What's Missing in Text-to-Image Generation? Current Models and Paths Forward.

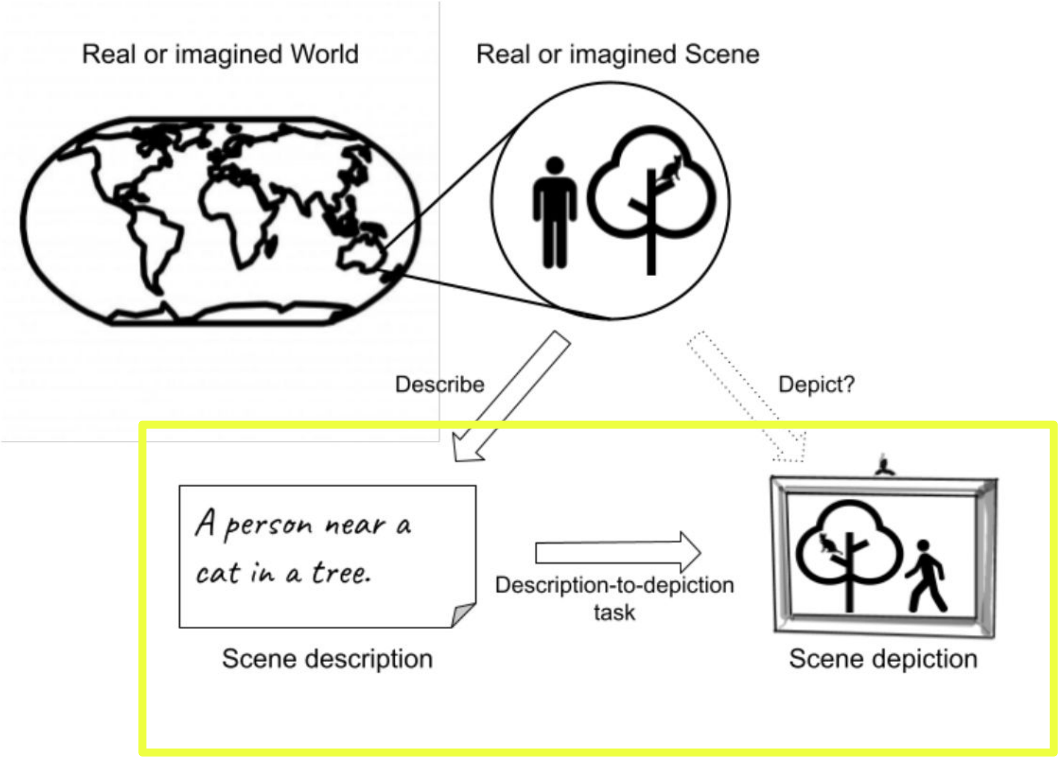
Jason Baldridge
Google Research

ICDM 2022 Workshop
Foundation Models in Vision and Language
November 28, 2022

Metal statue of a rabbit detective standing under a street light near a brick lined street on a rainy night. Bokeh.

Google Research

Description-to-depiction / Text-to-Image Generation



Real or imagined World

Real or imagined Scene

Describe

Depict?

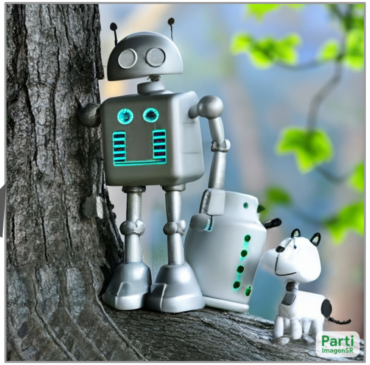
A person near a cat in a tree.

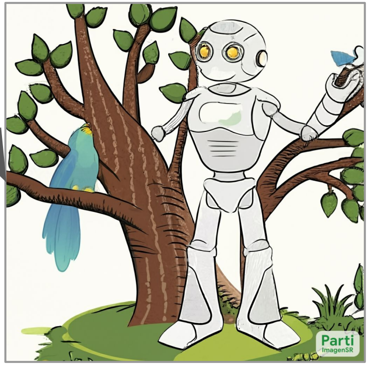
Description-to-depiction task

Scene description

Scene depiction

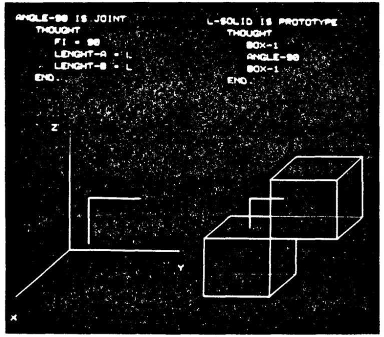
A robot and its pet in a tree.





Text-to-image generation, over time (sampled)

Adorni and Di Manzo **(1983)**



```
ANGLE-90 IS JOINT                L-SOLID IS PROTOTYPE
  THOUGHT                          THOUGHT
    FI = 90                          BOX-1
    LENGHT-A = L                     ANGLE-90
    LENGHT-B = L                     BOX-1
  END.                             END.
```

FIG.4 - Definition of a Simple Jointing Element and Use of this Element to build a More Complex Object

WordsEye
Coyne and Sproat **(2001)**

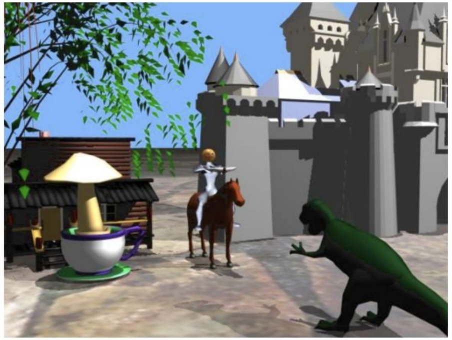
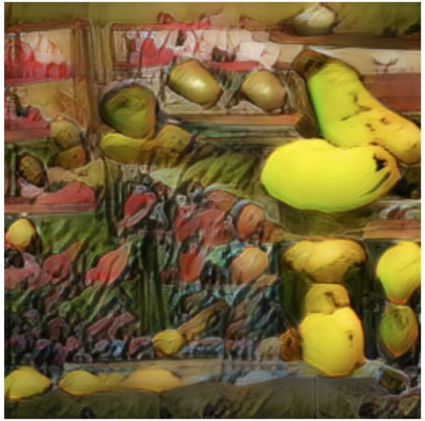


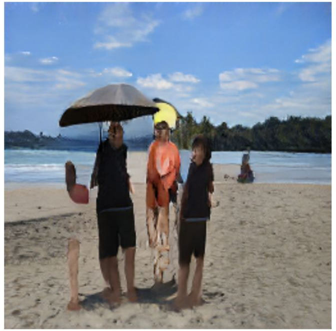
Figure 1: *John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.*

AttnGAN
Xu et al **(2017)**



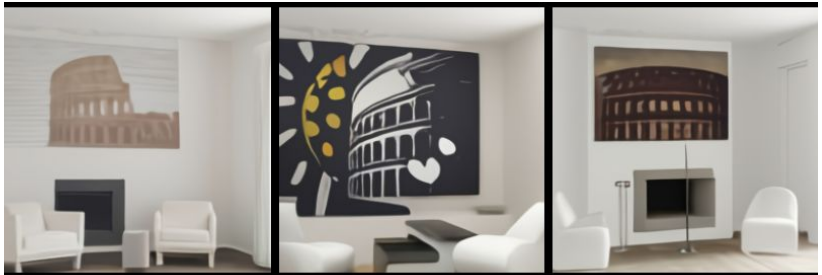
a fruit stand display with bananas and kiwi

XMC-GAN
Zhang, Koh et al **(2021)**



In this picture there are two members lying on the beach in the sand under an umbrella. There are some people standing here. In the background there is water

DALL-E
Ramesh et al **(2021)**

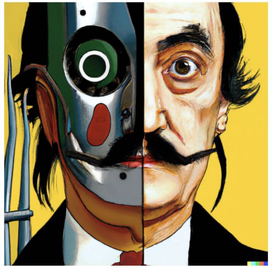


a living room with two white armchairs and a painting of the collosseum. the painting is mounted above a modern fireplace.

And many more!

2022: Text-to-image generation arrives!

DALL-E 2 Ramesh et al (**2022**)

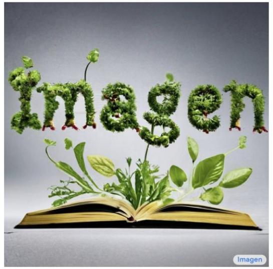


vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

IMAGEN Saharia et al (**2022**)



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

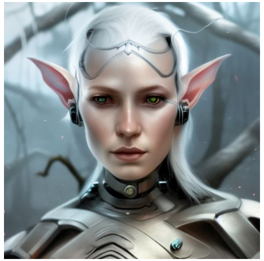
PARTI Yu et al (**2022**)

A. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

B. *A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The city of Los Angeles is in the background. Hi-res DSLR photograph.*

C. *A high-contrast photo of a panda riding a horse. The panda is wearing a wizard hat and is reading a book. The horse is standing on a street against a gray concrete wall. Colorful flowers and the word "PEACE" are painted on the wall. Green grass grows from cracks in the street. DSLR photograph. daytime lighting.*

STABLE DIFFUSION Rombach et al (**2022**)



Parti: Pathways Autoregressive Text-to-Image model

- We explore text-to-image generation via autoregressive seq-to-seq Transformer models, a high performing architecture on many tasks, including machine translation, speech recognition, conversational modeling, image captioning, and many others.

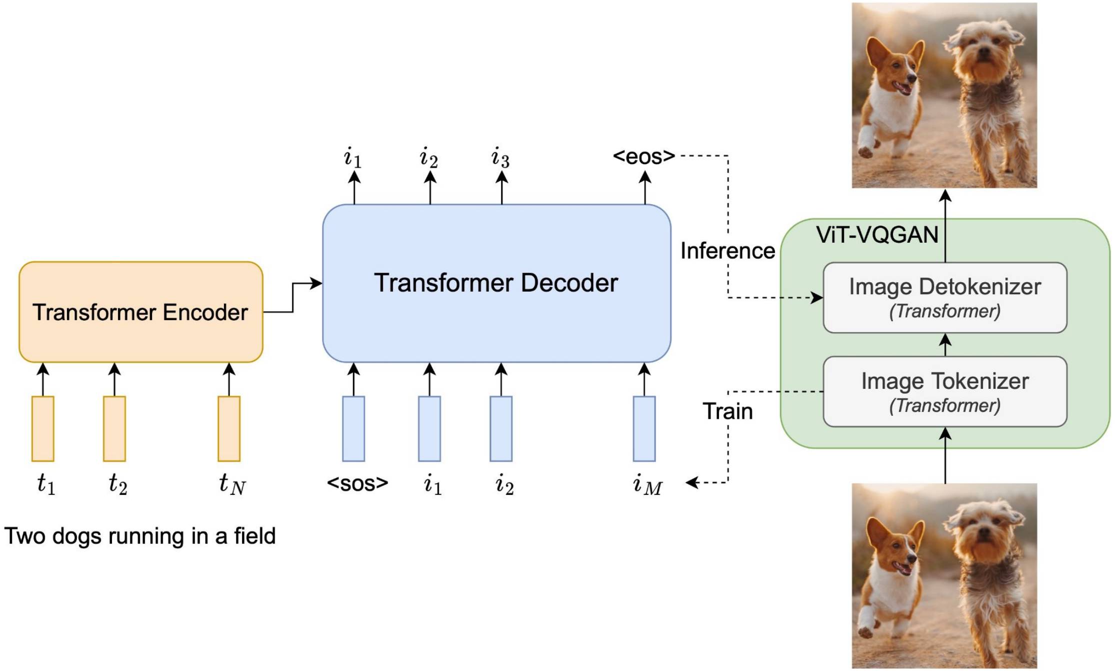
- *Why?* To better enable **scaling** and **unification** in language (e.g. GPT-3, GLaM, PaLM) and multimodal models (e.g., CoCa, PaLI: Image + Text ⇒ Text).
  - Scaling unlocks new capabilities / emergent abilities [1] through zero-shot, few-shot, transfer learning.
  - Unification delivers general-purpose ML models, amortizes training costs, and provides opportunities to scale further.

