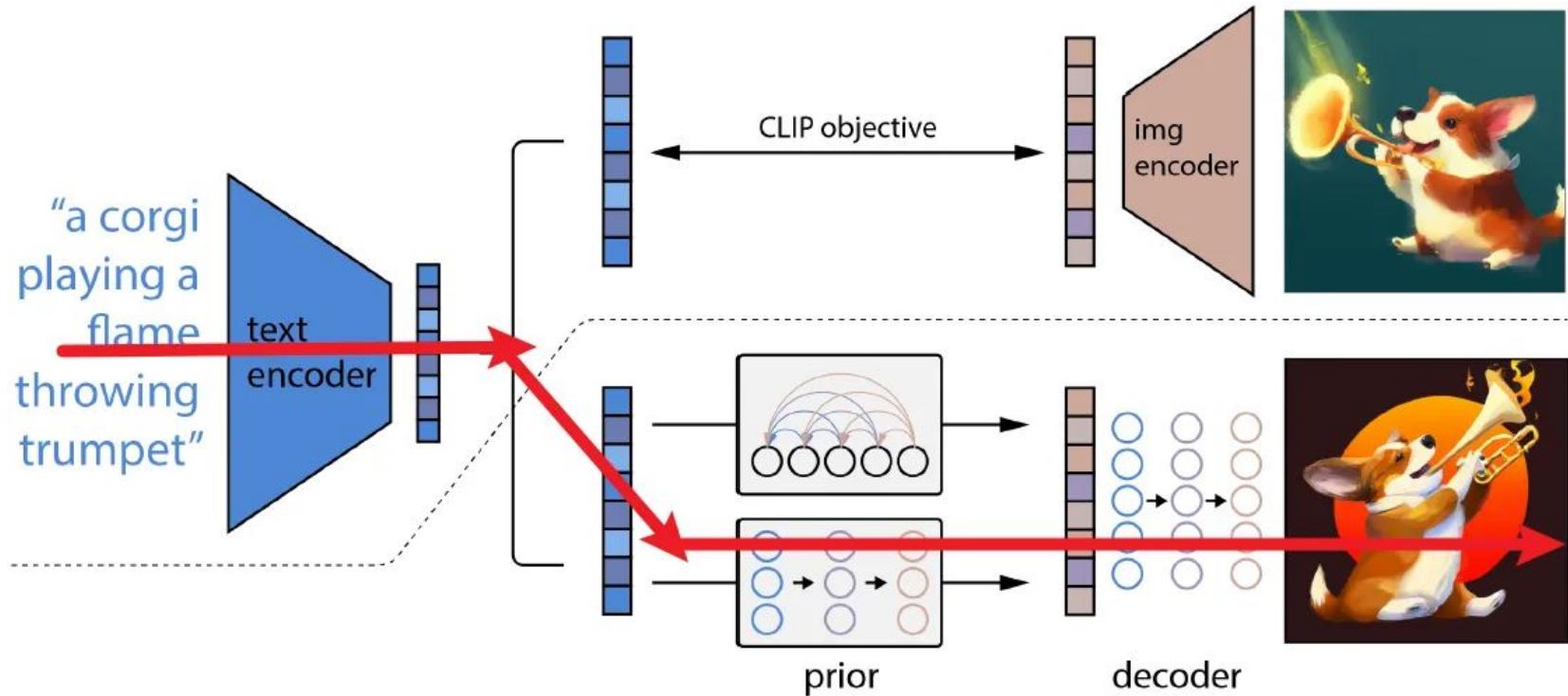
(b) Caption Similarity

Guidance	Photorealism	Caption
Unguided	-88.6	-106.2
CLIP guidance	-73.2	29.3
Classifier-free guidance	82.7	110.9

Road Map (OpenAI)



DALL·E - 2



Discussions

- Question 1
 - Is it possible to avoid retraining a CLIP model when using CLIP guidance?

- Question 2
 - What are the advantages and disadvantages of autoregressive generation and diffusion generation?



