# Imagen, a text-to-image diffusion model

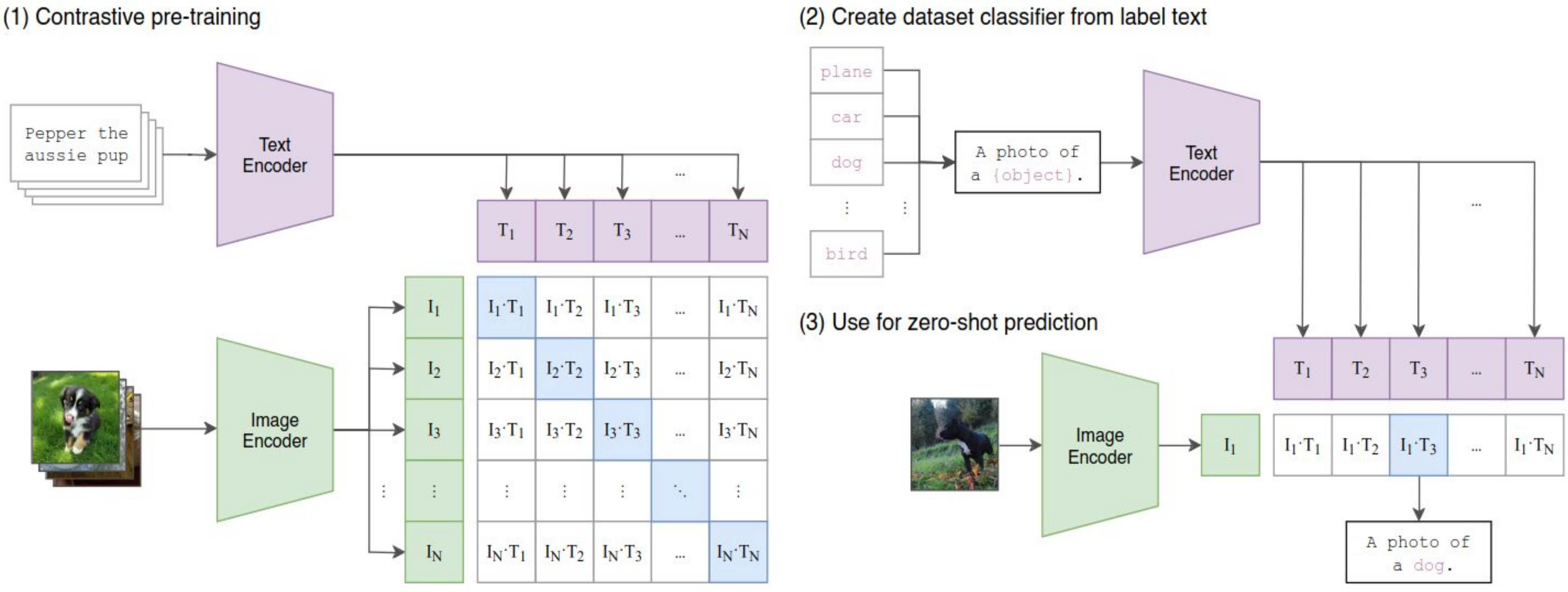
(1) Better semantic text encoding by LLMs

