Zero-Shot Text-to-Image Generation(Dall-e)

Aditya Ramesh 1 Mikhail Pavlov 1 Gabriel Goh 1 Scott Gray 1 Chelsea Voss 1 Alec Radford 1 Mark Chen 1 Ilya Sutskever 1

OpenAI

July, 1, 2021
구재원

Introduction

>> DALL-E: Zero-shot Image Generation by text prompt

- Input: Text tokens + image tokens (250 millions of text-image pairs)
- Output: image tokens
- Models: VQ-VAE + Transformer

https://openai.com/blog/dall-e/#fn2

Large-scale Generative Models

>> Problems: Still suffer from severe artifacts such as object distortion, illogical object placement, or unnatural blending of foreground and background elements

>> Recent advances of autoregressive transformers have achieved impressive results in several domains by large-scale of compute, model, and data

>> Text-to-image generation has typically been evaluated on relatively small datasets such as MS-COCO and CUB-200

Could dataset size and model size be the limiting factor of current approaches?

Contributions

>> Training a 12-billion params of an autoregressive transformer on 250 M image-text pairs

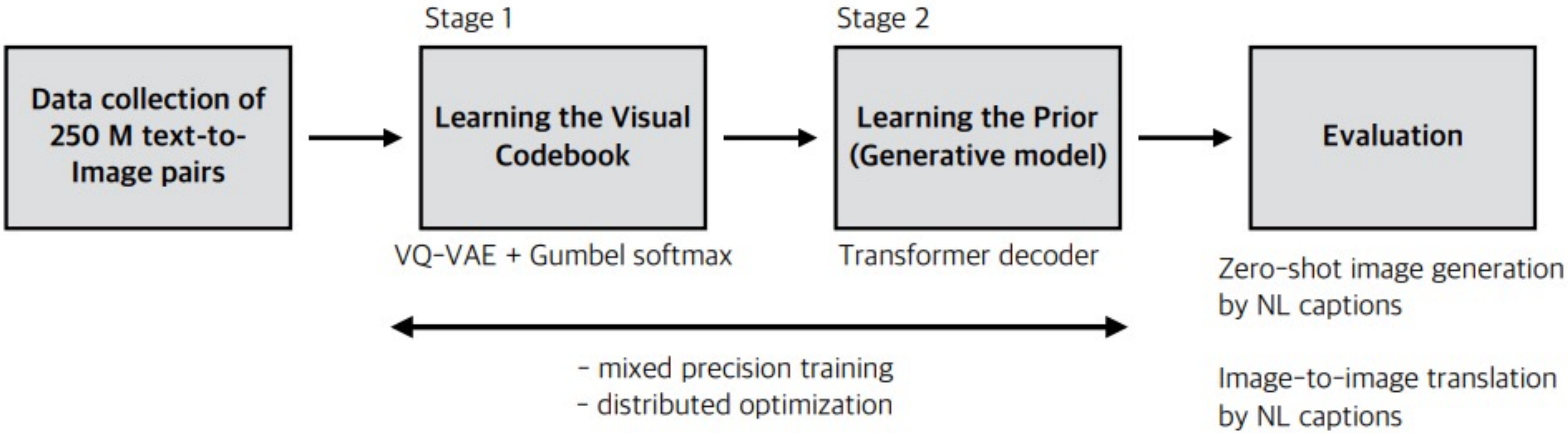
>> Flexible & high fidelity generative model of images controllable by natural language.

>> Zero-shot image generation on MS-COCO dataset.

>> 90 % of people prefer the DALL-E's images than those of prior work.

>> Image-to-image translation

Overall Method



[출처] https://drive.google.com/file/d/1QQH1Dyg-r_wlmZHw17sdQPunthDxg6MS/view

Training Objective

>> overall procedure can be views and maximizing the evidence of lower bound by the joint distribution of an Images x, captions y, and the image tokens z.

$$p_{\theta,\psi}(x, y, z) = \underbrace{p_\theta(x \mid y, z)}_{\text{VQ-VAE decoder}} \underbrace{p_\psi(y, z)}_{\text{Transformer decoder}}$$

$$\ln p_{\theta,\psi}(x, y) \geqslant \mathop{\mathbb{E}}_{z \sim q_\phi(z \mid x)} \Big(\ln p_\theta(x \mid y, z) - \beta \, D_{\mathrm{KL}}(q_\phi(y, z \mid x), p_\psi(y, z)) \Big),$$

Step 1: Learning the Visual Codebook

>> using pixels directly as image tokens would require an inordinate amount of memory for high-resolution images