A super math wizard cat, richly textured oil painting.

[1] Wei et al (2022). Emergent Abilities of Large Language Models.

Parti’s components

- Standard encoder-decoder autoregressive architecture, treating text-to-image generation akin to machine translation.
- ViT-VQGAN [1] as image tokenizer to encode images as sequences of discrete tokens.
- Scaling via GSPMD [2].
- Overall approach builds on and improves the original DALL-E and CogView models.



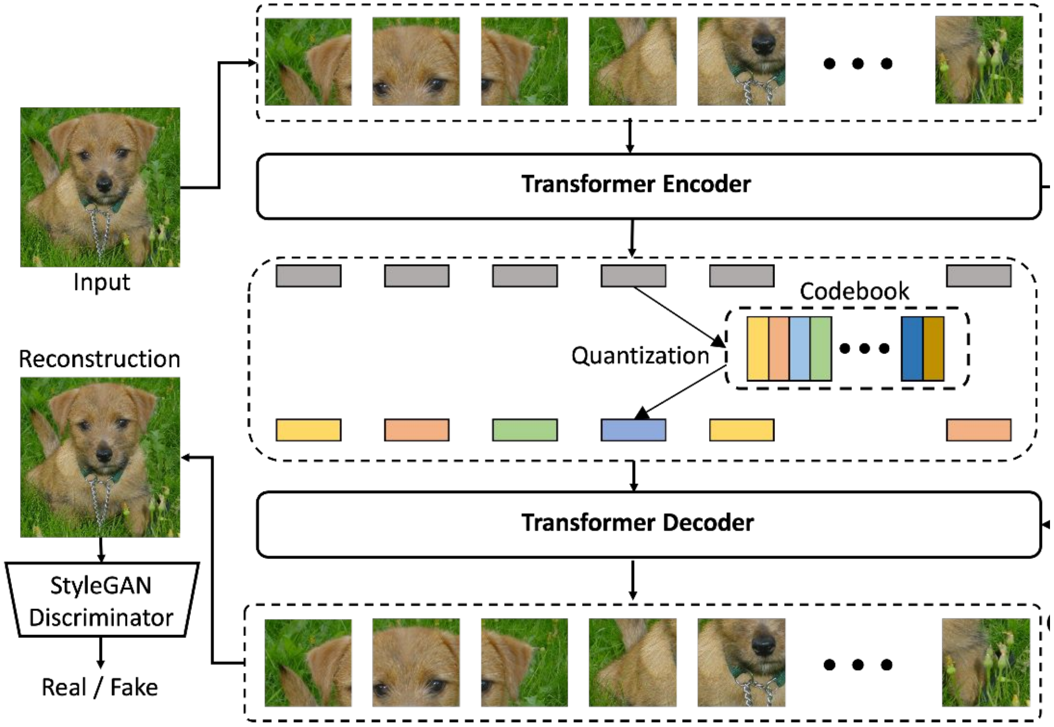
[1] Yu et al (2022). Vector-quantized Image Modeling with Improved VQGAN.
[2] Xu et al (2021). GSPMD: General and Scalable Parallelization for ML Computation Graphs.

ViT-VQGAN

- Goal: learn a visual tokenizer – a vocabulary for image patches – to enable efficient autoregressive modeling.
  - Convert 256x256 image into 32x32 latent codes.

- Builds on previous work such as discrete Variational Autoencoders and VQGAN.

- Uses vision transformers, additional losses and scaling to improve vocabulary learning and image reconstruction.



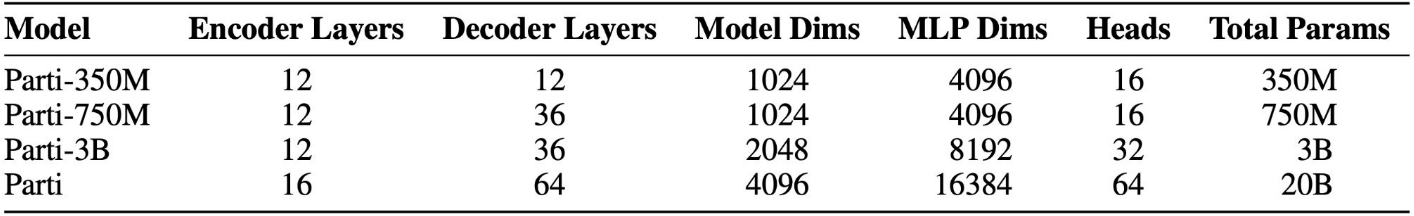
Additional important details

- Text Encoder pretraining, using both masked language modeling (BERT) and image-text contrastive objectives (CLIP, ALIGN, etc).

- Classifier-Free Guidance
  - Important for diffusion models (DALL-E 2, Imagen)
  - Adapted from diffusion for autoregressive by Katherine Crowson and used in Meta's Make-a-Scene model.

- CoCa Reranking: batch-sample 16 images per text input and rank based on image-text similarity using the Constrastive Captioners model.

Scaling Parti

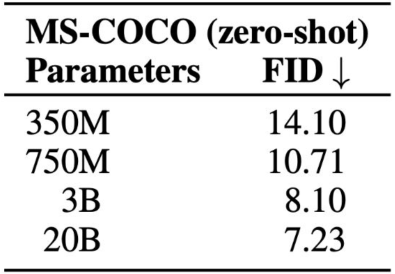


Model	Encoder Layers	Decoder Layers	Model Dims	MLP Dims	Heads	Total Params
Parti-350M	12	12	1024	4096	16	350M
Parti-750M	12	36	1024	4096	16	750M
Parti-3B	12	36	2048	8192	32	3B
Parti	16	64	4096	16384	64	20B

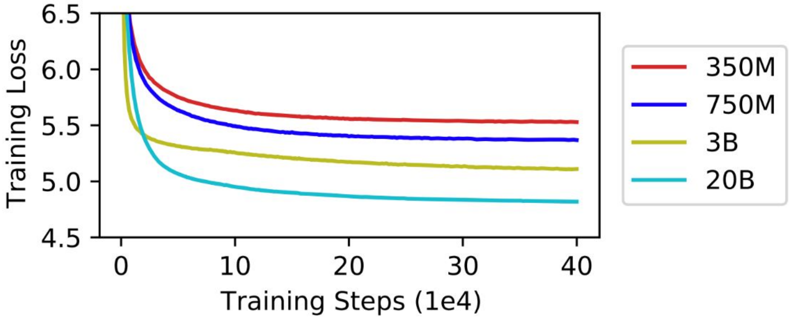
Size variants of Parti. Both encoder and decoder are based on Transformers.

Effect of scaling.

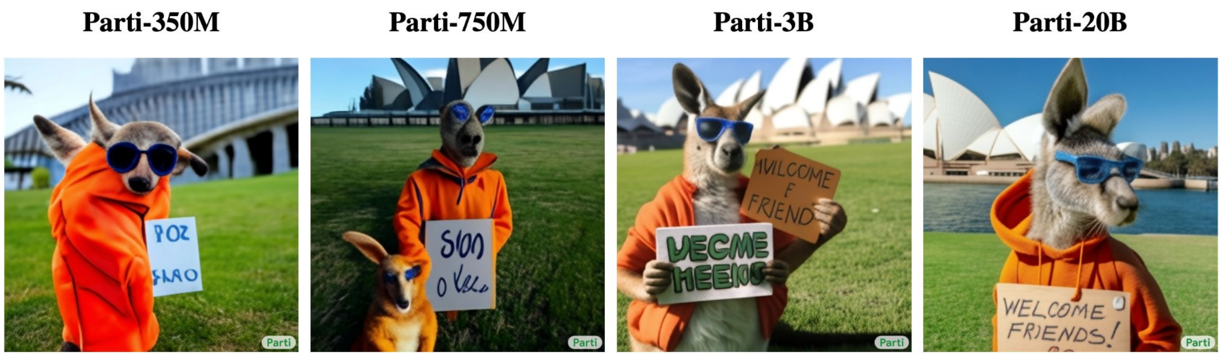
- Zero-shot FID scores (left) for MS-COCO (2014)
- Training loss curves of the corresponding models (right).



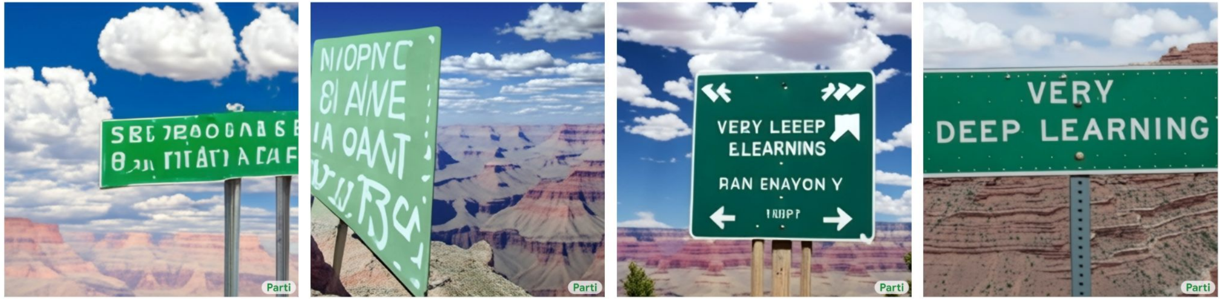
MS-COCO (zero-shot)	
Parameters	**FID ↓**
350M	14.10
750M	10.71
3B	8.10
20B	7.23



Scaling Parti


Parti-350M Parti-750M Parti-3B Parti-20B

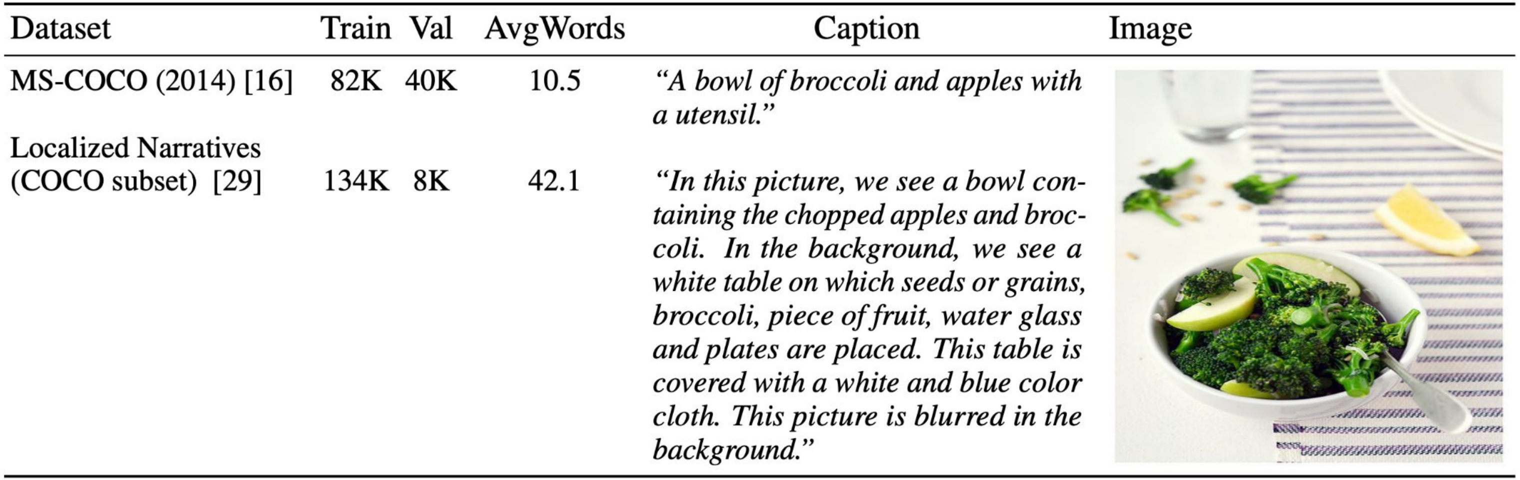
A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!



A green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon. Puffy white clouds are in the sky.

