Synthèse d'images à partir de descriptions en langage naturel

Séminaire TALia du 18 novembre 2022

Computer Science > Computer Vision and Pattern Recognition

arXiv:2103.00020 (cs)

[Submitted on 26 Feb 2021]

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever

https://arxiv.org/abs/2103.00020

Computer Science > Computer Vision and Pattern Recognition

arXiv:2204.06125 (cs)

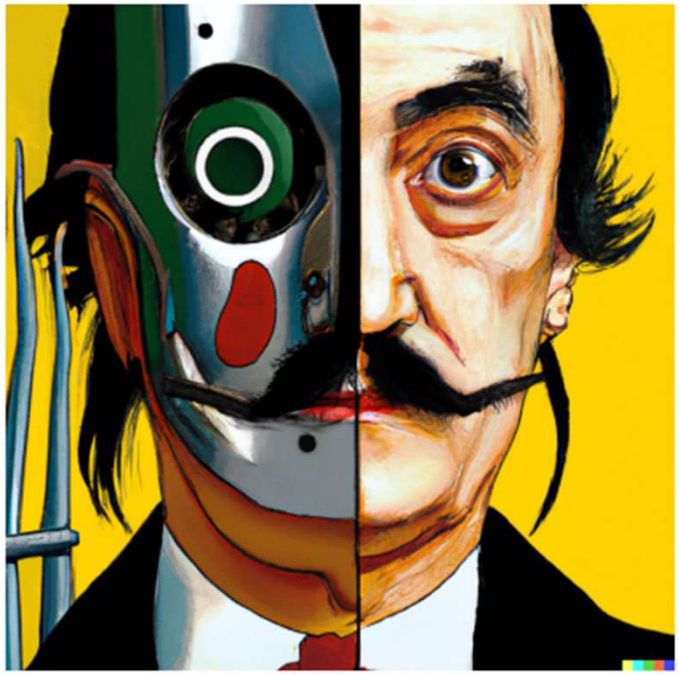
[Submitted on 13 Apr 2022]

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen

https://arxiv.org/abs/2204.06125

Le problème



vibrant portrait painting of Salvador Dalí with a robotic half face

a corgi's head depicted as an explosion of a nebula

a teddy bear on a skateboard in times square

Générer une image à partir d'un *prompt* en langage naturel

Le problème

Un sujet qui suscite beaucoup d'intérêt en 2022

Imagen

Stable Diffusion

Midjourney

DALL·E 2

Motivations :

- De nombreuses applications (édition de photographies, art numérique, aide au design, etc.)
- Une tâche difficile
  - grand écart sémantique entre les domaines
  - forte dimensionnalité de l'espace de sortie structuré

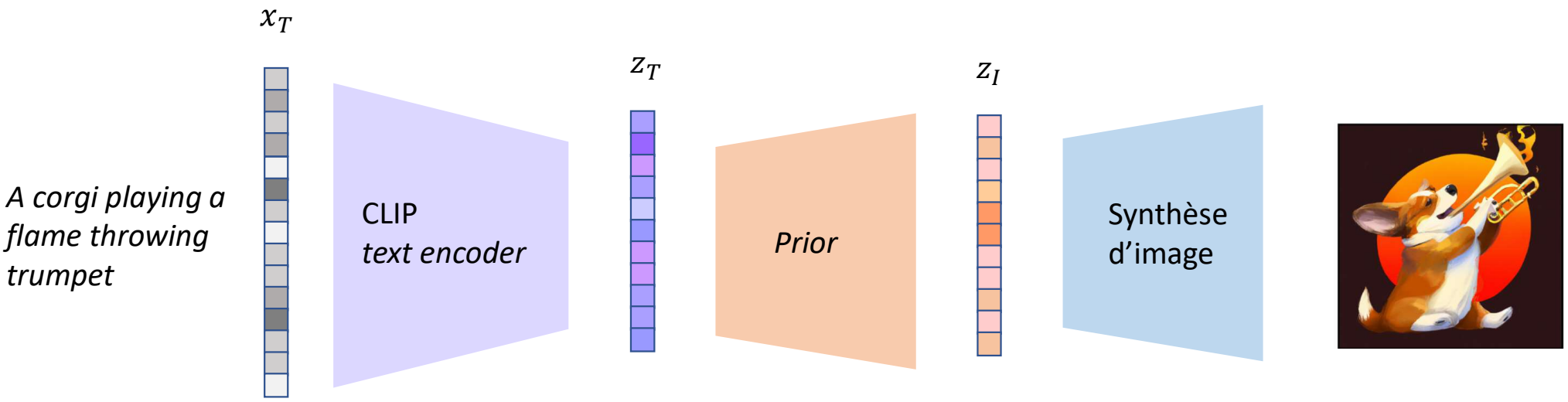
Un sujet qui va au-delà de la synthèse d'images