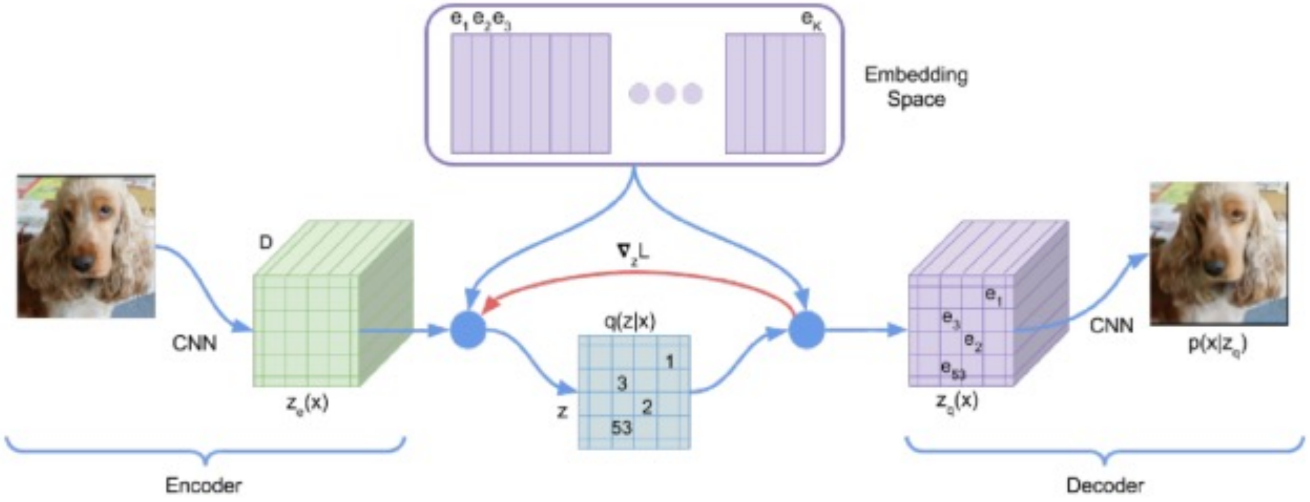
>> low frequency < high frequency

>> Downsampling of an input image is required for the scalable feasibility
  - ex) 256 x 256 -> 32 x 32

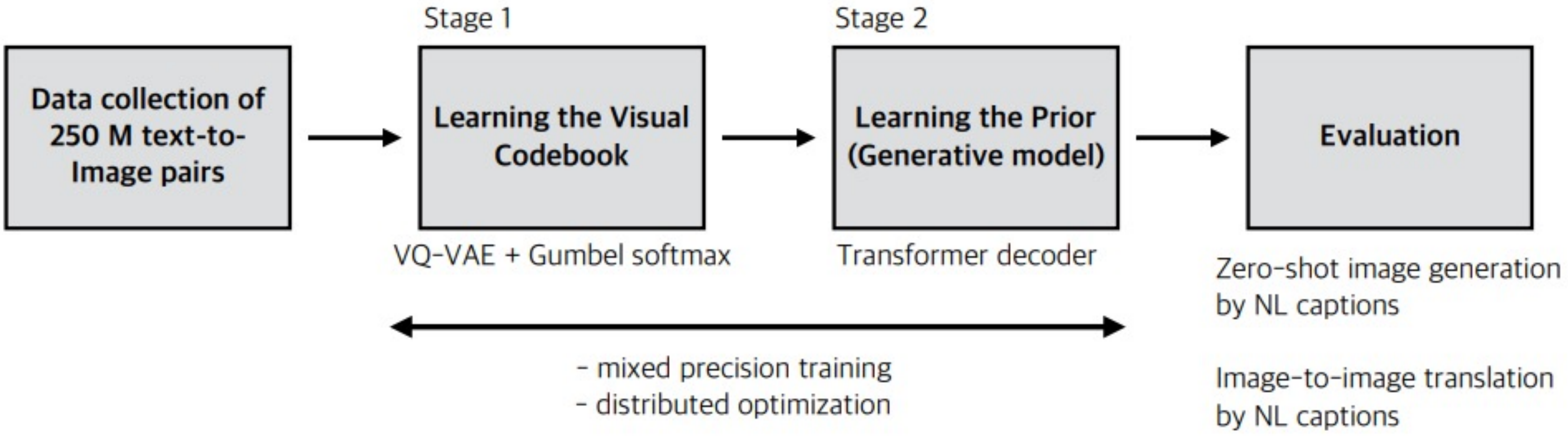
>> VQ-VAE is used to downsample

Step 1: Learning the Visual Codebook

>> VQ-VAE : The code book is updated to contain latent representations for the reconstruction of training samples.
>> gumbel softmax (Jang et al., ICLR'17) is used to backpropagate the reconstruction loss, beyond the stochastic layer with categorical random variables.