DATA 8005 Advanced Natural Language Processing

Chameleon: Mixed-Modal Early-Fusion Foundation Models



Mengzhao Chen

Fall 2024

Discussions

- The first sentence in **GPT-4o system card**:

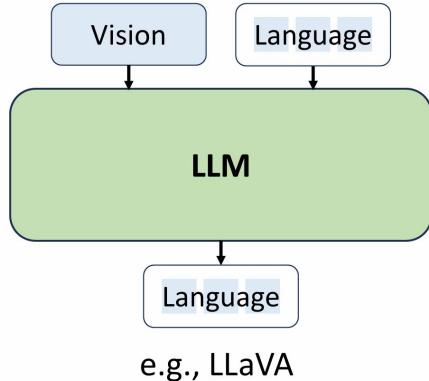
GPT-4o^[1] is an autoregressive omni model, which accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It's trained

- GPT-4o uses naive autoregressive model to tackle multi modalities
 - How to train it ?

Outlines

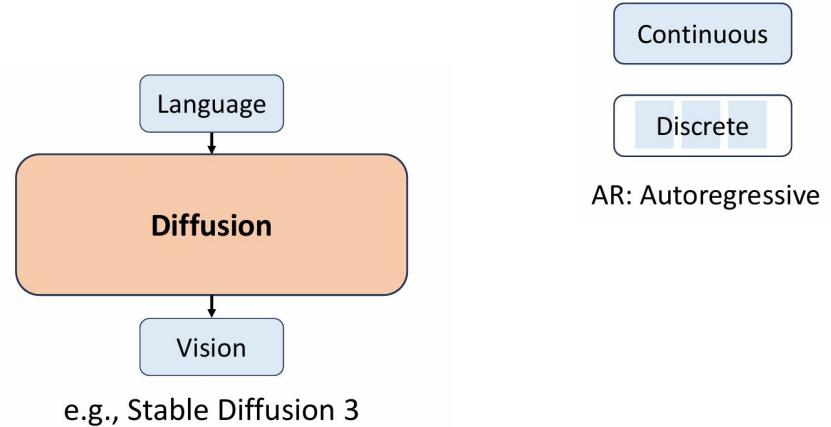
- Background
- Motivation
- Method
- Results
- Summary