(2) High-fidelity/photorealism by diffusion models

https://imagen.research.google/

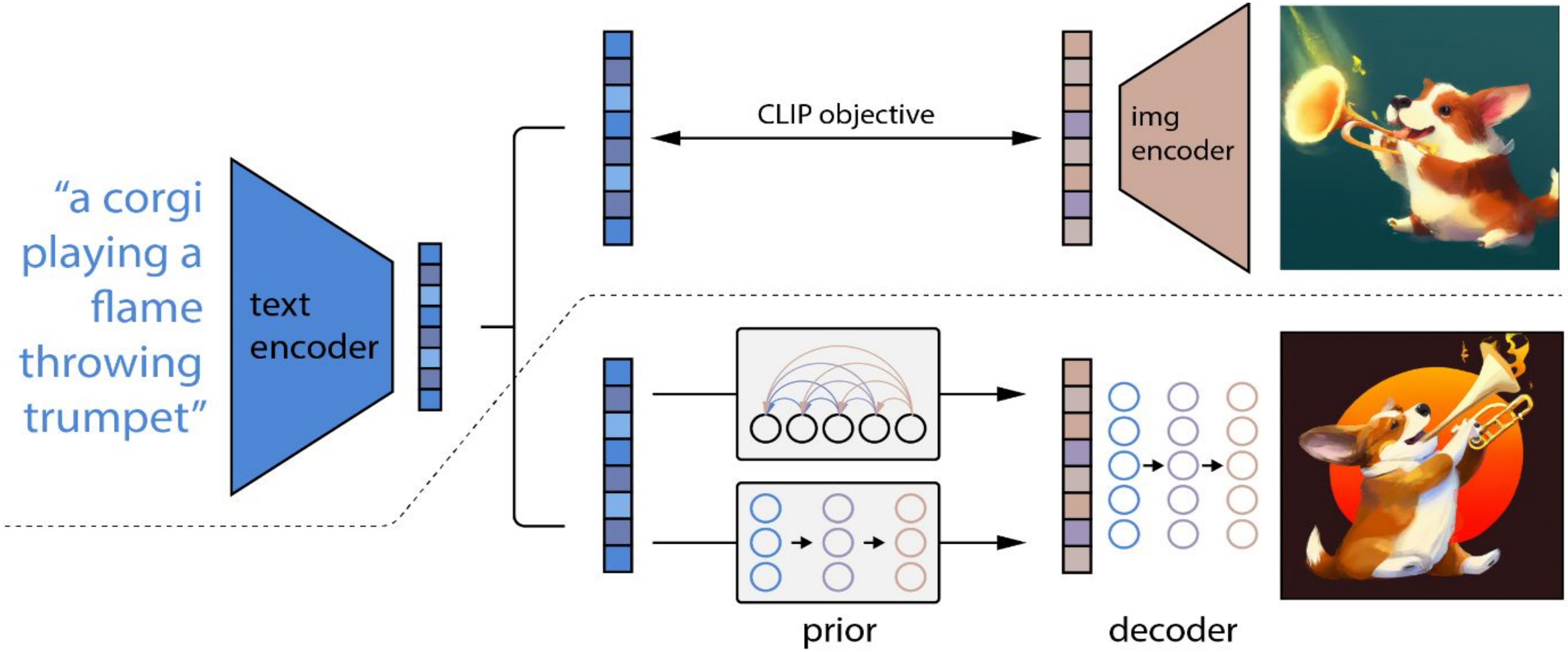
# Background of text-2-image: the CLIP model=ground-truth