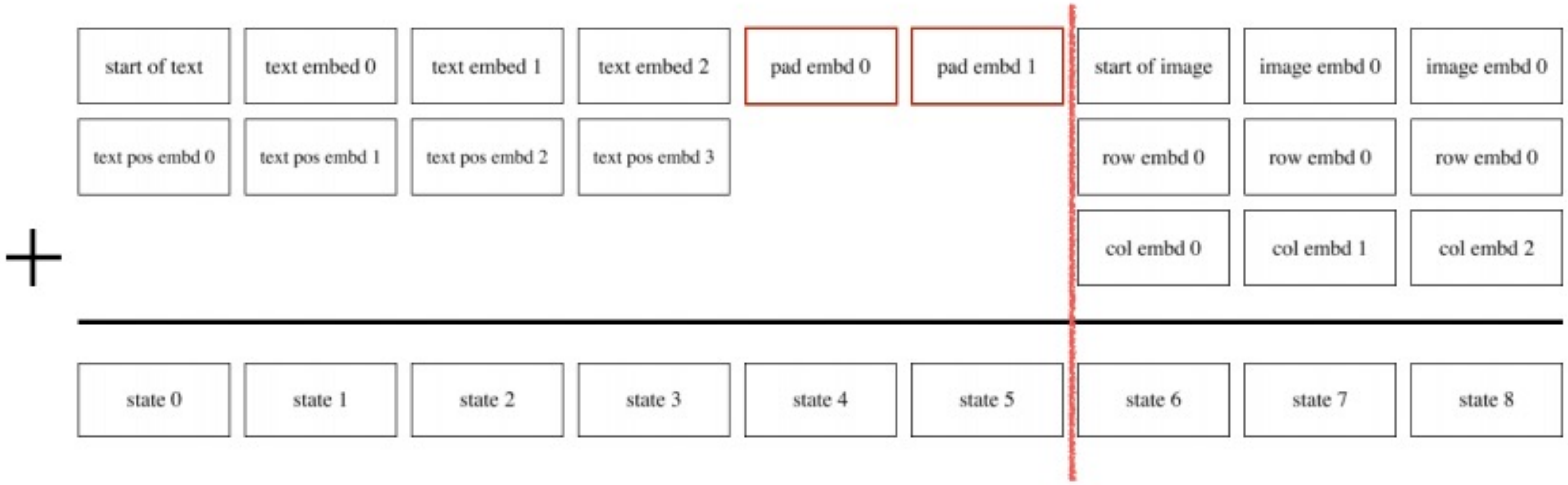
Overall Method



[출처] https://drive.google.com/file/d/1QQH1Dyg-r_wlmZHw17sdQPunthDxg6MS/view

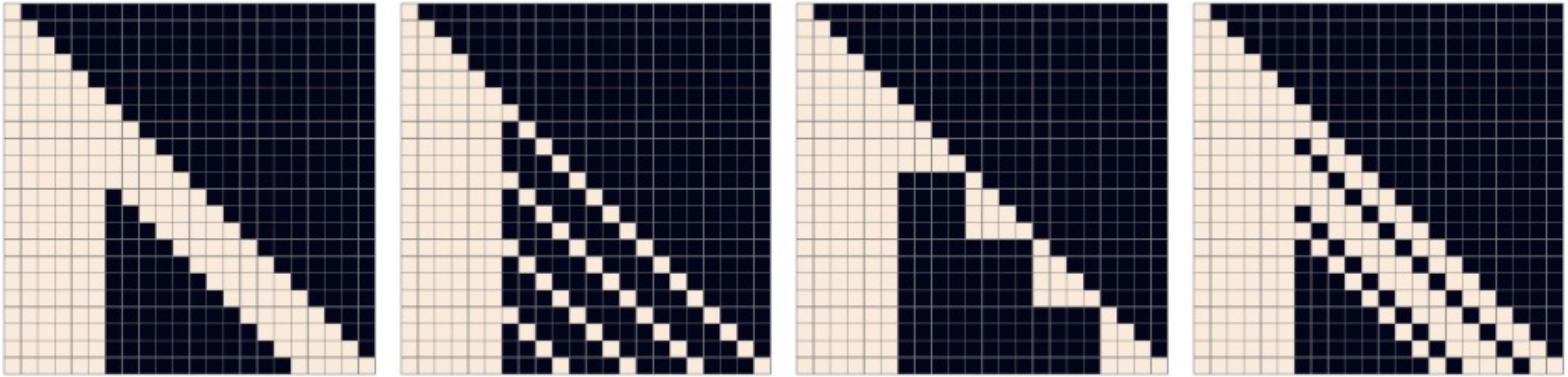
Step2: Learning the Prior

>> Text: BPE-encode using at most 256 tokens / Image: encode 32x32=1024 tokens