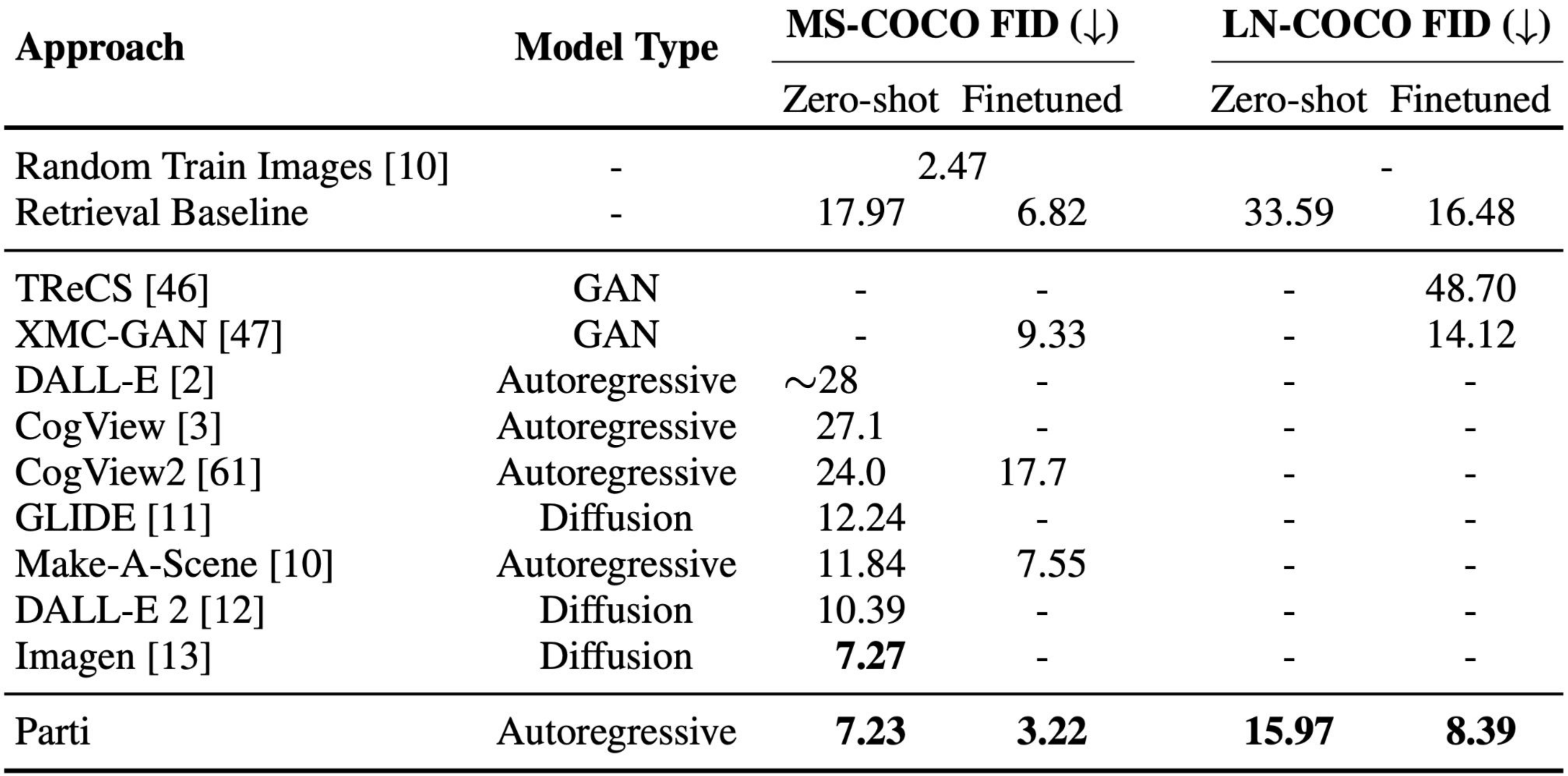
Evaluation with MS-COCO and Localized Narratives

- Evaluation of MSCOCO FID is standard, but has limitations.
- Localized Narratives: 4x longer texts on average.
  - Used LN-COCO subset.



Dataset	Train	Val	AvgWords	Caption	Image
MS-COCO (2014) [16]	82K	40K	10.5	*“A bowl of broccoli and apples with a utensil.”*	
Localized Narratives (COCO subset) [29]	134K	8K	42.1	*“In this picture, we see a bowl containing the chopped apples and broccoli. In the background, we see a white table on which seeds or grains, broccoli, piece of fruit, water glass and plates are placed. This table is covered with a white and blue color cloth. This picture is blurred in the background.”*	

Automatic evaluation (FID) for image realism



Approach	Model Type	MS-COCO FID (↓)		LN-COCO FID (↓)	
		Zero-shot	Finetuned	Zero-shot	Finetuned
Random Train Images [10]	-	2.47		-	
Retrieval Baseline	-	17.97	6.82	33.59	16.48
TReCS [46]	GAN	-	-	-	48.70
XMC-GAN [47]	GAN	-	9.33	-	14.12
DALL-E [2]	Autoregressive	∼28	-	-	-
CogView [3]	Autoregressive	27.1	-	-	-
CogView2 [61]	Autoregressive	24.0	17.7	-	-
GLIDE [11]	Diffusion	12.24	-	-	-
Make-A-Scene [10]	Autoregressive	11.84	7.55	-	-
DALL-E 2 [12]	Diffusion	10.39	-	-	-
Imagen [13]	Diffusion	**7.27**	-	-	-
Parti	Autoregressive	**7.23**	**3.22**	**15.97**	**8.39**

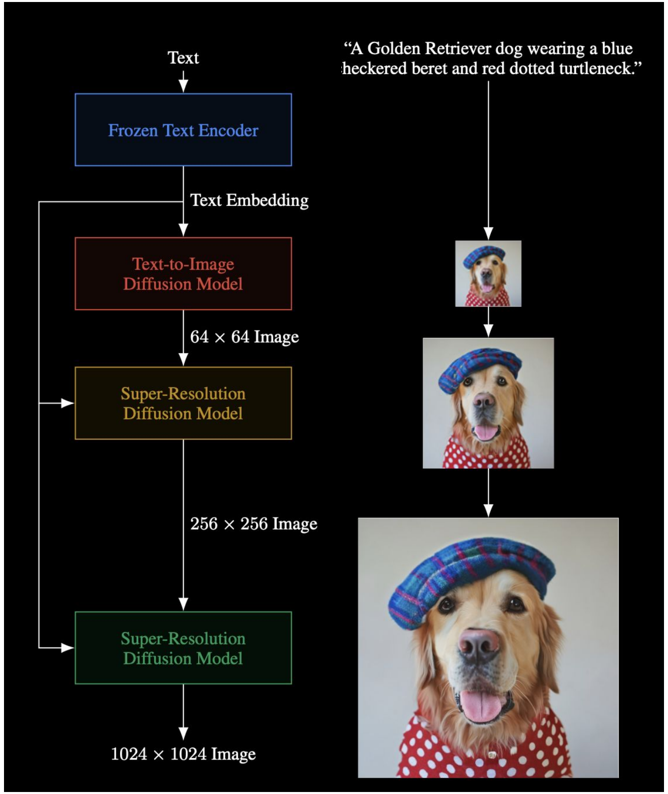
Imagen: Diffusion-based text-to-image generation

- Imagen [1] is another leading text-to-image generation model, also from Google Research.

- Uses diffusion, a general ML approach based on denoising that is popular in text-to-image generation and is being applied to a range of tasks.

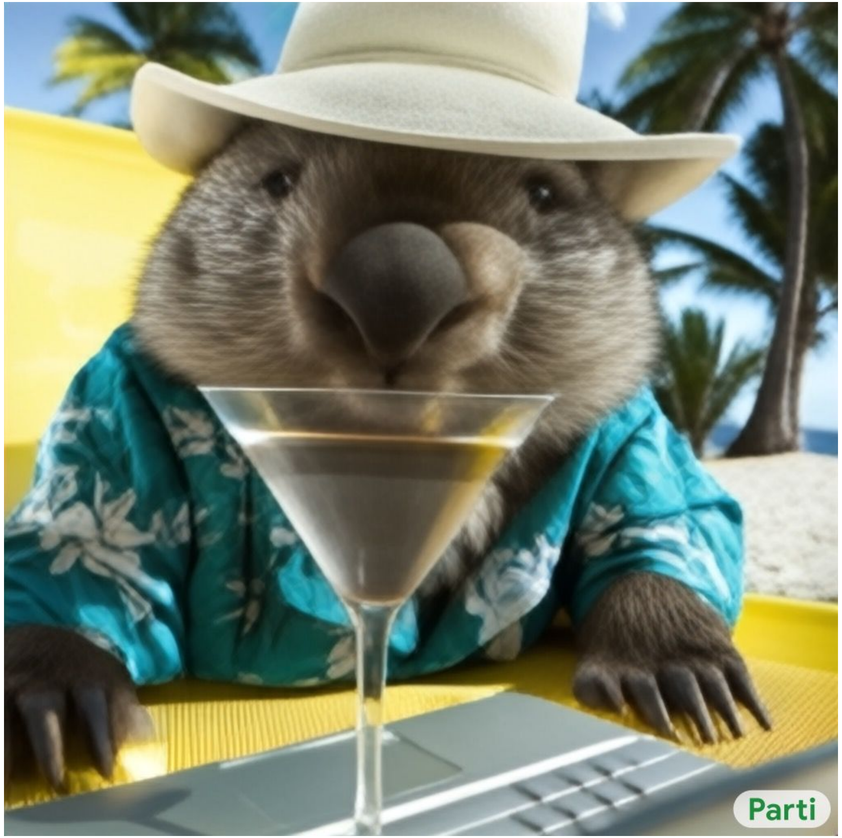
- A pipeline of three diffusion models:
  - Stage 1: Text → 64x64 image
  - Stage 2: Text + 64x64 → 256x256 image
  - Stage 3: Text + 256x256 → 1Kx1K image

- Achieves similar FID on MS-COCO to Parti and shows gains from scaling the text encoder (T5 variants).


Text
"A Golden Retriever dog wearing a blue heckered beret and red dotted turtleneck."
Frozen Text Encoder
Text Embedding
Text-to-Image Diffusion Model
64 × 64 Image
Super-Resolution Diffusion Model
256 × 256 Image
Super-Resolution Diffusion Model
1024 × 1024 Image

[1] Saharia et al (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

Strong super-resolution makes a big visual impact!