- CLIP dataset and pretraining on 400 million pairs -> **CLIP score, zero-shot**



# Background of text-2-image: the CLIP model=ground-truth

In DALL·E 2 (unCLIP), similar idea to VAE



# What does this paper do?

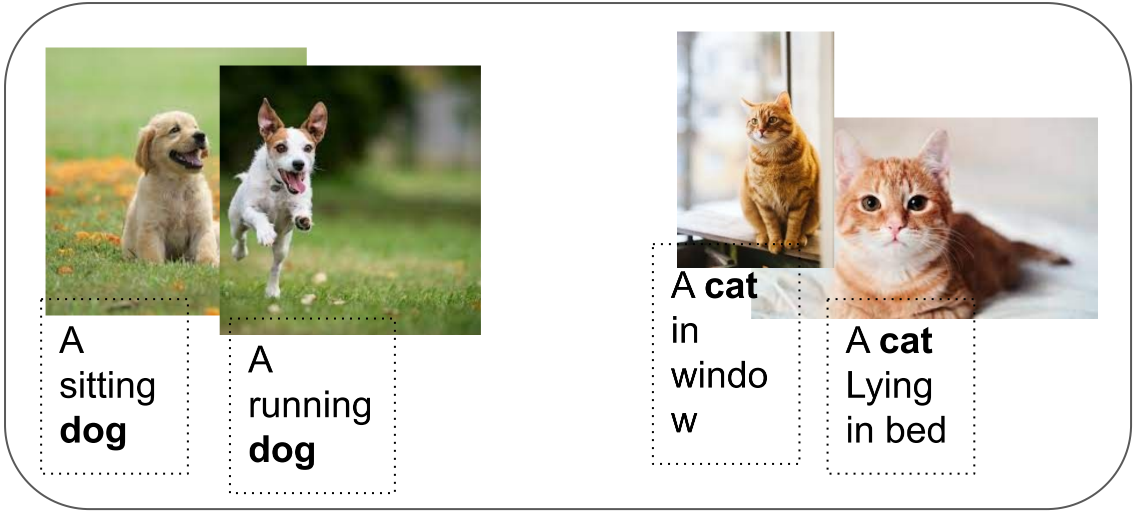
- Imagen (text-to-image)
  - Using **generic pretrained transformer LLMs** on text-only corpora
  - Using diffusion models for high-fidelity photorealism
- Evaluation metrics:
  - Smaller FID score = Higher fidelity/photorealism
  - Higher CLIP score = higher text-image alignment

  - Zero-shot transfer on COCO dataset
  - **Human raters on the new benchmark: DrawBench**

- Concurrent with DALL·E 2, and the idea is also similar, except for the prior model in DALL·E 2.

# 1st novelty and why

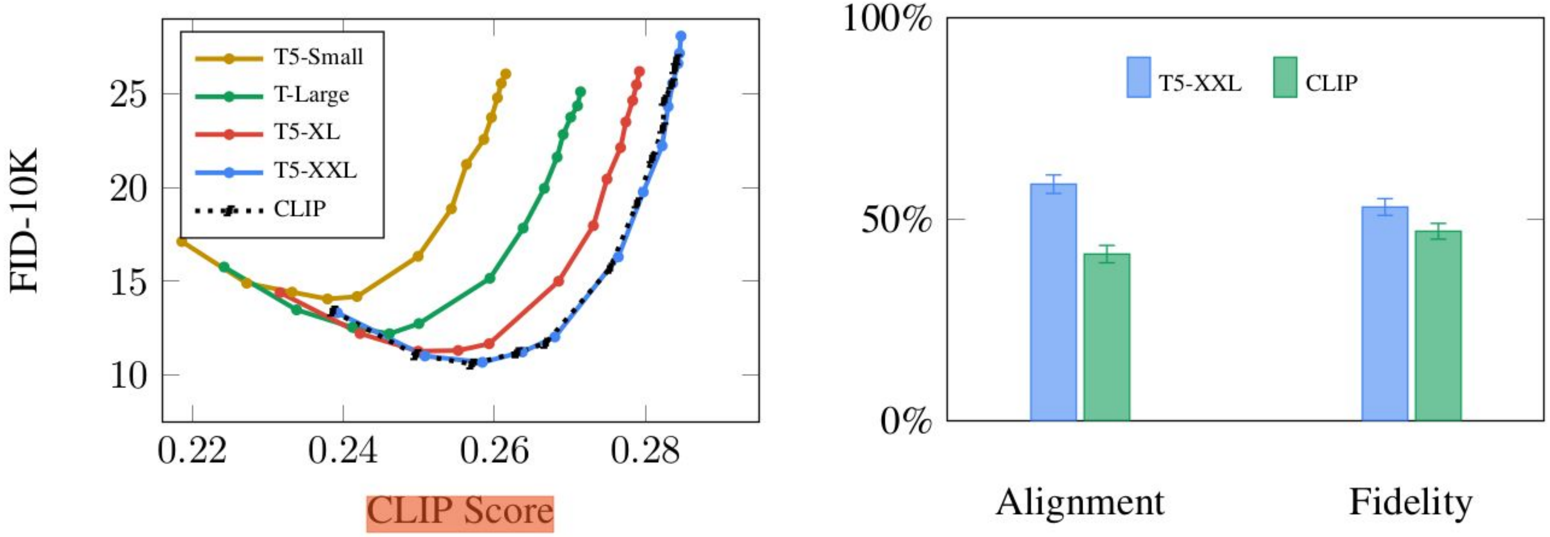
- Prior work:  uses **only image-text data**, for direct **intra-domain** and **inter-domain** learning.
- This work: Large **frozen** LMs trained **only on text data**, is effective enough.
- **Key insights**: we can learn and separate the one domain’s concepts first, then use them as anchors linking to the other domain’s concepts.