>> The text and image tokens are concatenated and used for the next token prediction.



Step2: Learning the Prior

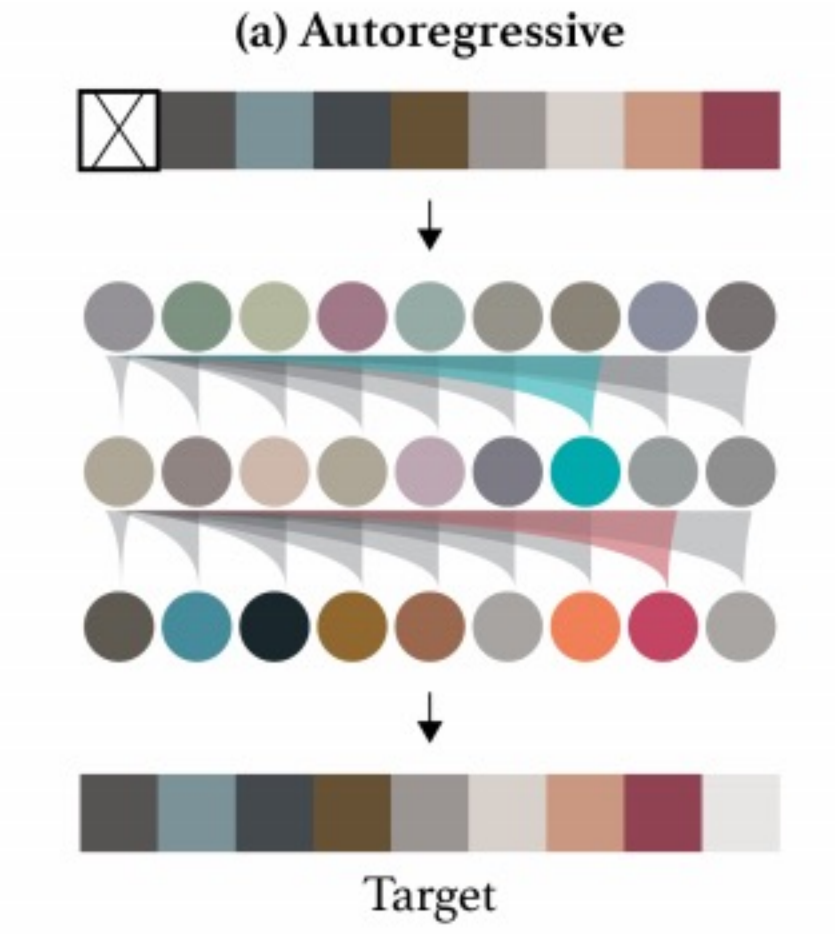
>> model uses three kinds of sparse attention masks