Parti super-resolution.
(40 million parameters)

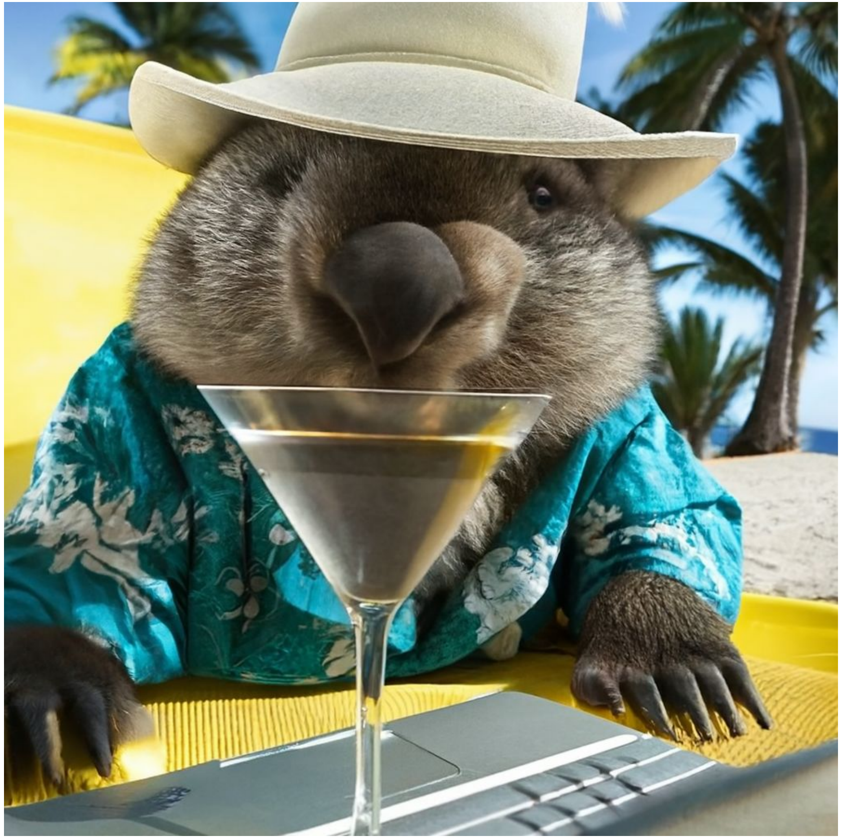
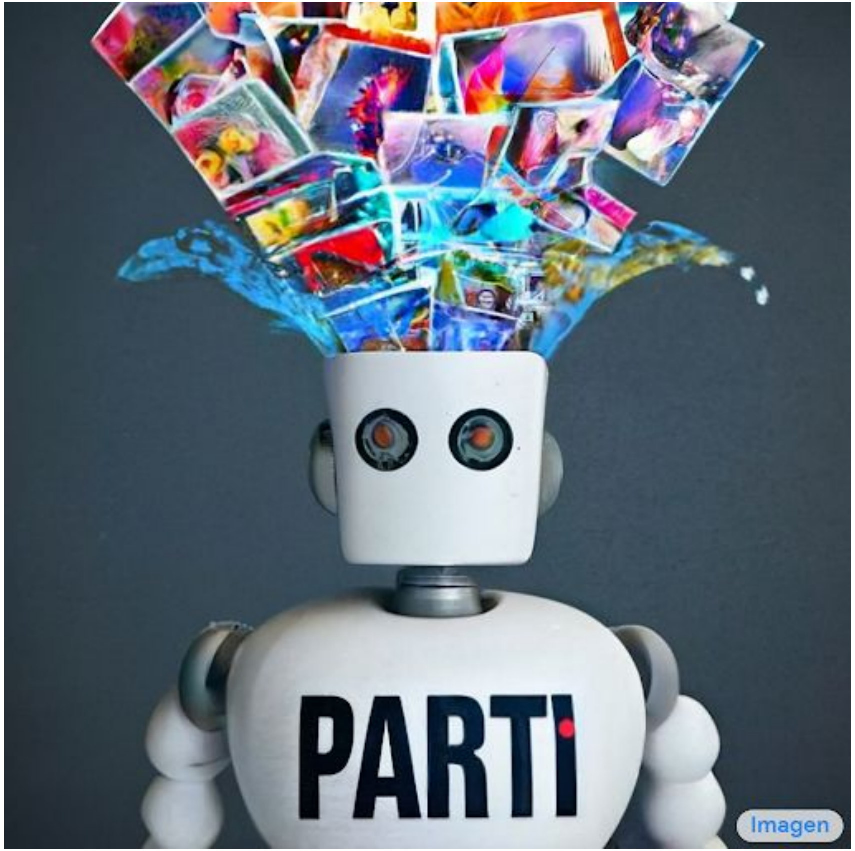
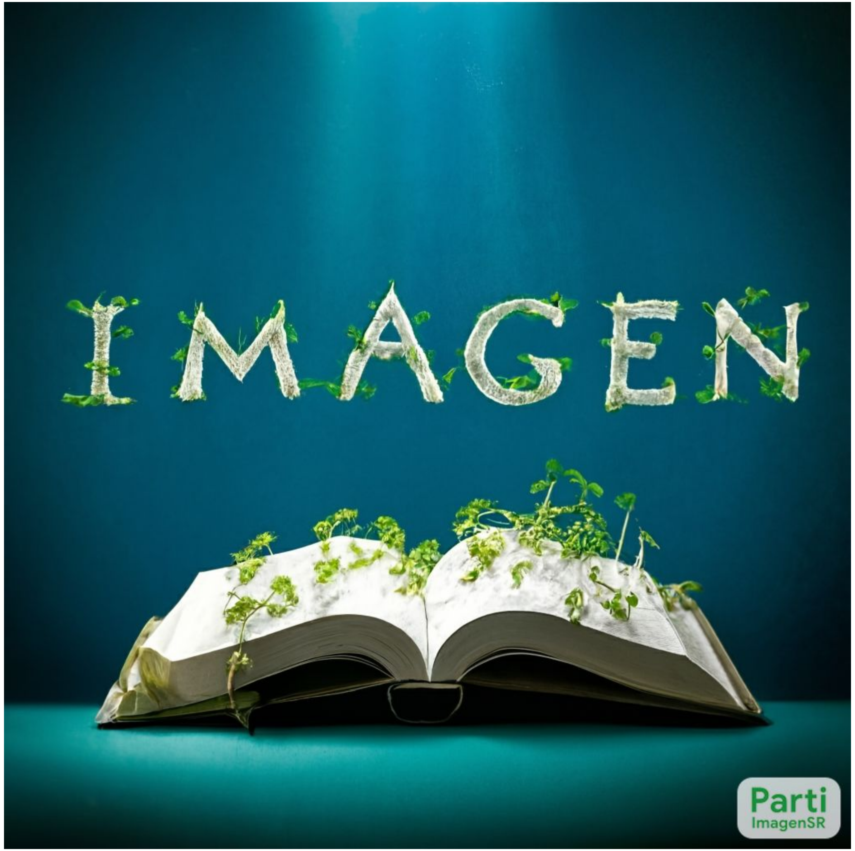


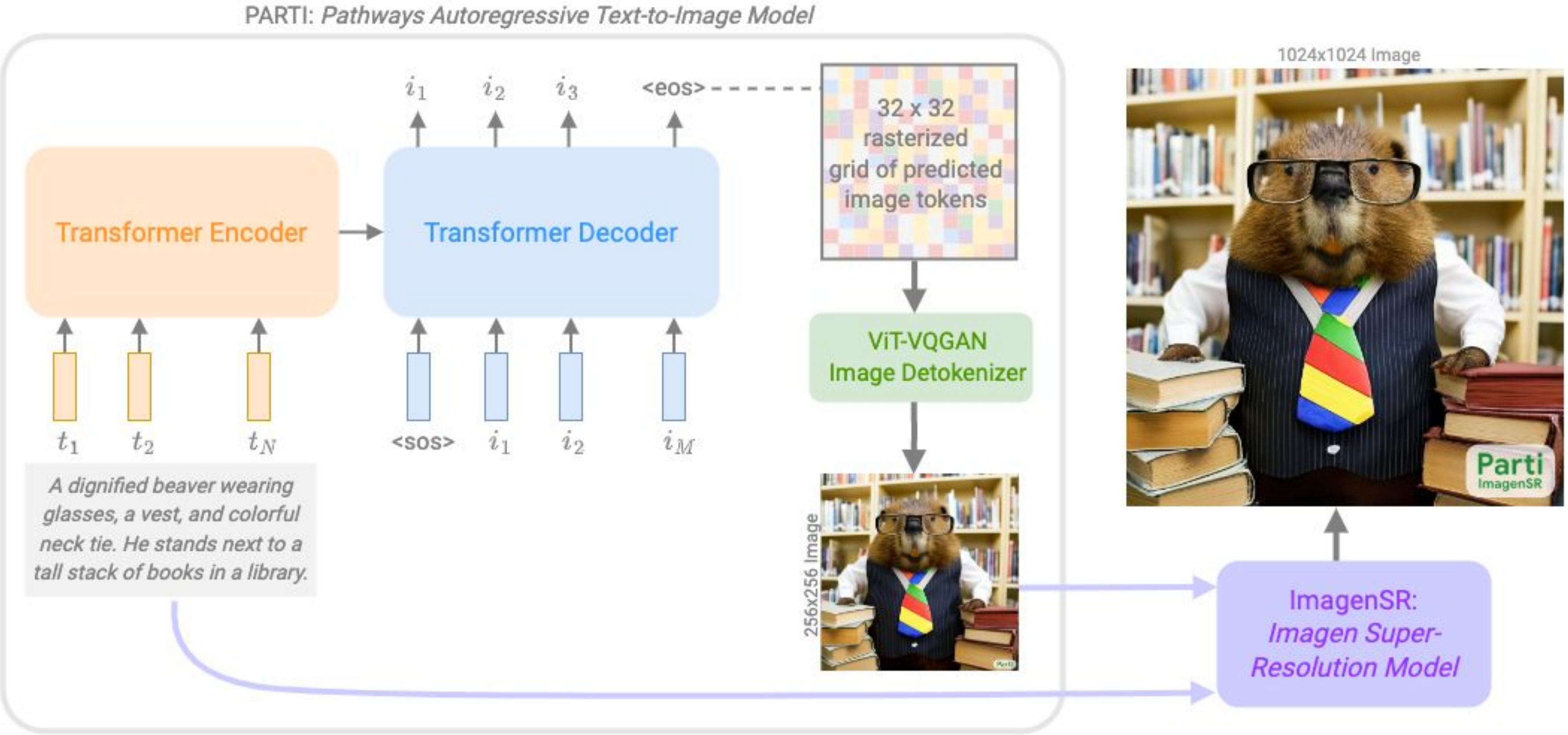
Imagen super-resolution on Parti 256x256 output.
(400 million parameters)

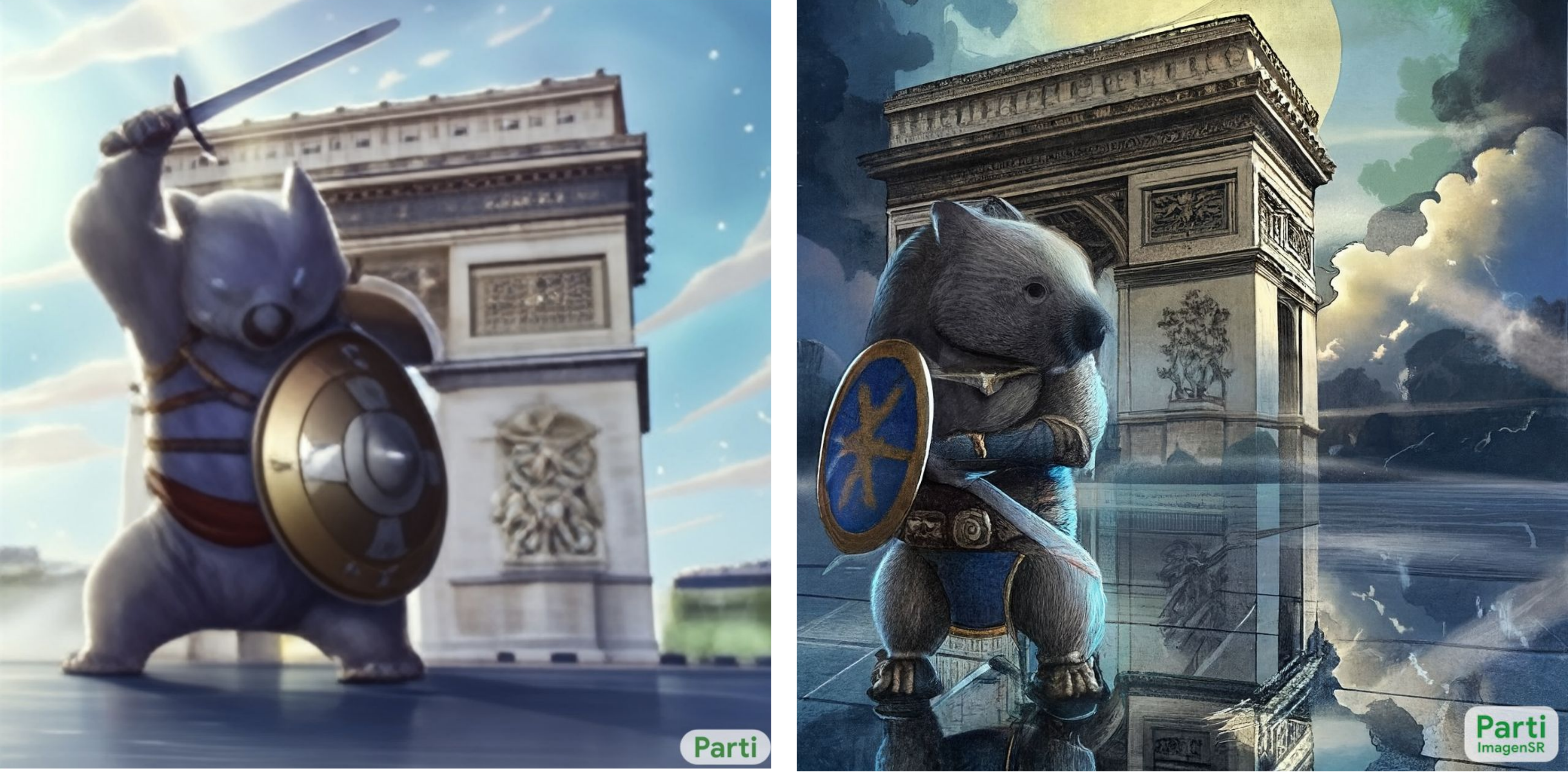
Imagen and Parti partnership!