DALL-E 2 - Overview

A corgi playing a flame throwing trumpet

OpenAI

DALL-E 2 - Overview

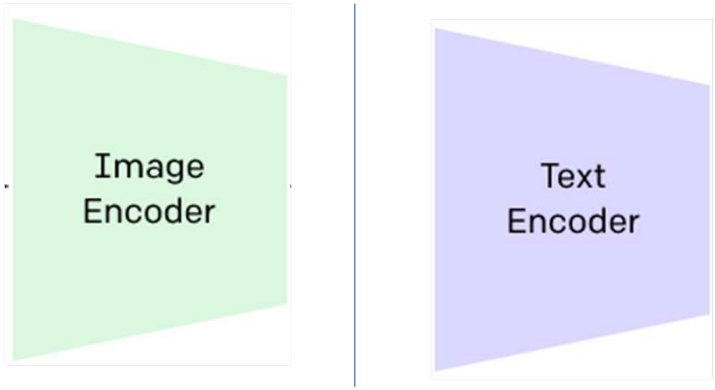


CLIP – *Contrastive Language-Image Pre-Training*

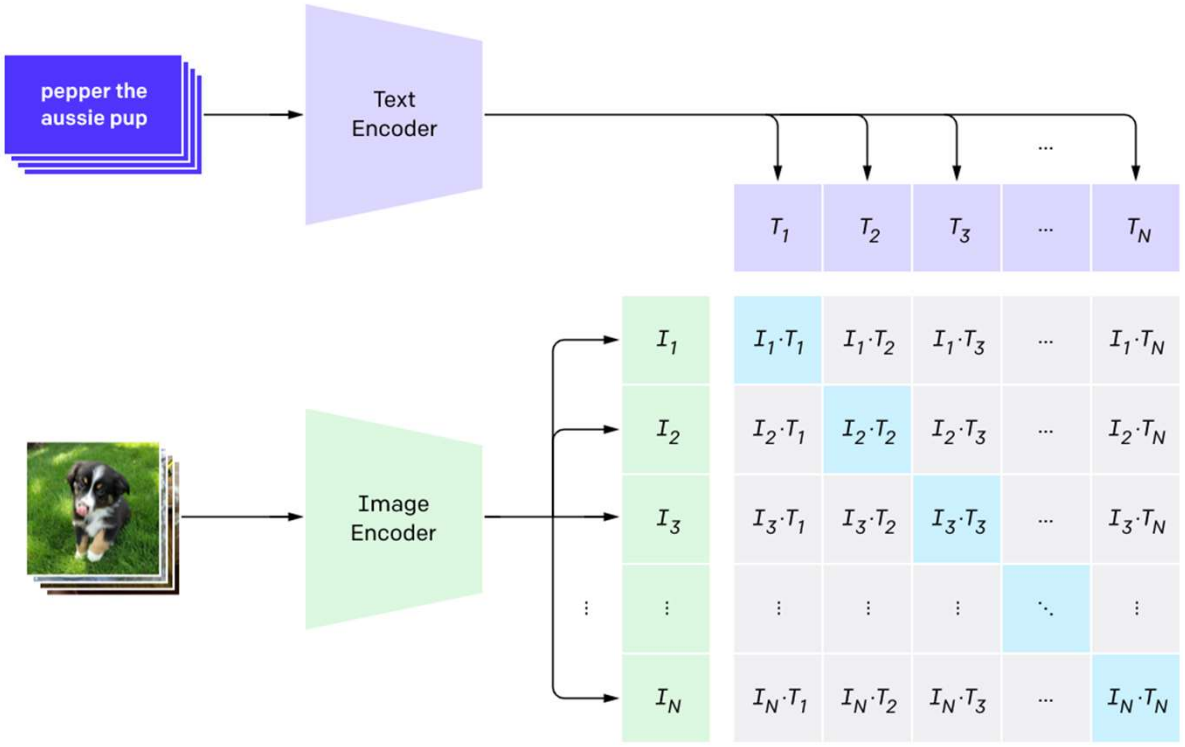
Le cœur de l'approche : réaliser un apprentissage d'éléments de 'perception' supervisé par le langage naturel.

Les avantages :

- Pas besoin de réaliser une tâche fastidieuse de labellisation
- Il ne s'agit pas seulement d'apprendre une représentation mais de faire le lien entre deux sémantiques différentes, de connecter des représentations d'images à des représentations de textes.



CLIP – Objectif contrastif



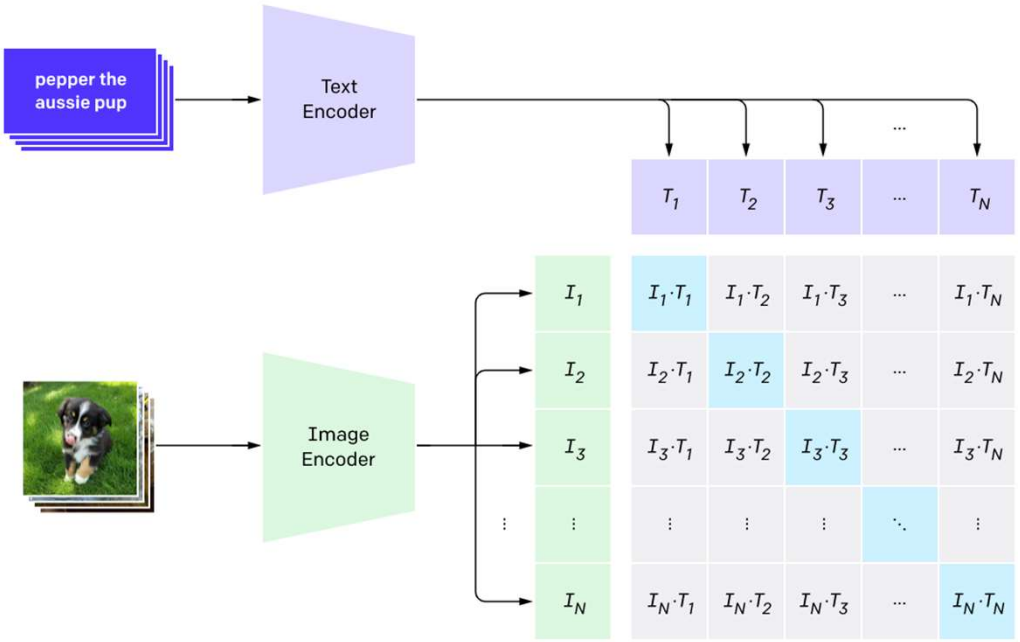
Avec un batch de N paires (image, texte), on peut fabriquer N^2 paires (image, texte).

CLIP est entraîné à prédire lesquelles parmi ces N^2 paires sont des vraies paires (les N paires initiales) et lesquelles sont des fausses paires (les $N^2 - N$ autres paires)

Pour réaliser ceci, 2 encodeurs sont entraînés conjointement à maximiser la similarité cosinus d'embeddings de textes et d'images des N vraies paires et à minimiser celle des $N^2 - N$ paires incorrectes