Background



Understanding Only with AR

- Be built Efficiently

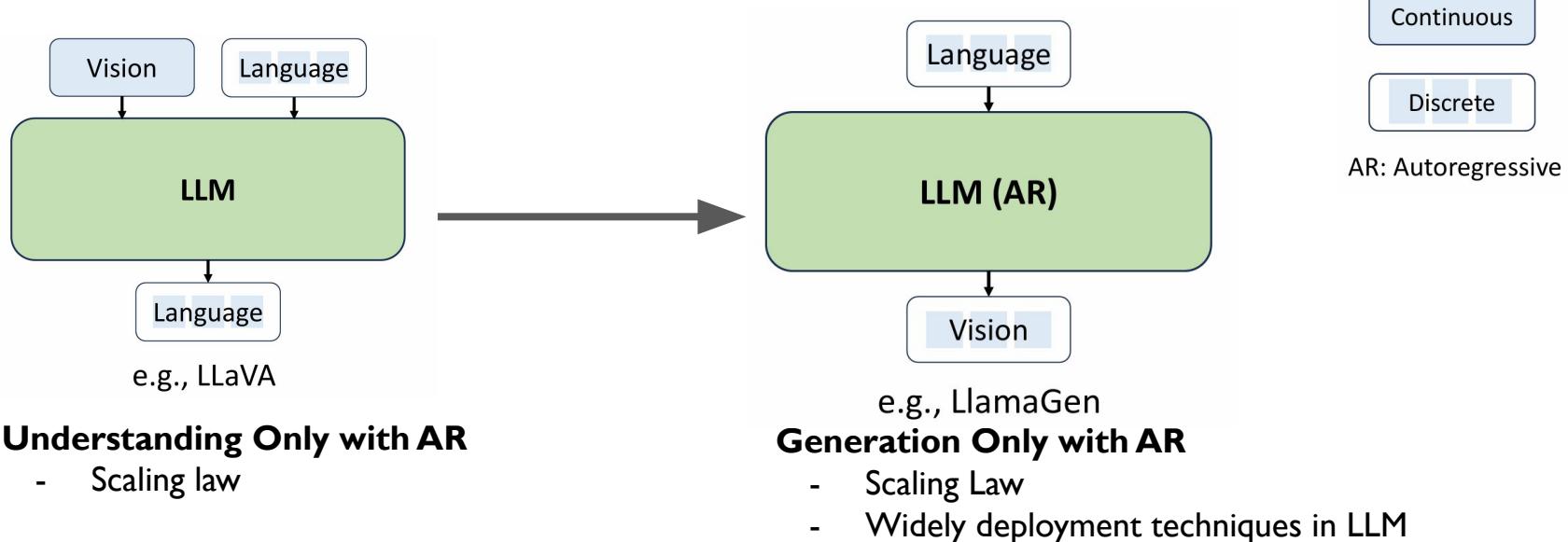


Generation Only with Diffusion

- Less information loss

Question 1: Can we use the same architecture for image understanding and generation?

Background



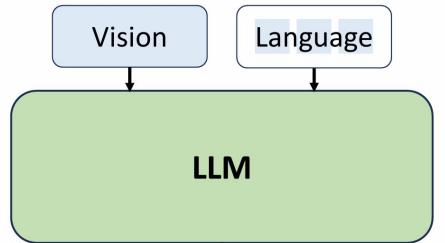
Understanding Only with AR

- Scaling law

Generation Only with AR

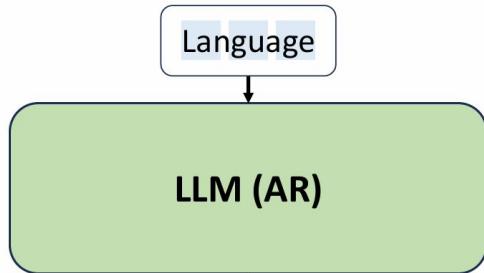
- Scaling Law
- Widely deployment techniques in LLM

Question 2: Can image understanding and generation be executed in one model?



e.g., LLaVA

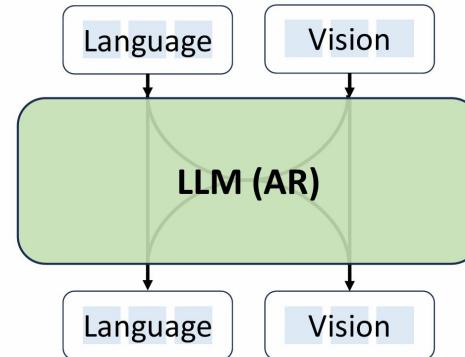
Understanding Only with AR



e.g., LlamaGen

Generation Only with AR

Motivation



e.g., LWM, Chameleon

Unified Model (Understanding & Generation)

- Powerful scaling law
 - Benefit from multi-modal & multi-task pre-training
- Widely deployment techniques in LLM
- **Transformer is all you need**
- **Next-token prediction is all you need**