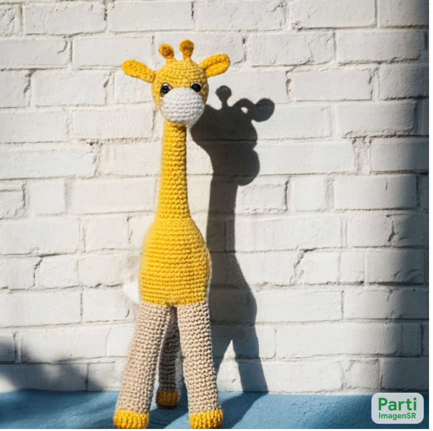


Combined system: Parti + Imagen Super Resolution!

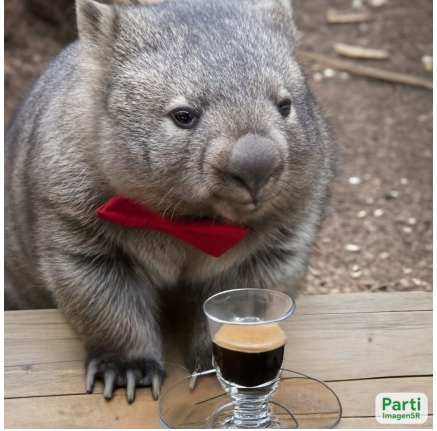

PARTI: *Pathways Autoregressive Text-to-Image Model*
i_1
i_2
i_3
<eos>
32 x 32 rasterized grid of predicted image tokens
Transformer Encoder
Transformer Decoder
ViT-VQGAN Image Detokenizer
t_1
t_2
t_N
<sos>
i_1
i_2
i_M
A dignified beaver wearing glasses, a vest, and colorful neck tie. He stands next to a tall stack of books in a library.
256x256 Image
1024x1024 Image
Parti
ImagenSR
ImagenSR:
Imagen Super-Resolution Model



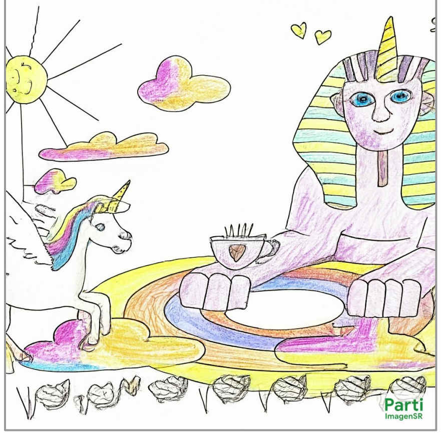
A warrior wombat holding a sword and shield in a fighting stance. The wombat stands in front of the Arc de Triomphe on a day shrouded mist with the sun high in the sky. Realistic anime illustration. (on right: +“by John Tenniel”)



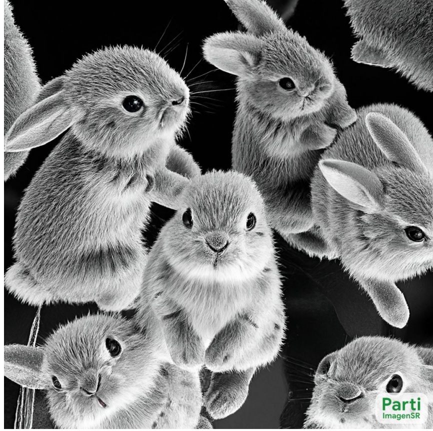
Amigurumi giraffe casting a shadow on a faded white brick wall.



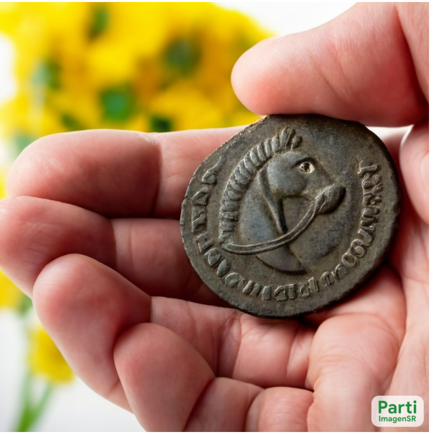
A wombat wearing a bowtie and drinking an espresso. DSLR photograph.



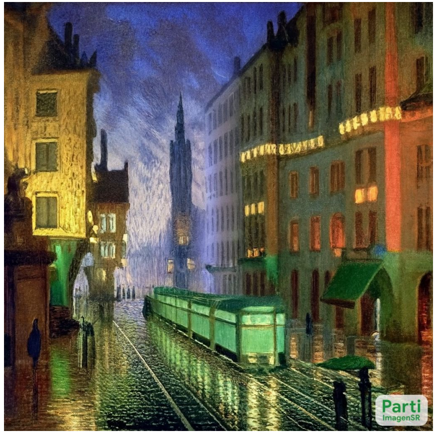
The great sphinx having tea with a unicorn. Kids drawing.



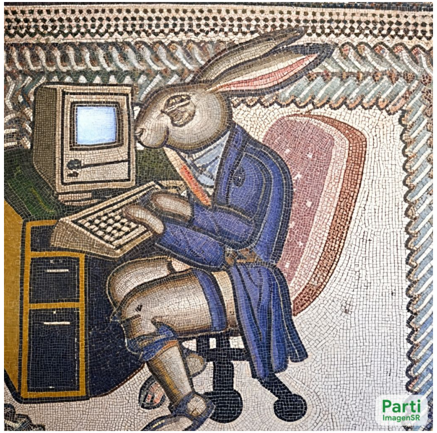
Electron microscope view of tiny rabbits jumping around.



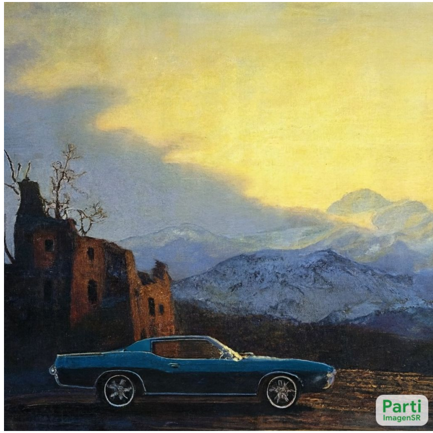
A hand holding an ancient coin with a horse's head profile. Vase with yellow flowers in background. Studio lighting. Bokeh. DSLR photograph.



A futuristic street train a rainy street at night in an old European city. Painting by David Friedrich, Claude Monet and John Tenniel.



A well dressed rabbit using a 1980's desktop computer. Byzantine mosaic.

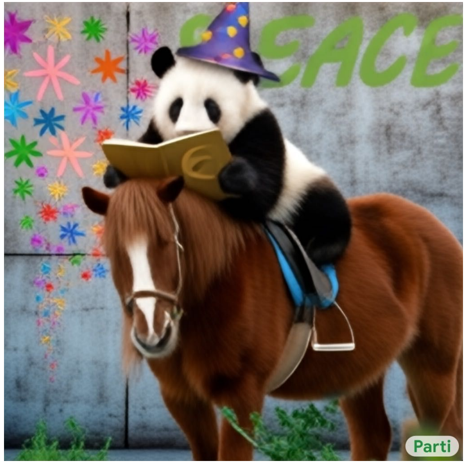


Distant view of a 1970s American muscle car near a ruined castle, with snow-capped mountains in the background at sunset. Highly detailed and masterful oil painting on canvas by Rembrandt.

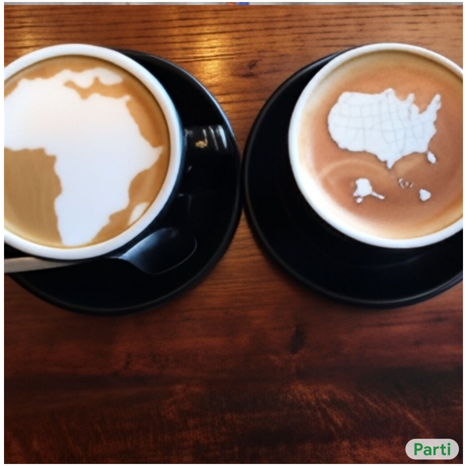
Key ingredients for text-to-image models

- Text pretraining
  - Large transformer language models (e.g. T5-XXL): Imagen
  - Multimodal dual encoders (CLIP): DALL-E2, Stable Diffusion
  - Combination of masked language modeling and multimodal contrastive learning: Parti
- Visual tokens (dVAE, VQGAN, ViT-VQGAN): Parti, Stable Diffusion
- Diffusion: DALL-E 2, Imagen, Stable Diffusion
- Autoregressive encoder-decoders: Parti
- Classifier-free guidance: DALL-E 2, Imagen, Parti, Stable Diffusion
- Scale: billions of parameters
- Data: hundreds of millions to several billion image-text pairs.

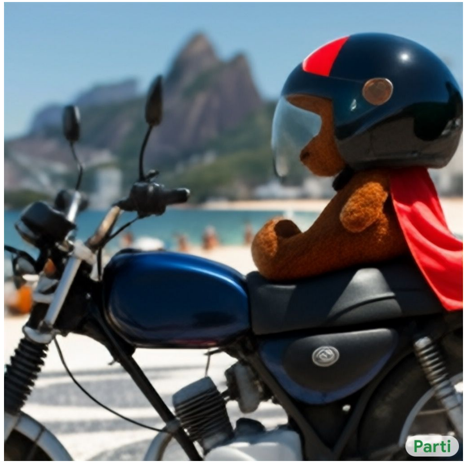
Scale matters: Reflecting detail and world knowledge



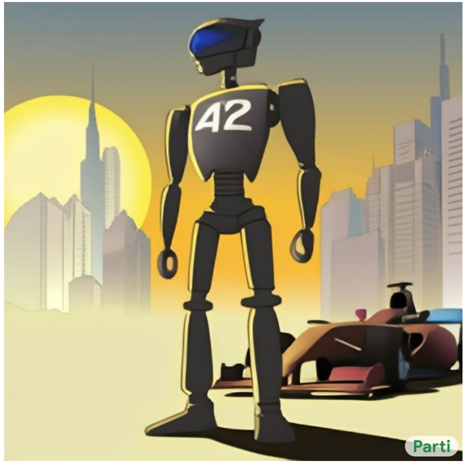
A high-contrast photo of a panda riding a horse. The panda is wearing a wizard hat and is reading a book. The horse is standing on a street against a gray concrete wall. Colorful flowers and the word "PEACE" are painted on the wall. Green grass grows from cracks in the street. DSLR photograph. daytime lighting.



Two cups of coffee, one with latte art of a map of the United States. The other has latte art of a map of Africa.



A teddy bear wearing a motorcycle helmet and cape is riding a motorcycle in Rio de Janeiro with Daos Irmãos in the background.