=> Objectif contrastif

CLIP – Entraînement du modèle



2 encodeurs à entraîner + 2 projections

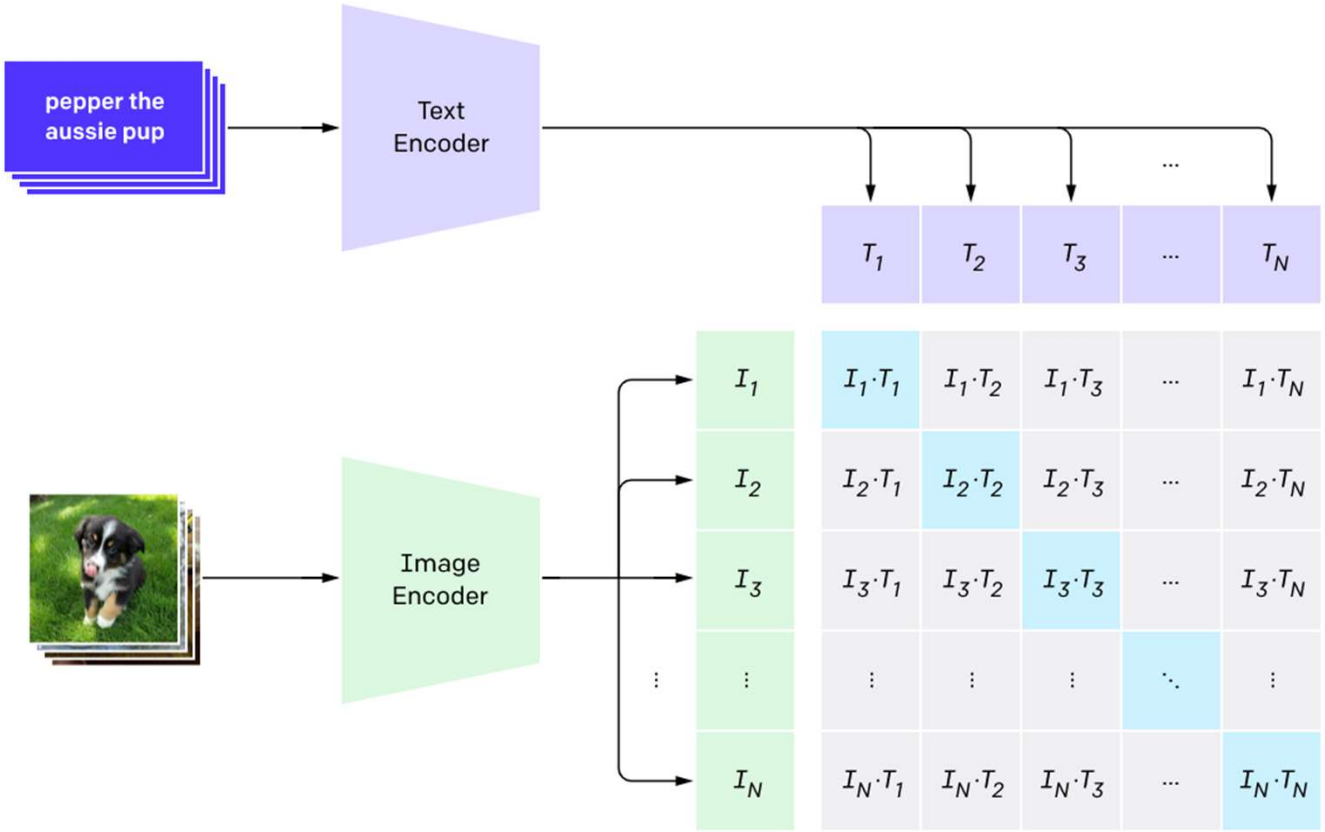
I_i et T_i projections normalisées sur un sous-espace de dimension d_e des représentations des images et des textes issues des encodeurs.

$$p_i^I = softmax(e^\tau I_i . T_1, e^\tau I_i . T_2, \dots, e^\tau I_i . T_N)$$

$$\left(p_i^I\right)_j = \frac{\exp(e^\tau I_i . T_j)}{\sum_{k=1}^{N} \exp(e^\tau I_i . T_k)} \qquad \sum_{k=1}^{N} \left(p_i^I\right)_k = 1$$

$$\mathcal{L}_i^I = -log\left(\left(p_i^I\right)_i\right) - \sum_{\substack{j=1 \\ j \neq i}}^{N} log\left(1 - \left(p_i^I\right)_j\right)$$

CLIP – Entraînement du modèle



$$p_i^I = softmax(e^\tau I_i.T_1, e^\tau I_i.T_2, \dots, e^\tau I_i.T_N)$$

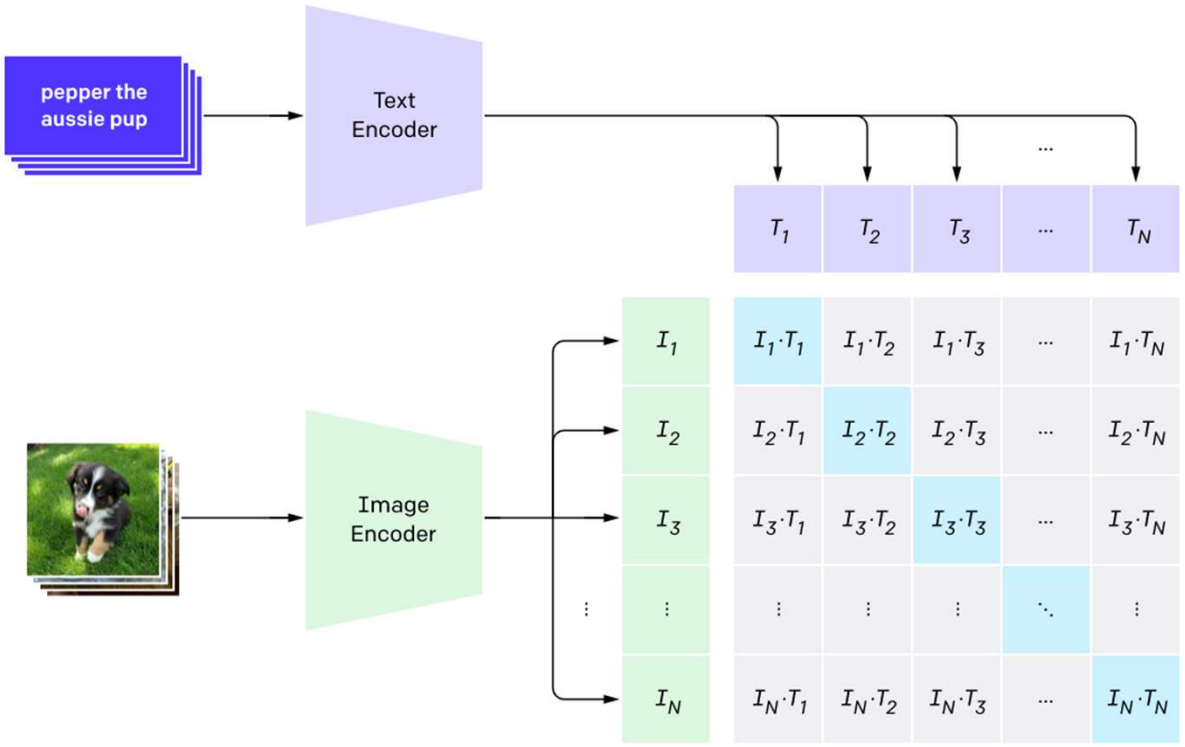
$$p_i^T = softmax(e^\tau I_1.T_i, e^\tau I_2.T_i, \dots, e^\tau I_N.T_i)$$