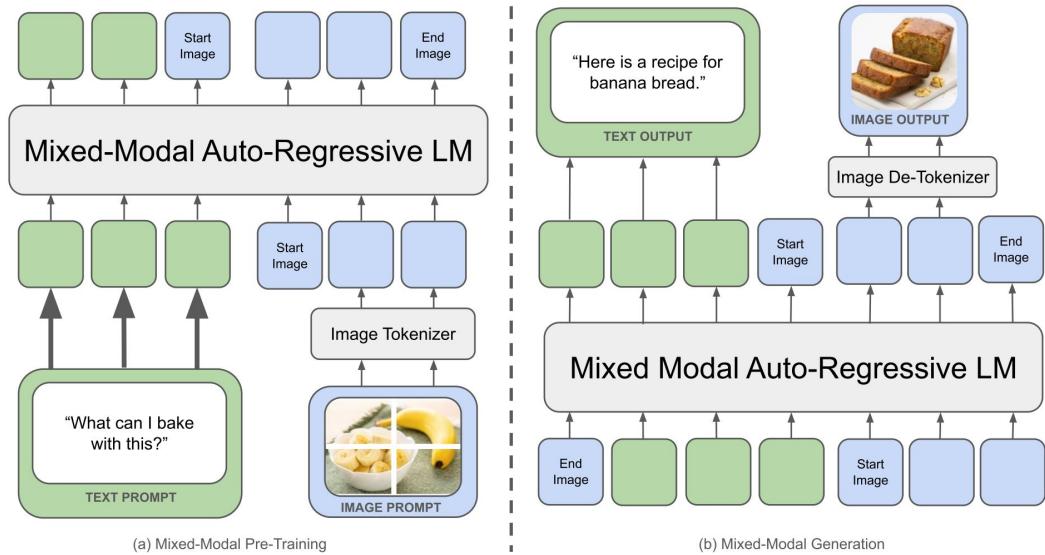
Method

Key feature:

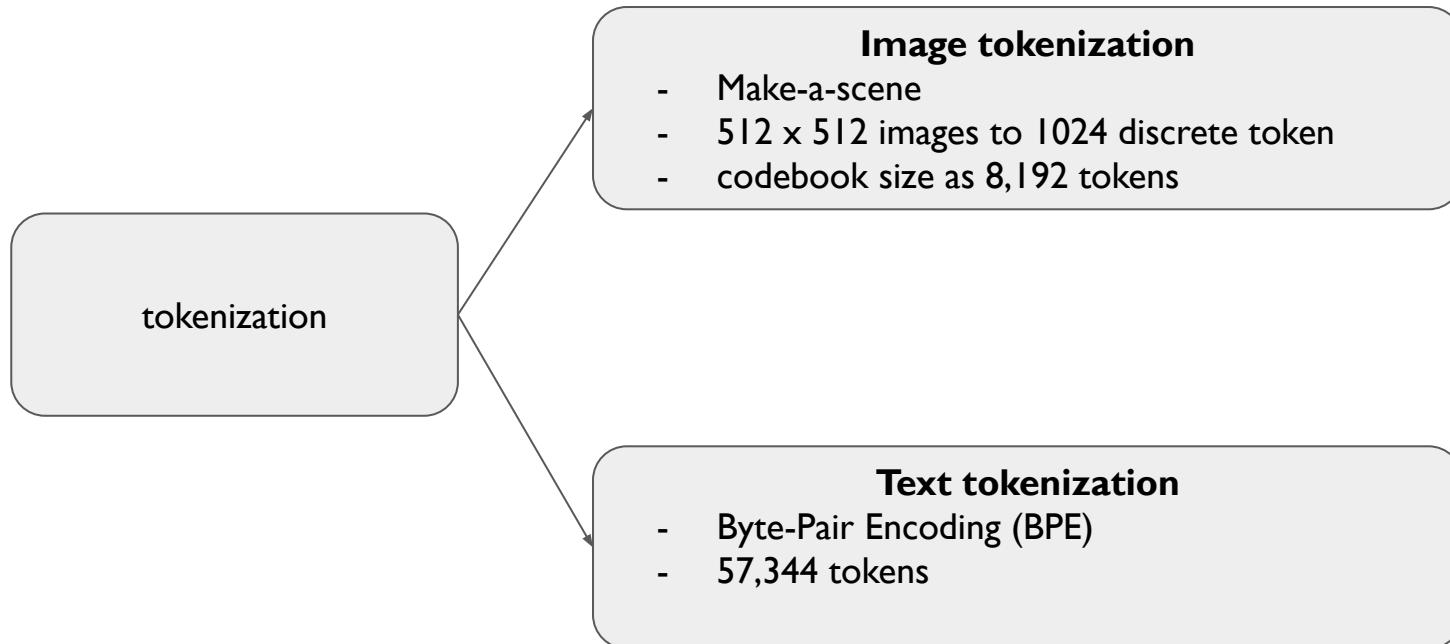
- Early fusion of T & I
- Any mixing of T & I
- Same training loss with LLM