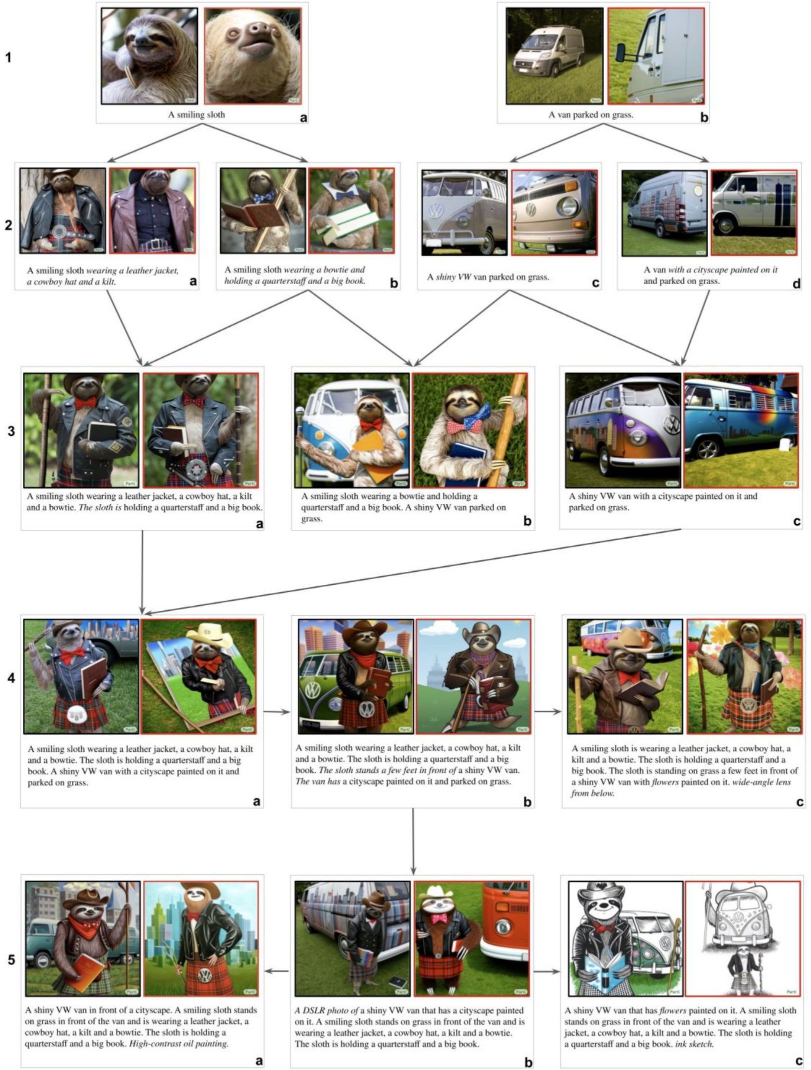
A robot with a black visor and the number 42 on its chest. It stands proudly in front of an F1 race car. The sun is setting on a cityscape in the background. wide-angle view. comic book illustration.

Growing a cherry tree

- Most outputs which are shared are the cherries – the better outputs.

- Usually, complex prompts and images are the output of a process of interactions that probe a model's capabilities. (See process visualization on right, Fig 14 of Parti paper.)

- The fact that compelling outputs can be produced at all is a huge statement – but it also gives an unrealistic impression of what models can deliver in general use.



1a: A smiling sloth

1b: A van parked on grass.

2a: A smiling sloth *wearing a leather jacket, a cowboy hat and a kilt.*

2b: A smiling sloth *wearing a bowtie and holding a quarterstaff and a big book.*

2c: A *shiny VW* van parked on grass.

2d: A van *with a cityscape painted on it* and parked on grass.

3a: A smiling sloth wearing a leather jacket, a cowboy hat, a kilt and a bowtie. *The sloth is* holding a quarterstaff and a big book.

3b: A smiling sloth wearing a bowtie and holding a quarterstaff and a big book. A shiny VW van parked on grass.

3c: A shiny VW van with a cityscape painted on it and parked on grass.

4a: A smiling sloth wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. A shiny VW van with a cityscape painted on it and parked on grass.

4b: A smiling sloth wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. *The sloth stands a few feet in front of* a shiny VW van. *The van has* a cityscape painted on it and parked on grass.

4c: A smiling sloth is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. The sloth is standing on grass a few feet in front of a shiny VW van with *flowers* painted on it. *wide-angle lens from below.*

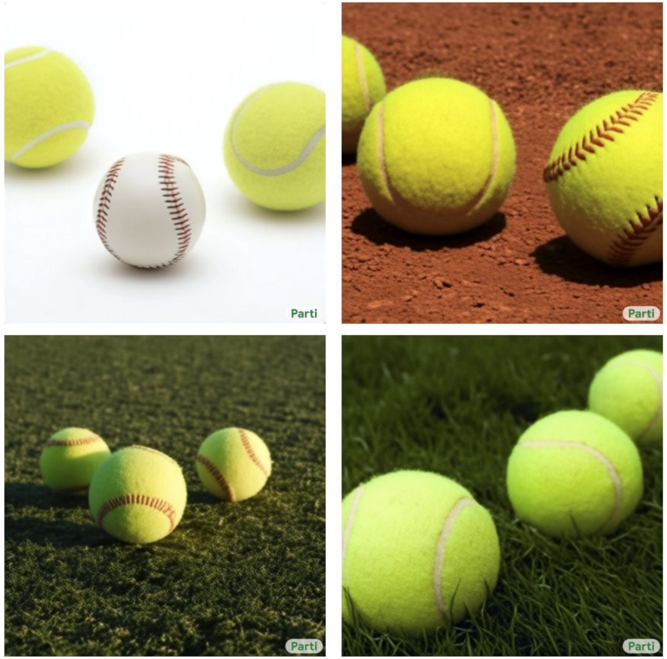
5a: A shiny VW van in front of a cityscape. A smiling sloth stands on grass in front of the van and is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. *High-contrast oil painting.*

5b: *A DSLR photo of* a shiny VW van that has a cityscape painted on it. A smiling sloth stands on grass in front of the van and is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book.

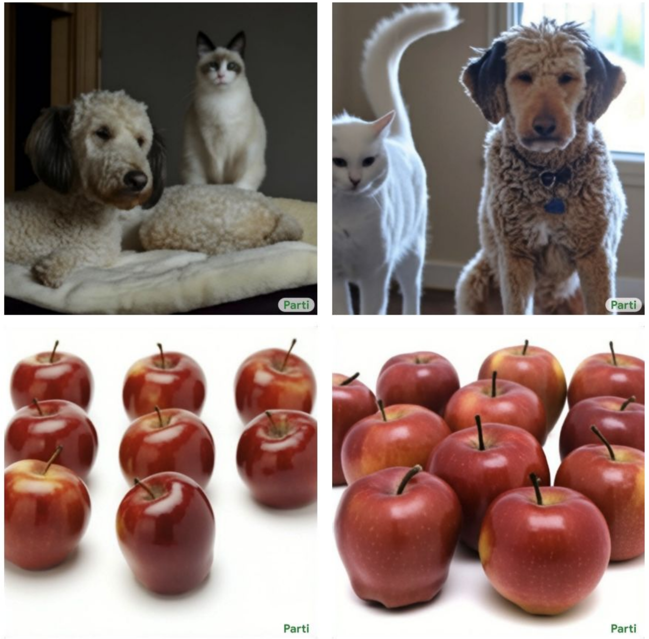
5c: A shiny VW van that has *flowers* painted on it. A smiling sloth stands on grass in front of the van and is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The sloth is holding a quarterstaff and a big book. *ink sketch.*

Limitations (Parti examples)

- Color bleeding
- Feature merging across multiple entities
- Counting
- Incorrect spatial relations
- Omission or hallucination of details



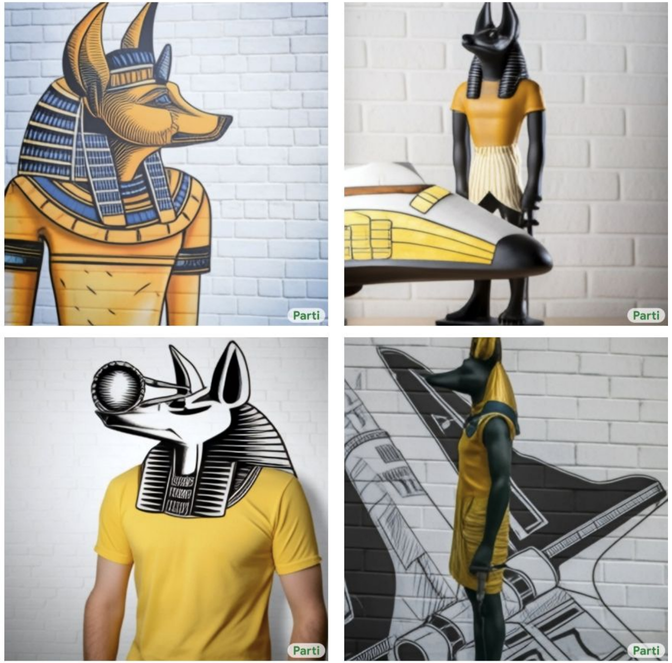
A. Four images generated in the same batch for the prompt *two baseballs to the left of three tennis balls*. **Failures**: color bleeding (b); feature merging (b,c); counting (a-d); spatial relations (a-d). (b-d) also include (arguably reasonable) hallucination of ground details such as gravel and grass.



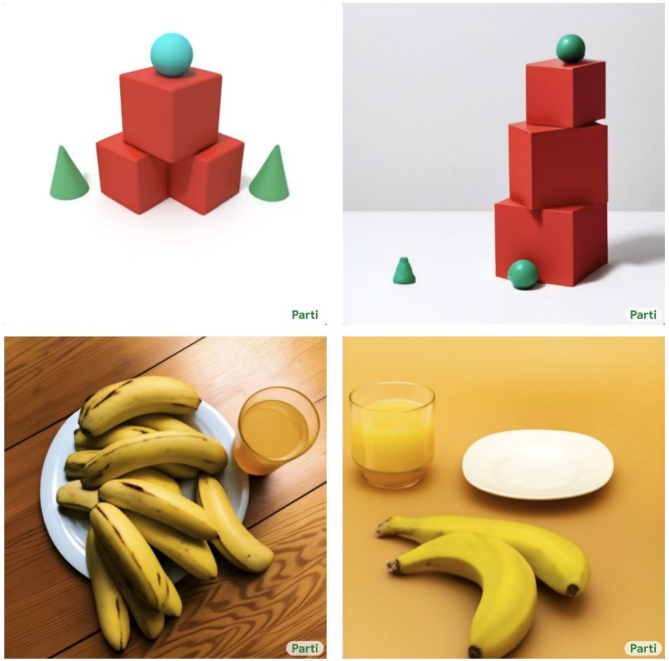
E. (a,b) Two images generated in the same batch for *a cream colored labradoodle next to a white cat with black-tipped ears*. (c,d) Two images generated in the same batch for *ten red apples*. **Failures**: hard to disentangle specific features assigned to multiple entities in the same description (a,b); incorrect count of 8 (a) and 11 (b). (Note that some correctly had ten apples.)

Limitations (Parti examples)

- Ignoring negation and mentioned absence.
- Incorrect attribute-entity binding
- Incorrect spatial relations
- Media blending
- Much more (see the Parti paper)



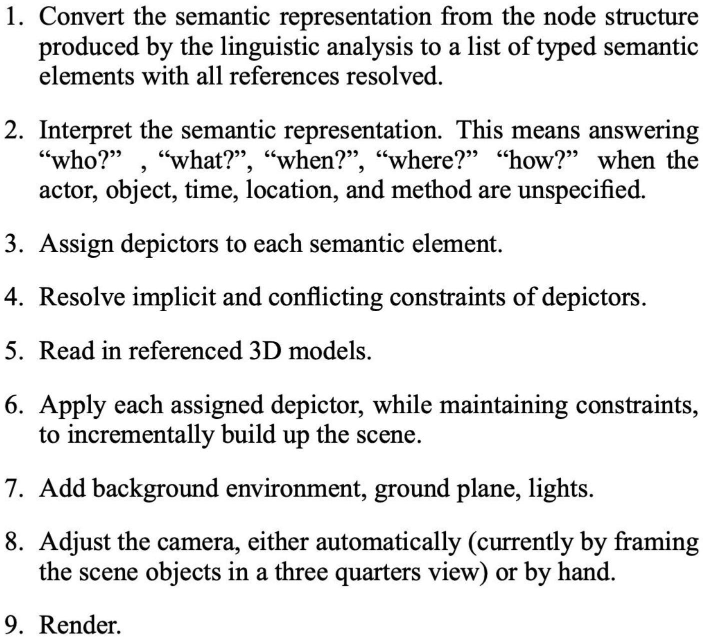
D. Four images generated in the same batch for the prompt *A portrait of a statue of Anubis with a crown and wearing a yellow t-shirt that has a space shuttle drawn on it. A white brick wall is in the background.* **Failures**: color bleeding (a,d); incorrect visual aspect (a,b,d); (unspecified) media blending (c); displaced positioning (b,d); missing details (a,c).