$$\mathcal{L}_i^I = -log\left(\left(p_i^I\right)_i\right) - \sum_{\substack{j=1 \\ j \neq i}}^{N} log\left(1 - \left(p_i^I\right)_j\right)$$

$$\mathcal{L}_i^T = -log\left(\left(p_i^T\right)_i\right) - \sum_{\substack{j=1 \\ j \neq i}}^{N} log\left(1 - \left(p_i^T\right)_j\right)$$

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}^I + \mathcal{L}^T\right)$$

CLIP – Le corpus de données



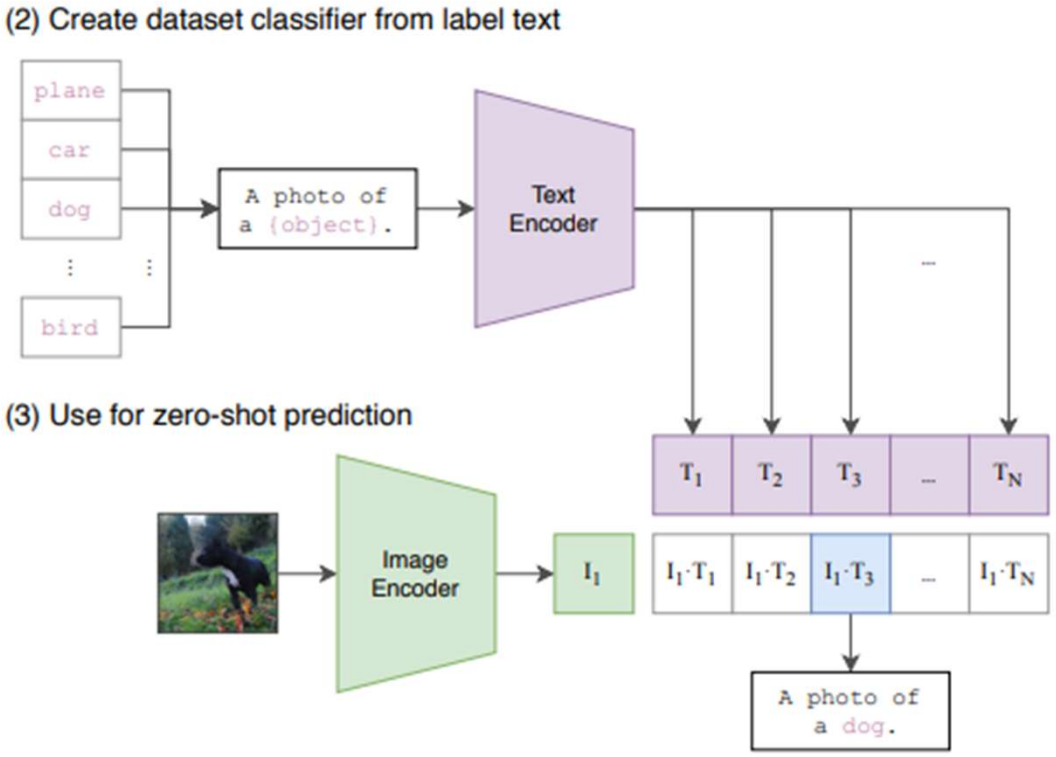
400 000 000 de paires (image, texte) :

- Dont les textes contiennent au moins une parmi 500 000 *queries*
- Equilibré approximativement en incluant au plus 20 000 paires (image, texte) par *query*

Les 500 000 *queries* correspondent :

- Tous les mots qui apparaissent **au moins 100 fois dans Wikipedia anglais** + des bi-grammes (sur une base d'information mutuelle).
- Tous les **titres d'articles Wikipedia** qui génèrent un grand volume de recherche
- Tous **les synsets Wordnet**

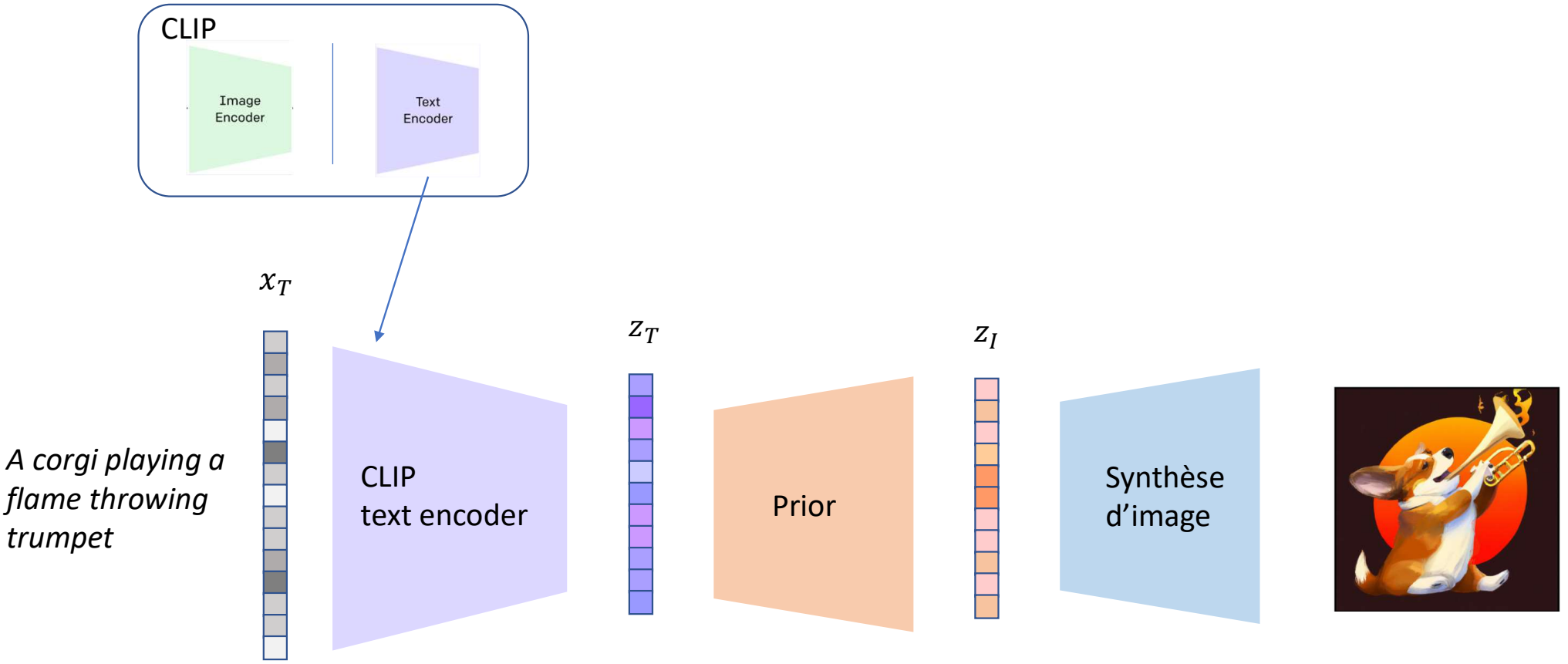
Utilisation de CLIP pour des tâches de classification zero-shot