Focus on

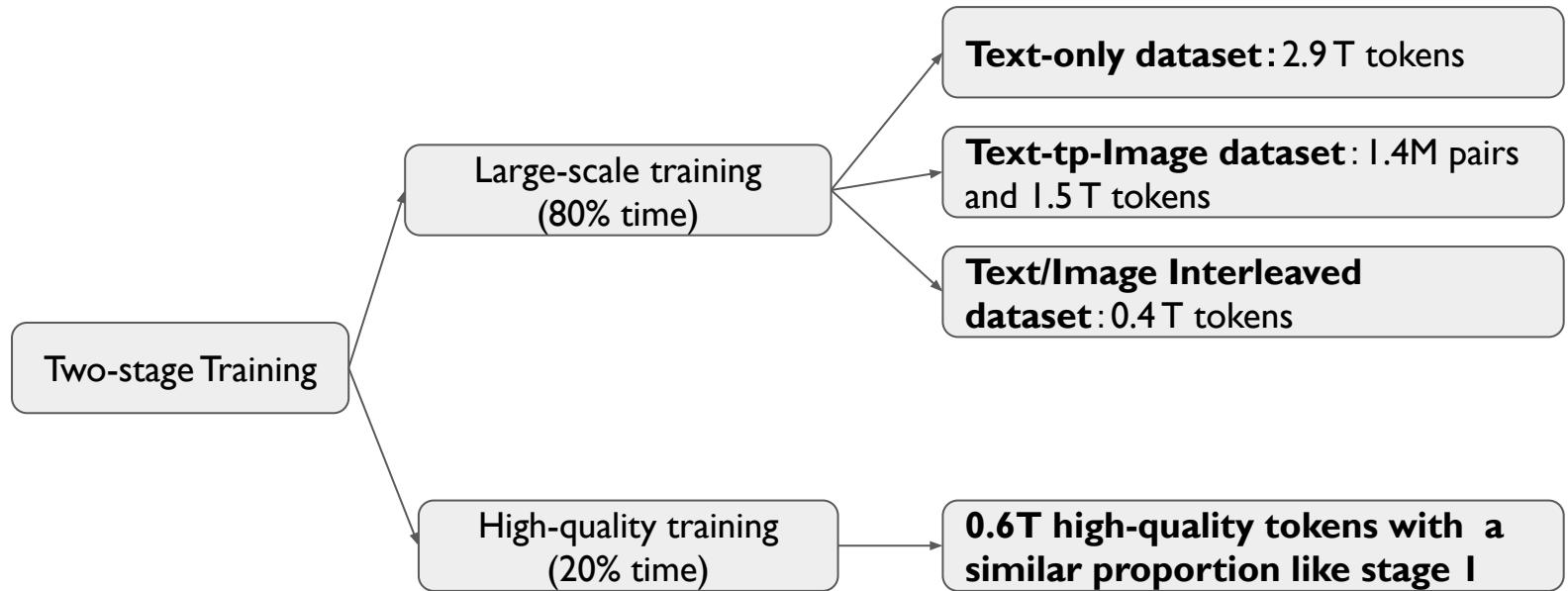
- Tokenization
- Training data & Pipeline
- Optimization stability



Tokenization



Training Data & Pipeline



Overall Optimization

Table 1 Summary of core architecture and optimization decisions made in Chameleon in contrast to LLaMa-1 and LLaMa-2.

Model	Params	Context Length	GQA	Tokens	LR	Epochs	Dropout	Zloss	Qknorm
LLaMa-1	7B	2k	✗	1.0T	3.0×10^{-4}	1.0	0.0	0.0	✗
	33B	2k	✗	1.4T	1.5×10^{-4}	1.0	0.0	0.0	✗
LLaMa-2	7B	4k	✗	2.0T	3.0×10^{-4}	1.0	0.0	0.0	✗
	34B	4k	✓	2.0T	1.5×10^{-4}	1.0	0.0	0.0	✗
Chameleon	7B	4k	✗	4.4T	1.0×10^{-4}	2.1	0.1	10^{-5}	✓
	34B	4k	✓	4.4T	1.0×10^{-4}	2.1	0.0	10^{-5}	✓