F. (a, b) Two images in the same batch for the prompt *a stack of three red cubes with a blue sphere on the right and two green cones on the left*. (c, d) Two images in the same batch for the prompt *a plate that has no bananas on it. there is a glass without orange juice next to it*. **Failures**: Incorrect relative positioning of objects (a,b,d). Incorrect coloring-to-attribute association (b). Hallucination (of objects specifically mentioned as absent) (c, d).

WordsEye (Coyne and Sproat 2001)

Text-to-image system based on symbolic representations and 3D rendering engine.
Supports a limited range of concepts and styles, but with greater precision than modern systems.



1. Convert the semantic representation from the node structure produced by the linguistic analysis to a list of typed semantic elements with all references resolved.
2. Interpret the semantic representation. This means answering “who?” , “what?”, “when?”, “where?” “how?” when the actor, object, time, location, and method are unspecified.
3. Assign depictors to each semantic element.
4. Resolve implicit and conflicting constraints of depictors.
5. Read in referenced 3D models.
6. Apply each assigned depictor, while maintaining constraints, to incrementally build up the scene.
7. Add background environment, ground plane, lights.
8. Adjust the camera, either automatically (currently by framing the scene objects in a three quarters view) or by hand.
9. Render.

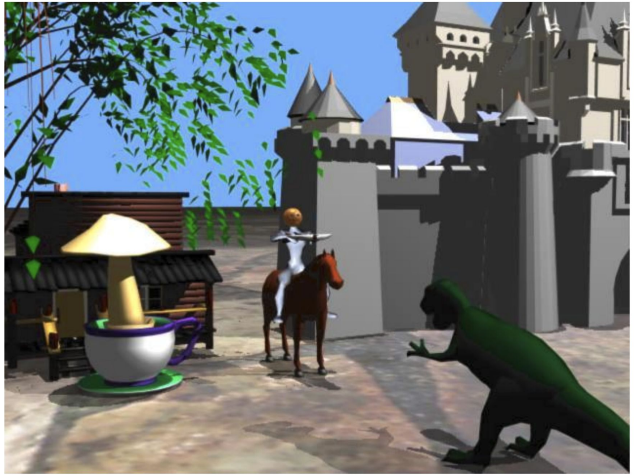


Figure 1: *John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.*

WordsEye: symbolic semantic representations

Hand-built knowledge and compositional construction of representations given prompt.

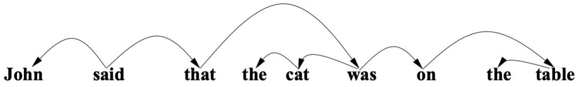
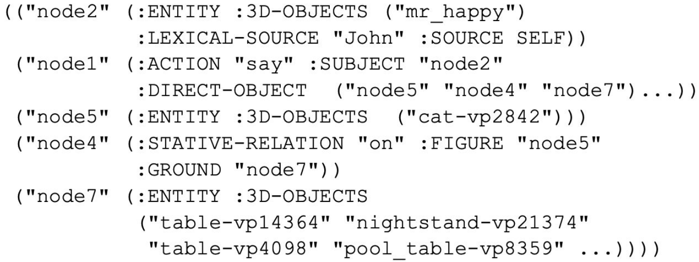
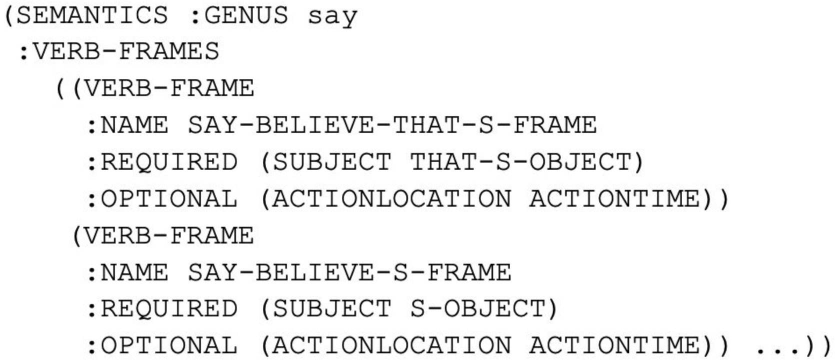


Figure 2: Dependency structure for *John said that the cat was on the table.*.



```
(("node2" (:ENTITY :3D-OBJECTS ("mr_happy")
           :LEXICAL-SOURCE "John" :SOURCE SELF))
 ("node1" (:ACTION "say" :SUBJECT "node2"
           :DIRECT-OBJECT  ("node5" "node4" "node7")...))
 ("node5" (:ENTITY :3D-OBJECTS  ("cat-vp2842")))
 ("node4" (:STATIVE-RELATION "on" :FIGURE "node5"
           :GROUND "node7"))
 ("node7" (:ENTITY :3D-OBJECTS
           ("table-vp14364" "nightstand-vp21374"
            "table-vp4098" "pool_table-vp8359" ...))))
```

Figure 3: Semantic representation for *John said that the cat was on the table.*



```
(SEMANTICS :GENUS say
 :VERB-FRAMES
   ((VERB-FRAME
     :NAME SAY-BELIEVE-THAT-S-FRAME
     :REQUIRED (SUBJECT THAT-S-OBJECT)
     :OPTIONAL (ACTIONLOCATION ACTIONTIME))
    (VERB-FRAME
     :NAME SAY-BELIEVE-S-FRAME
     :REQUIRED (SUBJECT S-OBJECT)
     :OPTIONAL (ACTIONLOCATION ACTIONTIME)) ...))
```

Figure 4: Semantic entry for *say*.

WordsEye: object database w/ spatial relations and affordances

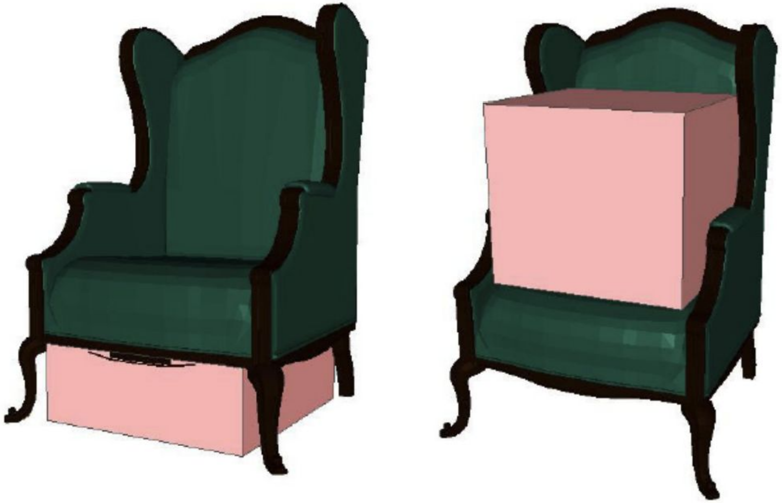


Figure 5: Spatial tag for "canopy area", indicated by the box under the lefthand chair; and "top surface", indicated by the box on the righthand chair.

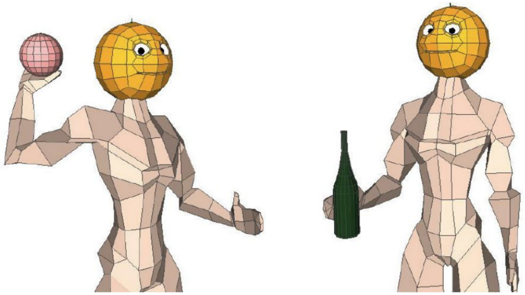


Figure 10: "Throw small object" pose and "hold wine bottle" grip.

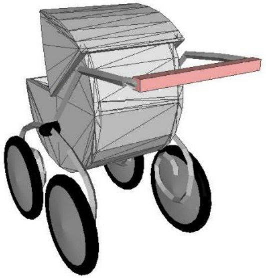


Figure 12: Spatial tag for "push handle" of baby carriage, indicated by the box around the handle.

WordsEye: explicit control over size and placement

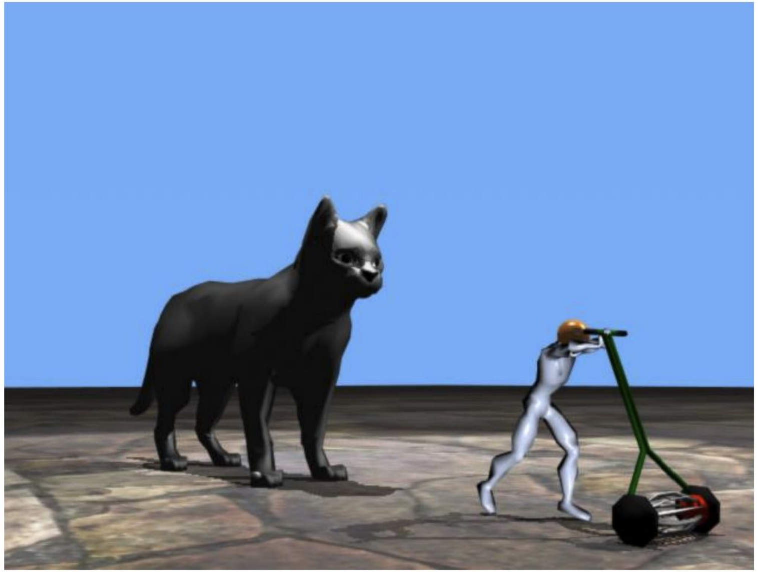
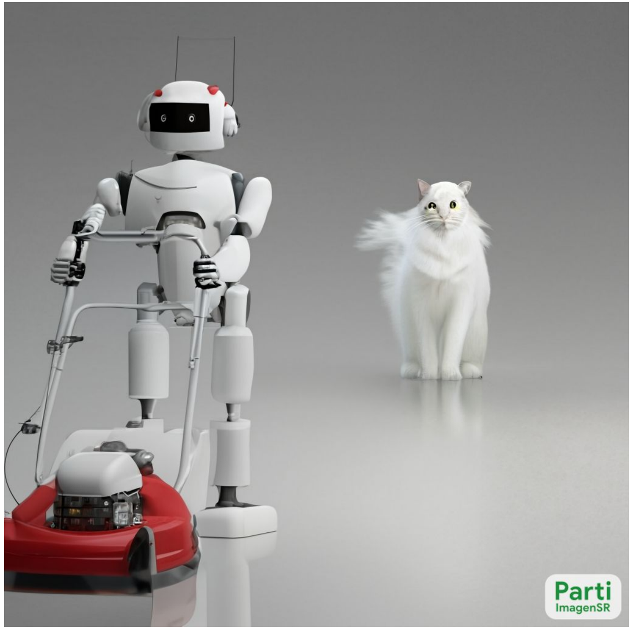


Figure 13: *The lawn mower is 5 feet tall. John pushes the lawn mower. The cat is 5 feet behind John. The cat is 10 feet tall.*



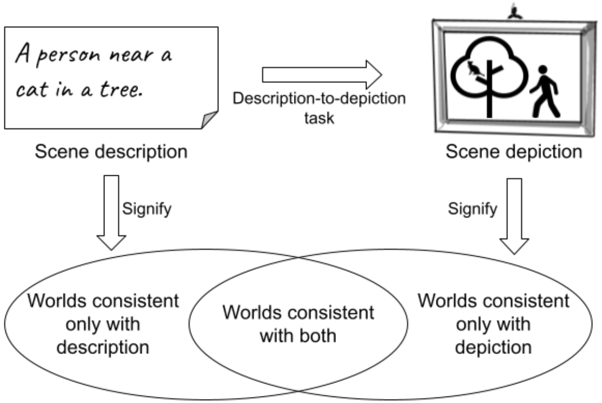
Parti generally fails (given this related prompt)
A robot pushes a lawn mower that is five-foot tall. A 10-foot-tall cat is standing 5 feet behind the robot. 3D computer graphics illustration.