If **text concepts** are well understood and **separated in the embedding space**, by LLMs,
->
**fix** the text embedding of concepts,
->
Just needs to **align the visual concepts.**

# 1st novelty and why

- Remark: Large frozen LMs, **is not better** except for the DrawBench dataset.




(a) Pareto curves comparing various text encoders. (b) Comparing T5-XXL and CLIP on DrawBench.

Figure A.5: Comparison between text encoders for text-to-image generation. For Fig. A.5a, we sweep over guidance values of $[1, 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10]$

# 2nd novelty and why

- Dynamic thresholding, (a new diffusion sampling technique), enabling "***large guidance weight samplers***"
  - Significantly better **photorealism**
  - Better image-text **alignment**, especially **when using very large guidance weights**.

(*Personally, this is the <u>core performance trick and most important contribution</u>*)

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) = w\epsilon_\theta(\mathbf{z}_t, \mathbf{c}) + (1 - w)\epsilon_\theta(\mathbf{z}_t). \quad (2)$$

Here, $\epsilon_\theta(\mathbf{z}_t, \mathbf{c})$ and $\epsilon_\theta(\mathbf{z}_t)$ are conditional and unconditional $\epsilon$-predictions, given by $\epsilon_\theta := (\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_\theta)/\sigma_t$, and $w$ is the *guidance weight*. Setting $w = 1$ disables classifier-free guidance, while increasing $w > 1$ strengthens the effect of guidance. Imagen depends critically on classifier-free guidance for effective text conditioning.

training data $\mathbf{x}$, i.e. within $[-1, 1]$

ons to exceed these bounds. This

# Other contributions

Key contributions of the paper include:

1. We discover that large frozen language models trained only on text data are surprisingly very effective text encoders for text-to-image generation, and that scaling the size of frozen text encoder improves sample quality significantly more than scaling the size of image diffusion model.
2. We introduce *dynamic thresholding*, a new diffusion sampling technique to leverage high guidance weights and generating more photorealistic and detailed images than previously possible.
3. We highlight several important diffusion architecture design choices and propose *Efficient U-Net*, a new architecture variant which is simpler, converges faster and is more memory efficient.
4. We achieve a new state-of-the-art COCO FID of 7.27. Human raters find Imagen to be on-par with the reference images in terms of image-text alignment.
5. We introduce DrawBench, a new comprehensive and challenging evaluation benchmark for the text-to-image task. On DrawBench human evaluation, we find Imagen to outperform all other work, including the concurrent work of DALL-E 2 [56].

# Applications in the biomedical science

- Biomedical image-text **pair data modelling**: e.g.,

CXR115_IM-0102-1001

**Impression:**

COPD. No acute pulmonary disease.

**Findings:**

the lungs are clear. there is hyperinflation of the lungs. there is no pleural effusion or pneumothorax. the heart and mediastinum are normal. the skeletal structures are normal.

**Labels:**

hyperinflation; chronic obstructive; copd; pulmonary disease

# Applications in the biomedical science

- **Conditional generative models**

  for molecule design