Different training settings with pure LLMs:

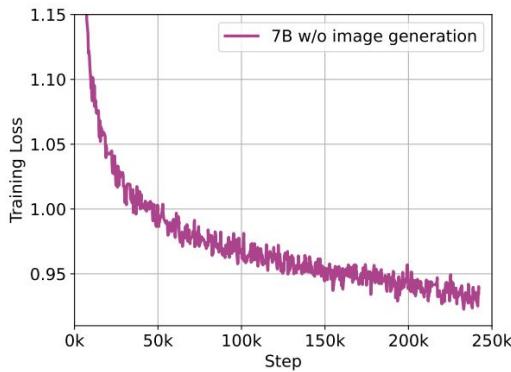
- More training tokens
- Stronger regularization

Question: Why need so strong regularization?

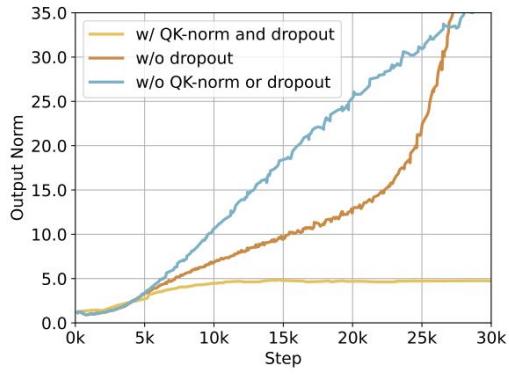
Optimization stability

Observation: It was challenging to maintain stable training when scaling the Chameleon models above 8B parameters and 1T tokens

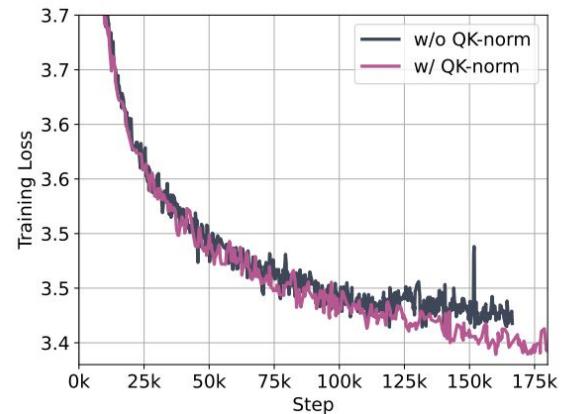
Solution: leveraging a number of regularization to constraint the norm increasing



stable training with
single modality



Blue line: Norm increasing
results unstable training



Black line:unstable training

Alignment

SFT on high-quality dataset for instruction-following ability

- Text: from Llama-2
- Code: from CodeLLaMa
- Image Generation
- Visual Chat
- Image text interleaved generation
- Safety: match unsafe prompt with a refusal response (e.g.“I can’t help with that.”).

Prompt: Can you write a poem using rhyming couplets?

Response: Sure, here's a short poem using rhyming couplets:

The sun sets in the sky,
A beautiful sight to the eye.
The stars come out to play,
A magical sight to behold.

The wind whispers through the trees,
A gentle breeze that brings ease.
The birds sing their sweet melodies,
A soothing sound that calms the soul.

The world is full of wonder,
A place of beauty and thunder.
So let us cherish every moment,
And bask in the beauty all around us.

(a) Text

Prompt: Given a list of integers, find the first number that appears at least twice.

Response:

```
def find_duplicate(my_list):  
    seen = set()  
    for num in my_list:  
        if num in seen:  
            return num  
        seen.add(num)  
    return None
```

(b) Code

Prompt: Renaissance square of Stepanakert

Response:



(c) Image Generation



Prompt: Which elements contribute to this flower arrangement's distinctiveness, and what sets its surroundings apart?

Response: The image features an autumnal-themed floral arrangement in a classically designed vase. The arrangement stands out with its striking combination of red maple leaves, purple flowers, and pink-red roses. The vase's rosettes give it an antique design. In the background, blurry green vegetation is visible, creating a contrast to the autumnal theme of the bouquet. The red maple leaves, both in the arrangement and scattered around the vase, evoke autumn, while the green vegetation suggests a more summer-like setting. The outdoor setting is bright and sunny, adding to the contrast between the arrangement and its settings.