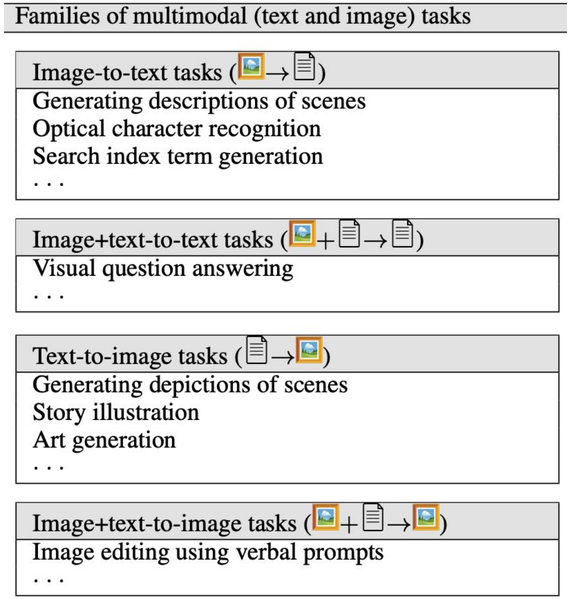
Key question: How to get this sort of control along with the broader capabilities of modern text-to-image systems?

Text-to-image generation needs greater clarity [1]

- Text-to-image itself is not itself a task, but rather encompasses many related tasks.

- Underspecification in both linguistic and visual representations must be attended to both for quality and responsibility.

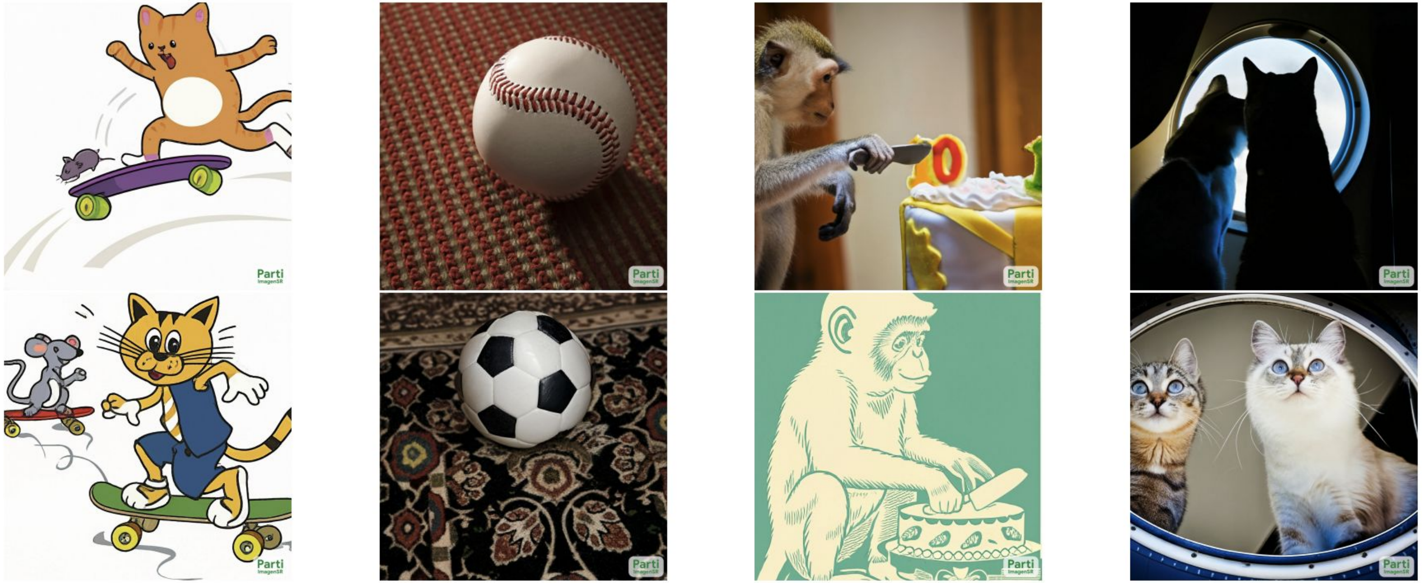

A person near a cat in a tree.
Description-to-depiction task
Scene description
Scene depiction
Signify
Signify
Worlds consistent only with description
Worlds consistent with both
Worlds consistent only with depiction



Families of multimodal (text and image) tasks
Image-to-text tasks (🖼→📄)
Generating descriptions of scenes Optical character recognition Search index term generation . . .
Image+text-to-text tasks (🖼+📄→📄)
Visual question answering . . .
Text-to-image tasks (📄→🖼)
Generating depictions of scenes Story illustration Art generation . . .
Image+text-to-image tasks (🖼+📄→🖼)
Image editing using verbal prompts . . .

[1] Hutchinson, Baldridge and Prabhakaran (2022). Underspecification in Scene Description-to-Depiction Tasks.

Multimodal underspecification → many kinds of ambiguity



Syntactic

(a) Outputs for “A cat chasing a mouse on a skateboard.” The number of boards and which animal is on any given board is ambiguous.

Details

(b) Outputs for “A ball on a rug.” The types and visual details of balls and rugs are unspecified.

Style

(c) Outputs for “A monkey cutting a cake.” The cutting instrument is unspecified, as is the style.

Perspective

(d) Outputs for “Two cats looking out of a space shuttle window. DSLR photograph.” Perspective is unspecified.

Images generated by Parti + Imagen Super Resolution.

Risks

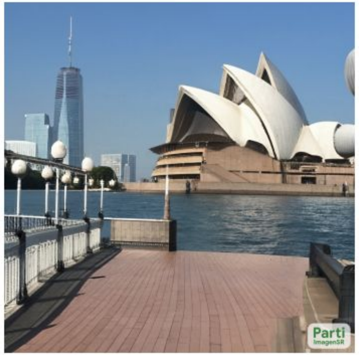
Bias

(a) Outputs for "Wedding attire displayed on a mannequin" may show gender and Western cultural biases.

Taboos and offense

(b) Outputs for "Graffiti on the New York Public library. DSLR photo." might cause offence to bibliophiles.

Misinformation



(c) Outputs for "A photo of a famous city with opera house" may spread misinformation.

Safety

(d) Outputs for "A photo of a non-venomous Australian spider" may have safety risks for animal lovers.

Task dependence: given the goals and scope of a particular deployed system (e.g. generating drawings for children, ideas for room decoration, or artwork), such risks may or may not be a problem.

Dealing with input ambiguity

- All linguistics inputs are ambiguous, *because language*. The question is how a system responds to and reflects those ambiguities.
- Two approaches:
    - *Ambiguity In, Ambiguity Out* (**AIAO**): outputs attempt to retain visual ambiguity, e.g. by using stick figures or blurring.
    - *Ambiguity In, Diversity Out* (**AIDO**): multiple outputs capture a range of specific depictions that provide visual ambiguity in aggregate.
- To implement either of these entails having the means to represent and differentiate input ambiguities.
    - This is standard in (brittle) “old school” language representations (such as logical forms), but requires more development for current neural models.
- Such design choices are generally important, and especially matter for Responsible AI, including fairness and diversity.

Clarifying tasks and capabilities

- When collaborating to create comics, a writer must understand the style and capability of the artist: the same should be true for human-machine text-to-image collaboration.
- Similarly: developers of multimodal systems should aim to understand and communicate how they may be directed via language and the range of output forms and styles they support.
  - Manage user expectations and ability to interpret system behaviors.
  - Helps mitigate risks of misuse.
- Understanding and characterizing visual capabilities requires engaging with experts in visual disciplines, including photographers, artists, designers and curators.
- This interacts with training and test data, and the kinds of semantic and pragmatic relationships that hold between linguistic and visual pairings.

Broadening evaluations

- FID on MS-COCO should become the MNIST for text-to-image generation.
- We need a suite of evaluations that test model generalization and a broader suite of task-specific applications.
  - New benchmarks testing a range of capabilities (e.g. DALL-Eval)
  - New/adapted datasets with longer, more challenging descriptions or targeting specific styles and/or applications.
  - New types of human judgments (timing or task based, A/B tests).
  - Better automatic evaluations (that correlate with human judgments).
  - Evaluations that cover multiple generations for the same prompt.
  - Datasets that explicitly evaluate responsibility matters (fairness, bias, and beyond).
- Surprisingly, anecdotal evaluations are still quite valuable as existence proofs of model capabilities.
  - BUT: we need to take care with cherry picking and setting expectations.

Beyond single interaction image generation

- Many models support text-guided image editing (DALL-E 2, Stable Diffusion, Prompt-to-Prompt, Imagic)
- Segmentation and text-guided image generation: Make-A-Scene, SpaText
- Image-and-text guided image generation: DreamBooth, Unitune, Paint by Example
- Text-to-3D: DreamFusion
- Text-to-story generation: StoryDALL-E
- Text-to-video generation: Make-a-Video, Imagen Video, Phenaki

- And much more… this whole space is moving incredibly fast and is hard to keep up with!

Final note

- With the fast pace of development, **it's important not to forget the text in text-to-image:** and working to better ensure models can respond to and accurately reflect textual descriptions (and ambiguity)!
- We need to explore:
  - Improved language representations and the means for visual components to exploit them.
  - Retrieval-based methods to address scale and adaptation of concepts. (e.g. RE-Imagen)
  - Entity representations and coherence/persistence of visual appearance (over multiple generations, across multiple panels, or in an extended video).
  - More benchmarks that test specific language capabilities, such as DALL-Eval (spatial relations) and Winoground (word order).
  - And more!

THANK YOU!
Parti
ImagenSR

"THANK YOU!" written above a wombat giving a thumbs up outdoors. DSLR photo.

And many thanks to so many others for their help in making Parti possible!

We would like to thank Elizabeth Adkison, Fred Alcober, Tania Bedrax-Weiss, Krishna Bharat, Nicole Brichtova, Yuan Cao, William Chan, Zhifeng Chen, Eli Collins, Claire Cui, Andrew Dai, Jeff Dean, Emily Denton, Toju Duke, Dumitru Erhan, Brian Gabriel, Zoubin Ghahramani, Jonathan Ho, Michael Jones, Sarah Laszlo, Quoc Le, Lala Li, Zhen Li, Sara Mahdavi, Kathy Meier-Hellstern, Kevin Murphy, Paul Natsev, Paul Nicholas, Mohammad Norouzi, Ruoming Pang, Niki Parmar, Fernando Pereira, Slav Petrov, Vinodkumar Prabhakaran, Utsav Prabhu, Evan Rapoport, Keran Rong, Negar Rostamzadeh, Chitwan Saharia, Gia Soles, Austin Tarango, Ashish Vaswani, Tao Wang, Tris Warkentin, Austin Waters, Ben Zevenbergen for helpful discussions and guidance, Peter Anderson, Corinna Cortes, Tom Duerig, Douglas Eck, David Ha, Radu Soricut and Rahul Sukthankar for paper review and feedback, Erica Moreira and Victor Gomes for help with resource coordination, Tom Small for designing the Parti watermark, Google ML Data Operations team for collecting human evaluations on our generated images and others in the Google Brain team and Google Research team for support throughout this project.

We would also like to give particular acknowledgments to the Imagen team, especially Mohammad Norouzi, Chitwan Saharia, Jonathan Ho and William Chan, for sharing their near complete results prior to releasing Imagen; their findings on the importance of CF guidance were particularly helpful for the final Parti model. We also thank the Make-a-Scene team, especially Oran Gafni, for helpful discussion on CF-guidance implementation in autoregressive models. We thank the DALL-E 2 authors, especially Aditya Ramesh, for helpful discussion on MS-COCO evaluation. We also thank the DALL-Eval authors, especially Jaemin Cho, for help with reproducing their numbers.