	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

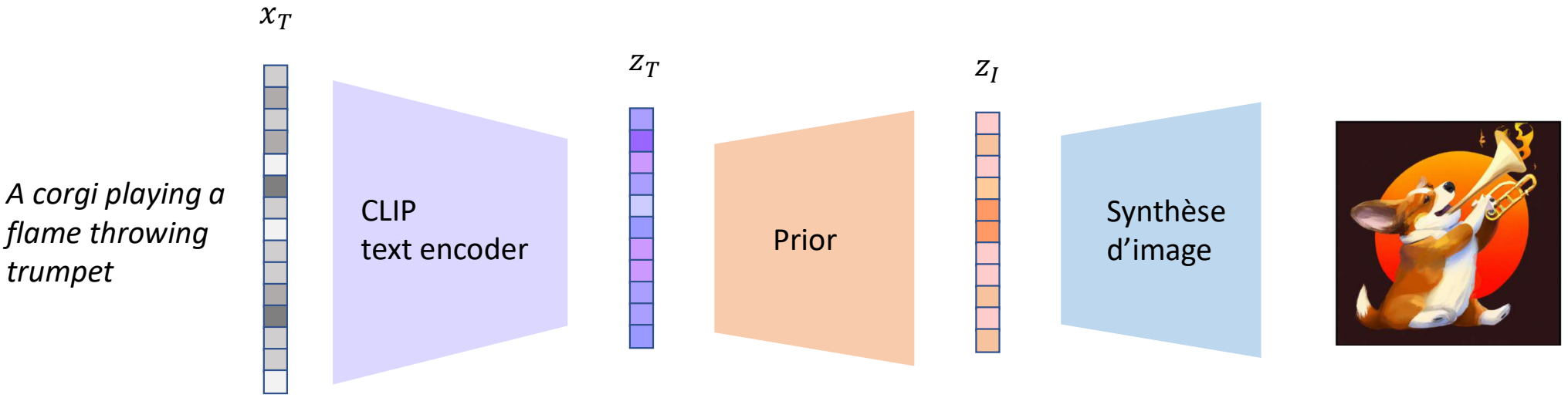
Table 1. Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).

DALL-E 2 – Overview



DALL-E 2 – Le prior

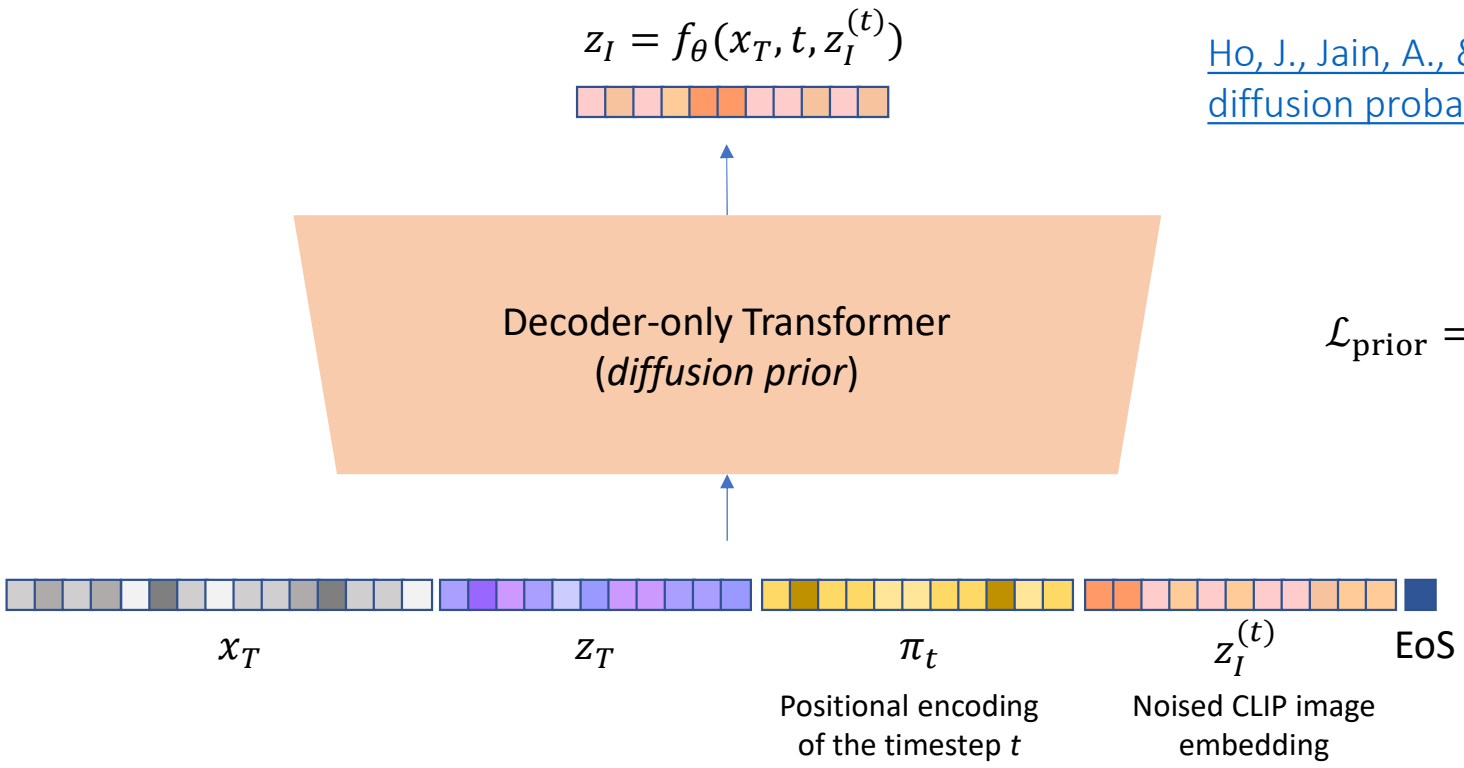
Le *prior* traduit la représentation z_T en une représentation d'image z_I
On souhaite que z_I ait les 'bonnes propriétés' des représentations d'images générées via l'encodeur d'images de CLIP