(d) Visual Chat

Gyumri is an urban municipal community and the second-largest city in Armenia, serving as the administrative center of Shirak Province in the north-western part of the country.



Archaeological excavations conducted throughout the Soviet period have shown that the area of modern-day Gyumri has been populated since at least the third millennium BC.



(e) Interleaved Text/Image Generation

Inference Example

Interleaved image and text generation from Chameleon

Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2-3 sentences.



Response: Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird:

Puffin Puffins are small, aquatic birds that are found in the northern hemisphere. They have distinctive black and white feathers, as well as a brightly colored beak that is used for attracting mates and defending their territory. Puffins are known for their distinctive "puffin" call, which sounds like a loud, guttural chuckle. They feed on fish, which they catch by diving into the water from the air. Puffins are colonial birds, and they breed in large colonies on rocky cliffs and islands. Here is an image of this adorable bird:



Golden Pheasant The Golden Pheasant is a brightly colored bird native to the forests of China. It is a member of the pheasant family and is known for its striking plumage, which features a golden-yellow body, red face, and green tail feathers. The Golden Pheasant is a ground-dwelling bird, and it feeds on a diet of insects, seeds, and fruits. It is a popular game bird, and is often hunted for its meat and feathers. In captivity, the Golden Pheasant is sometimes kept as an ornamental bird, due to its striking appearance. Here is an image of this beautiful bird:

I hope these images and descriptions have helped to brighten your day! If you have any other questions or requests, please don't hesitate to ask.

Inference Example

Image understanding and image generation



Prompt: What can I bake with this? Give me the full recipe and show me an image with the final dish.

Response: Bananas are a versatile fruit that can be used in various baked goods, such as muffins, bread, cakes, and cookies. They add natural sweetness, moisture, and a subtle banana flavor to the finished product. Additionally, bananas are rich in potassium, fiber, and other nutrients, making them a healthy addition to your baking repertoire.

Here is a simple recipe for banana bread that uses 3 ripe bananas:

Ingredients:

- 3 ripe bananas, mashed
- 1 1/2 cups all-purpose flour
- 1 tsp baking powder
- 1/2 tsp baking soda
- 1/2 tsp salt
- 1/4 cup unsalted butter, softened
- 1 cup granulated sugar
- 2 large eggs
- 1 tsp vanilla extract
- 1/2 cup chopped walnuts (optional)

Instructions:

- Preheat your oven to 350°F (180°C). Grease a 9x5-inch loaf pan and set it aside.
- In a medium bowl, whisk together the flour, baking powder, baking soda, and salt. Set aside.
- In a large mixing bowl, use an electric mixer to cream together the butter and sugar until light and fluffy, about 2-3 minutes.
- Beat in the eggs one at a time, followed by the mashed bananas and vanilla extract.



- Gradually mix in the dry ingredients until just combined, being careful not to overmix.
- If using walnuts, fold them into the batter.
- Pour the batter into the prepared loaf pan and smooth the top.
- Bake for 50-60 minutes, or until a toothpick inserted into the center of the bread comes out clean.
- Remove the bread from the oven and let it cool in the pan for 10 minutes before transferring it to a wire rack to cool completely.

Results

Human evaluation with 12 tasks

Advice: 10.2%

What does a meningitis rash look like? What are the other symptoms I should be on the lookout for?

How-to: 12.5%

How do I properly clean my TV screen? I used Windex and now there are towel fibers and wipe marks all over. Show me some reference photos.

Explanation: 14.4%

I've been studying classical French art, and my favorite so far is his painting seen here:
Could you please give me a few images of other contemporary artworks that have this same aesthetic?



Hypothetical: 5.6%

What would the modern-day vehicle look like if oil had never been discovered?

Brainstorming: 18.6%

Show me a Middle Eastern alternative to these dishes. <img1> <img2>



Article: 3.1%

Write me an introduction to a story about knick-knacks, and finish the story by shifting the focus with an image.

Story: 3.9%

Can you create and illustrate a short story for children about an octopus that can't stop eating pizza?

Identification: 9.3 %



Is the below image a Shetland Pony? If not, what is it, and can you show me a Shetland Pony?

Comparison: 9.6%

Please tell me what the difference between these two creatures is, and show me some more examples. <img1> <img2>



Report: 5.4%

Who designed the church in the image below, and what's the name of the Church? Can you please provide me with additional photos of famous landmarks designed by the same architect?



Other: 5.2%

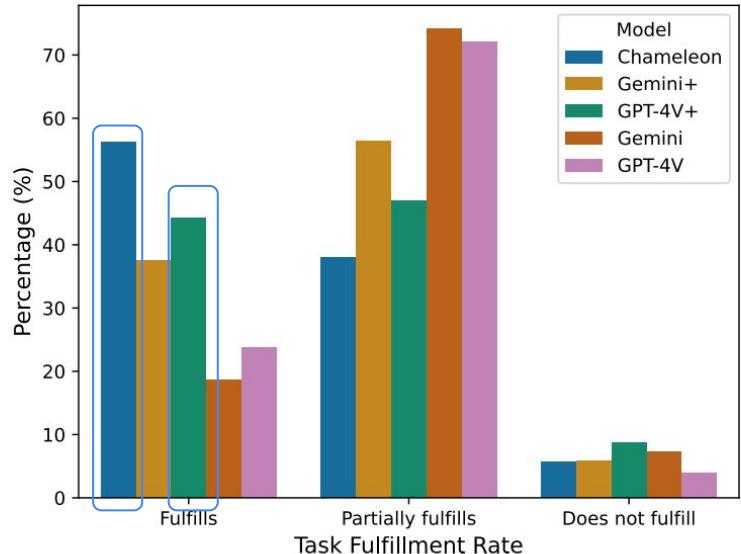
Create a decal for my truck that features running horses as well as the TRD insignia. Use black to gray gradients.

Reasoning: 2.1%

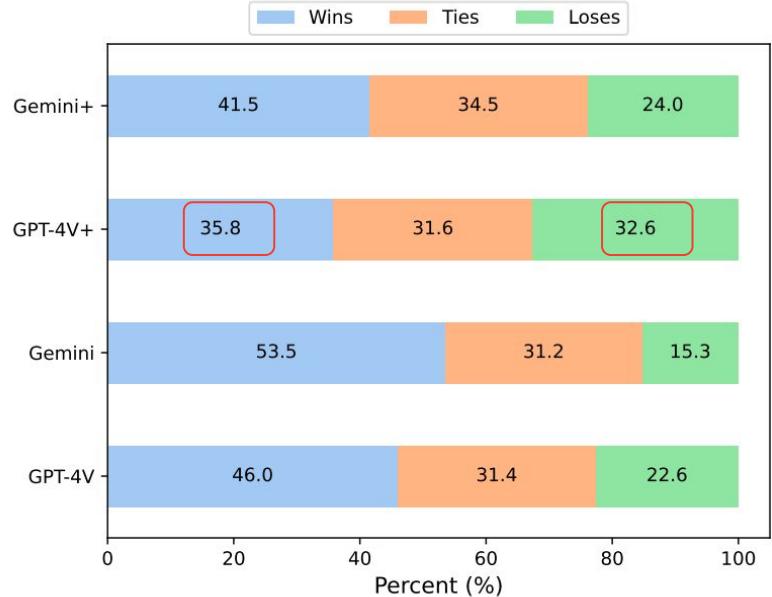
What is typically found at a construction site? Show me a construction site that has a crane.

Results

Higher fulfillment rate and win rate



(a) The prompt task fulfillment rates.



(b) Chameleon vs. the baselines: Gemini+, GPT-4V+, Gemini, GPT-4V.

Summary

- Chameleon achieve unified visual understanding and generation
- The key to Chameleon's success is its fully token-based architecture, which allows for seamless information integration across modalities.
- Chameleon introduces novel techniques for stable and scalable training of early-fusion models.