DALL-E 2 – Génération d'un embedding d'image (3)

https://huggingface.co/blog/annotated-diffusion

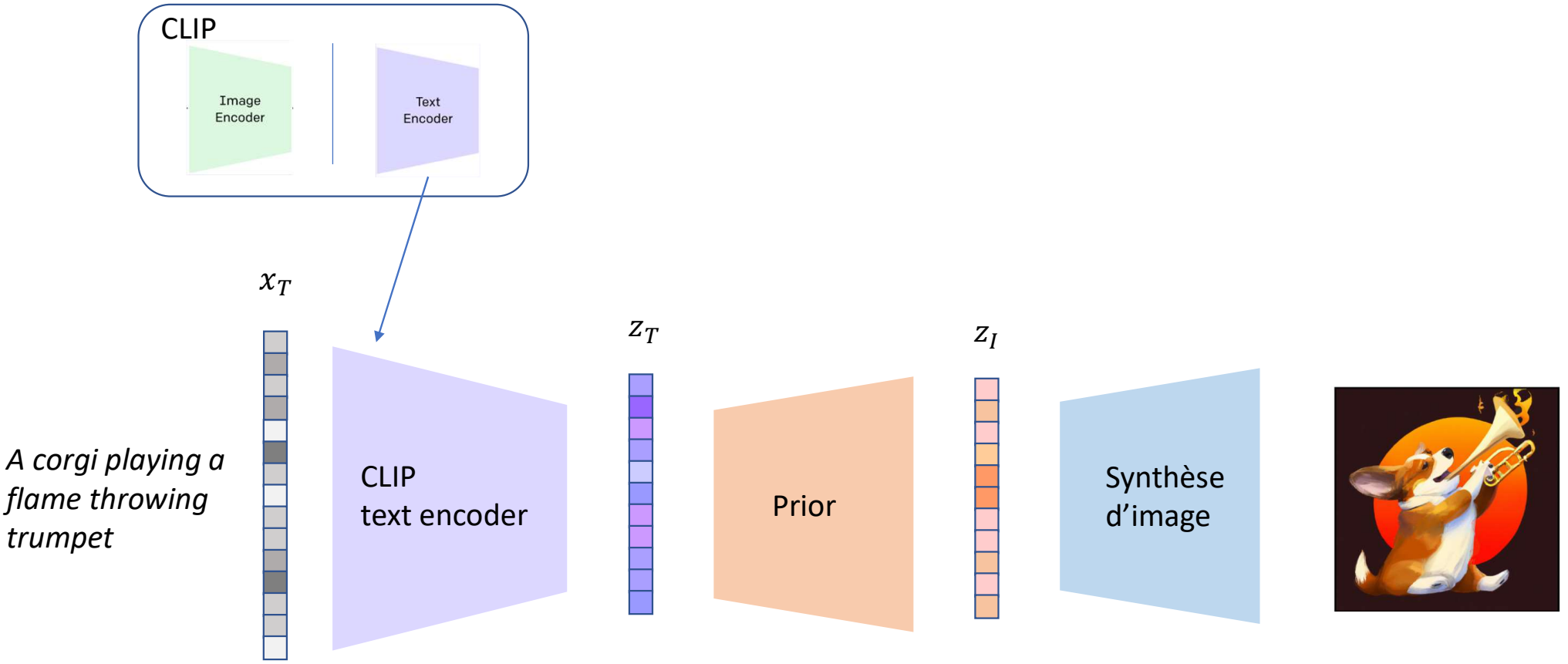
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS 2020*.



$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{t\sim[1,T], z_i^{(t)}\sim q_t} \left\| f_\theta(x_T, t, z_I^{(t)}) - z_I \right\|^2$$

Les auteurs choisissent de ne pas conditionner la génération de z_I par z_T. z_I.
Pour un même z_T, 2 embeddings z_I^1 et z_I^2 sont générés et celui qui minimise la similarité cosinus avec z_T est sélectionné.

DALL-E 2 – Overview



DALL-E 2 – Synthèse de l'image

Decoder-only Transformer

… ou on le met à 0
… ou on utilise z_T

Le décodeur utilisé pour la synthèse d'image est conditionné par :

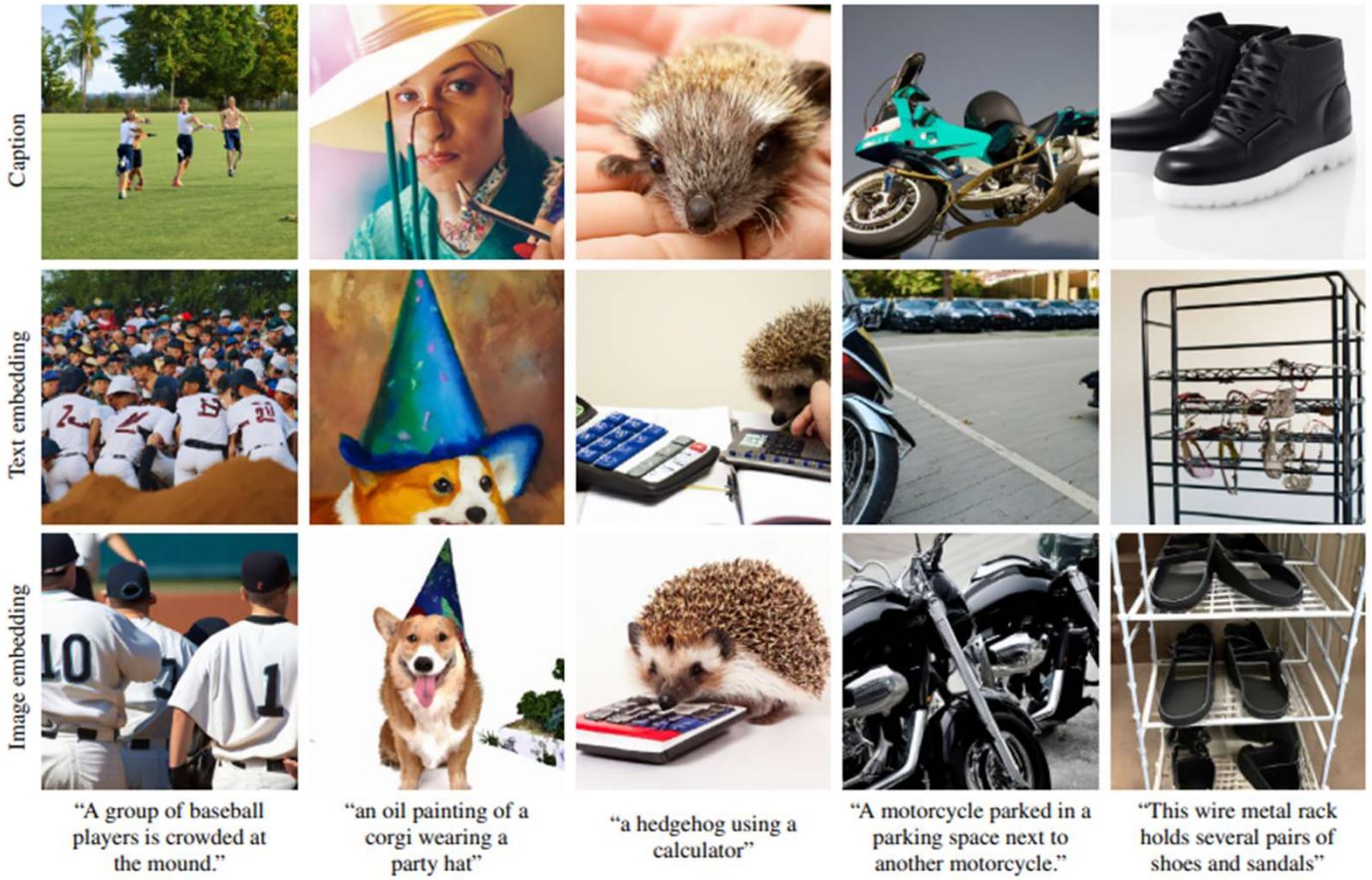
- Un encodage de la description en langage naturel
- L'embedding d'image calculé à l'aide du prior

En phase d'entraînement, l'embedding d'image est *dropped* (mis à zéro) 5% du temps (*classifier-free guidance*)

Le prior n'est pas 'nécessaire' au sens strict

- On peut utiliser uniquement l'encodage de la description
- On peut utiliser l'embedding de texte issu de l'encodeur de CLIP comme embedding d'image

DALL-E 2 – Importance du *prior*



DALL-E 2 – *Text diffs*

concept space de Word2vec :

“queen” ~ “woman” + “king” − “man”

On peut tenter une operation similaire avec les representations d’image de CLIP :

(image of victorian house) + “a modern house” − “a victorian house”

$$
\begin{aligned}
z_{i0} &= f_i(\text{house.png}),\\
z_{t0} &= f_t(\text{“a photo of a victorian house”}),\\
z_{t1} &= f_t(\text{“a photo of a modern house”}), \quad \text{and}\\
z_d &= (z_{t1} - z_{t0})/\|z_{t1} - z_{t0}\|,
\end{aligned}
$$

DALL-E 2 – *Text diffs* (exemples)

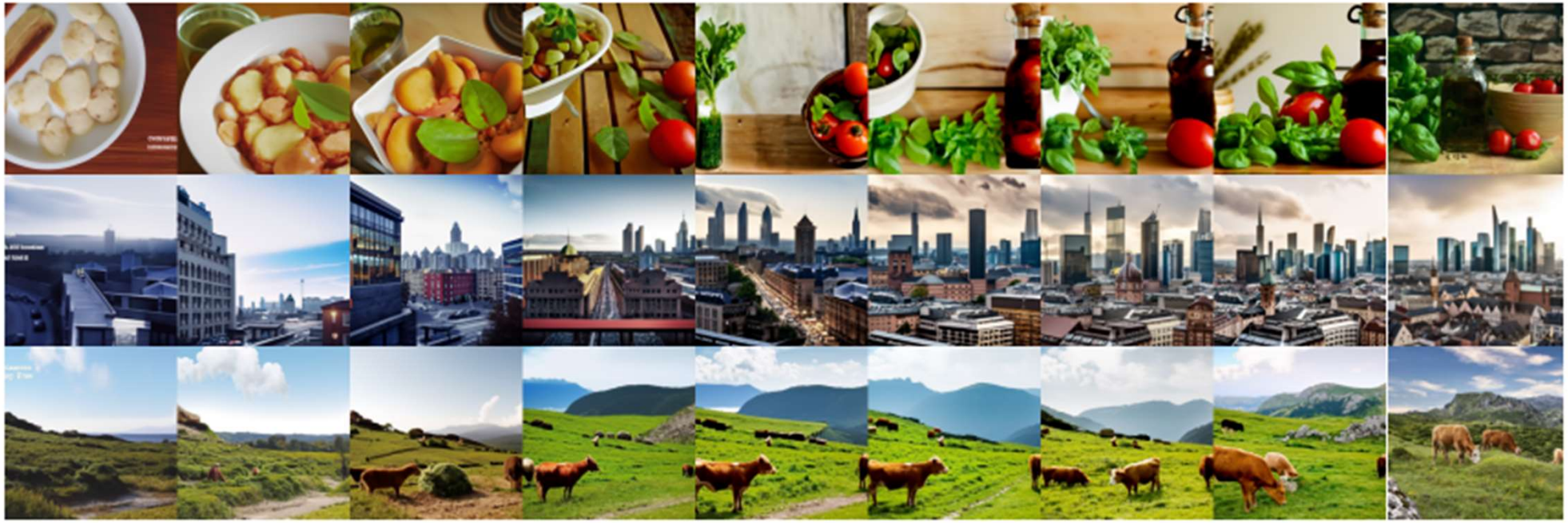
a photo of a cat → an anime drawing of a super saiyan cat, artstation

a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

DALL-E 2 – Analyse en composantes principales



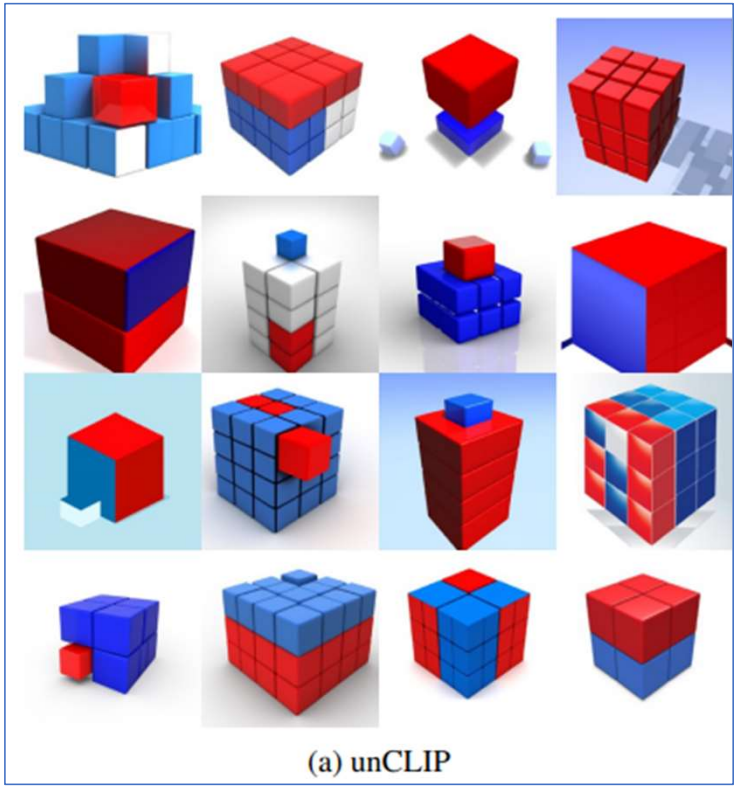
DALL-E 2 – Typographic attacks

Granny Smith: 100%
iPod: 0%
Pizza: 0%

Granny Smith: 0.02%
iPod: 99.98%
Pizza: 0%

Granny Smith: 94.33%
iPod: 0%
Pizza: 5.66%

DALL-E 2 – Limites : composition de l'image



(a) unCLIP

"a red cube on top of a blue cube"

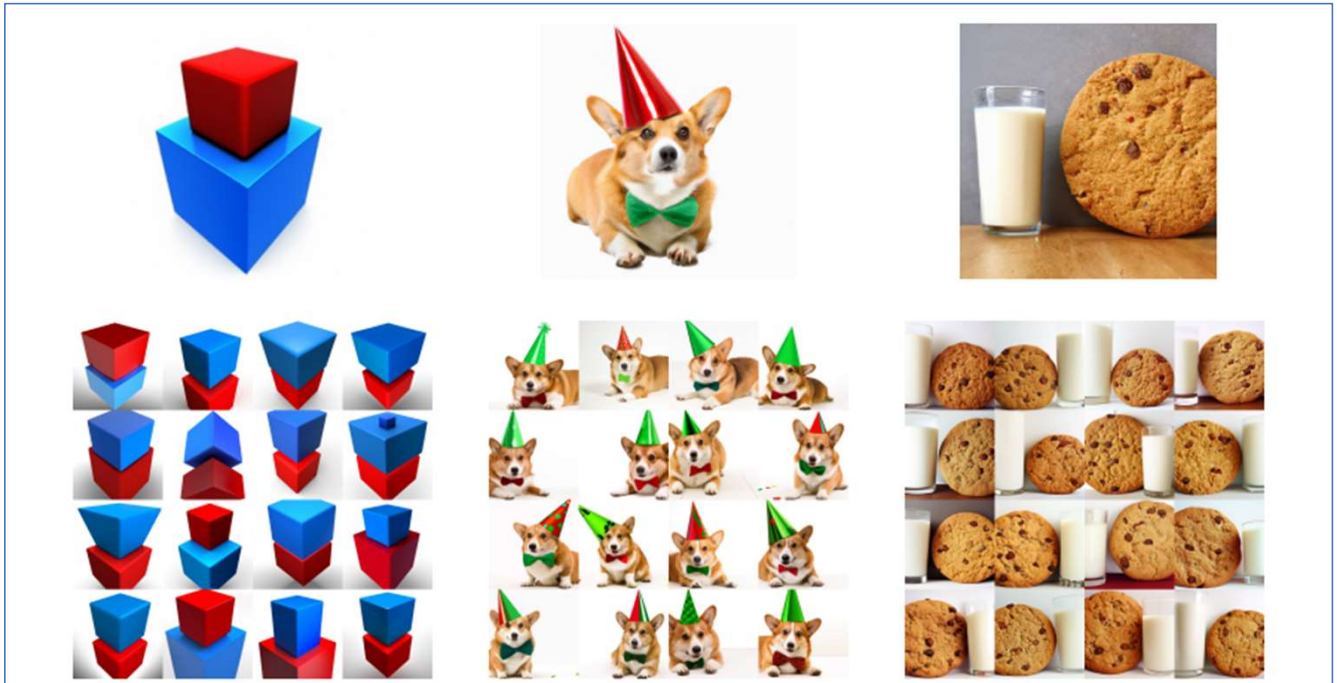


Figure 15: Reconstructions from the decoder for difficult binding problems. We find that the reconstructions mix up objects and attributes. In the first two examples, the model mixes up the color of two objects. In the rightmost example, the model does not reliably reconstruct the relative size of two objects.

DALL-E 2 – Limites : reconstruction de texte dans les images

Figure 16: Samples from unCLIP for the prompt, “A sign that says deep learning.”