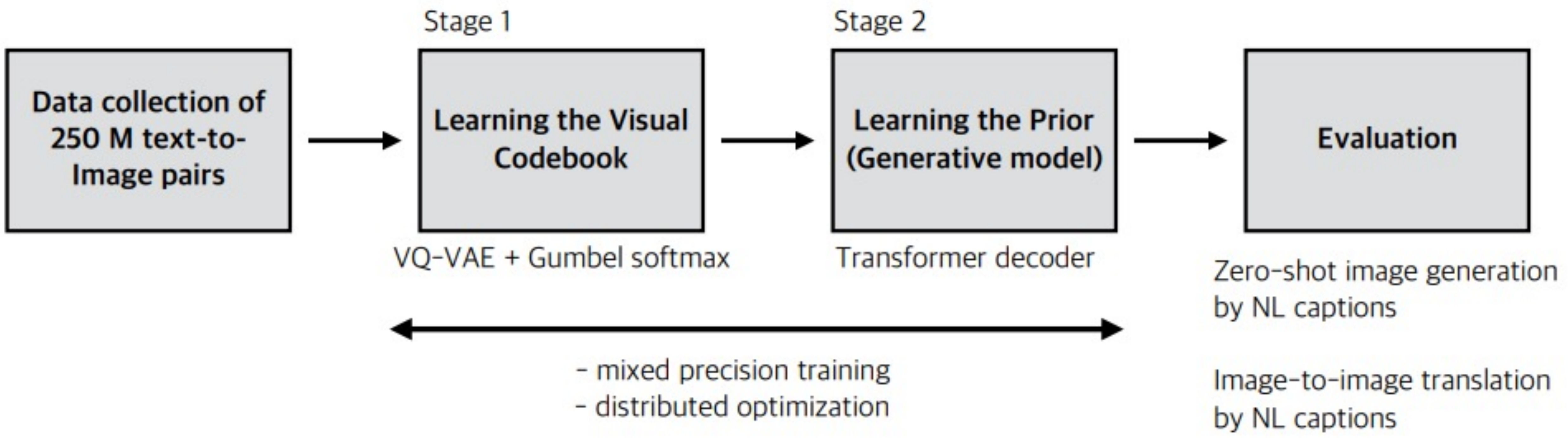
(a) Row attention mask. (b) Column attention mask. (c) Column attention mask with transposed image states. (d) Convolutional attention mask.

Step2: Learning the Prior

>> The transformer decoder is trained to predict next tokens



Overall Method

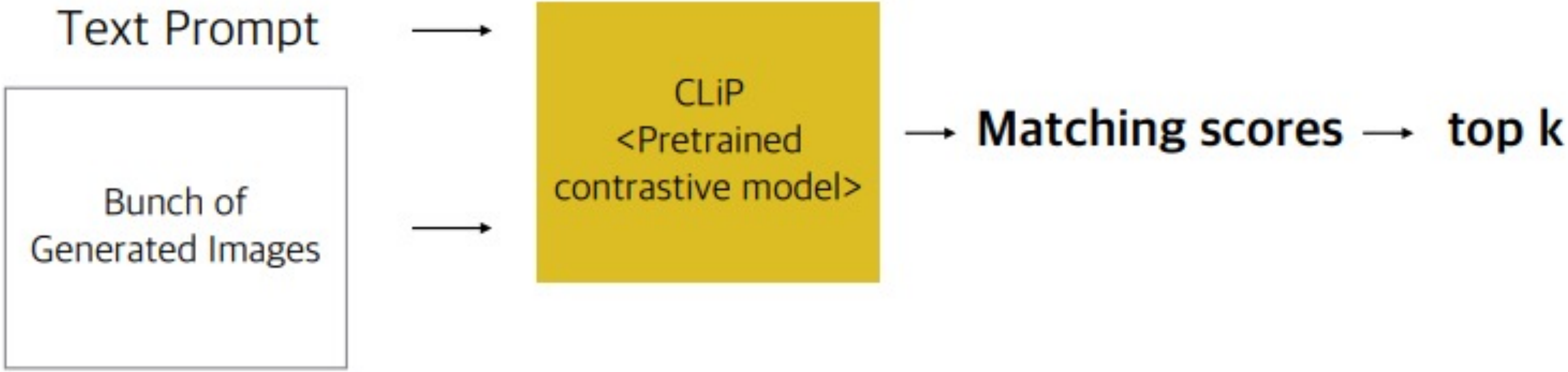


[출처] https://drive.google.com/file/d/1QQH1Dyg-r_wlmZHw17sdQPunthDxg6MS/view

Results : Sample Generation

>> At the test time, an image is generated by the text prompts with/without the part of given images

>> The samples drawn from the transformer are reranked by a retrained contrastive model (CLIP).



Results

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

Figure 2. With varying degrees of reliability, our model appears to be able to combine distinct concepts in plausible ways, create anthropomorphized versions of animals, render text, and perform some types of image-to-image translation.

Results

(a) "the exact same cat on the top as a sketch on the bottom"

(b) "the exact same photo on the top reflected upside-down on the bottom"

(c) "2 panel image of the exact same cat. on the top, a photo of the cat. on the bottom, an extreme close-up view of the cat in the photo."

(d) "the exact same cat on the top colored red on the bottom"

(e) "2 panel image of the exact same cat. on the top, a photo of the cat. on the bottom, the cat with sunglasses."

(f) "the exact same cat on the top as a postage stamp on the bottom"

Figure 14. Further examples of zero-shot image-to-image translation.

Results

Figure 8. Zero-shot samples from our model on the CUB dataset.

Results

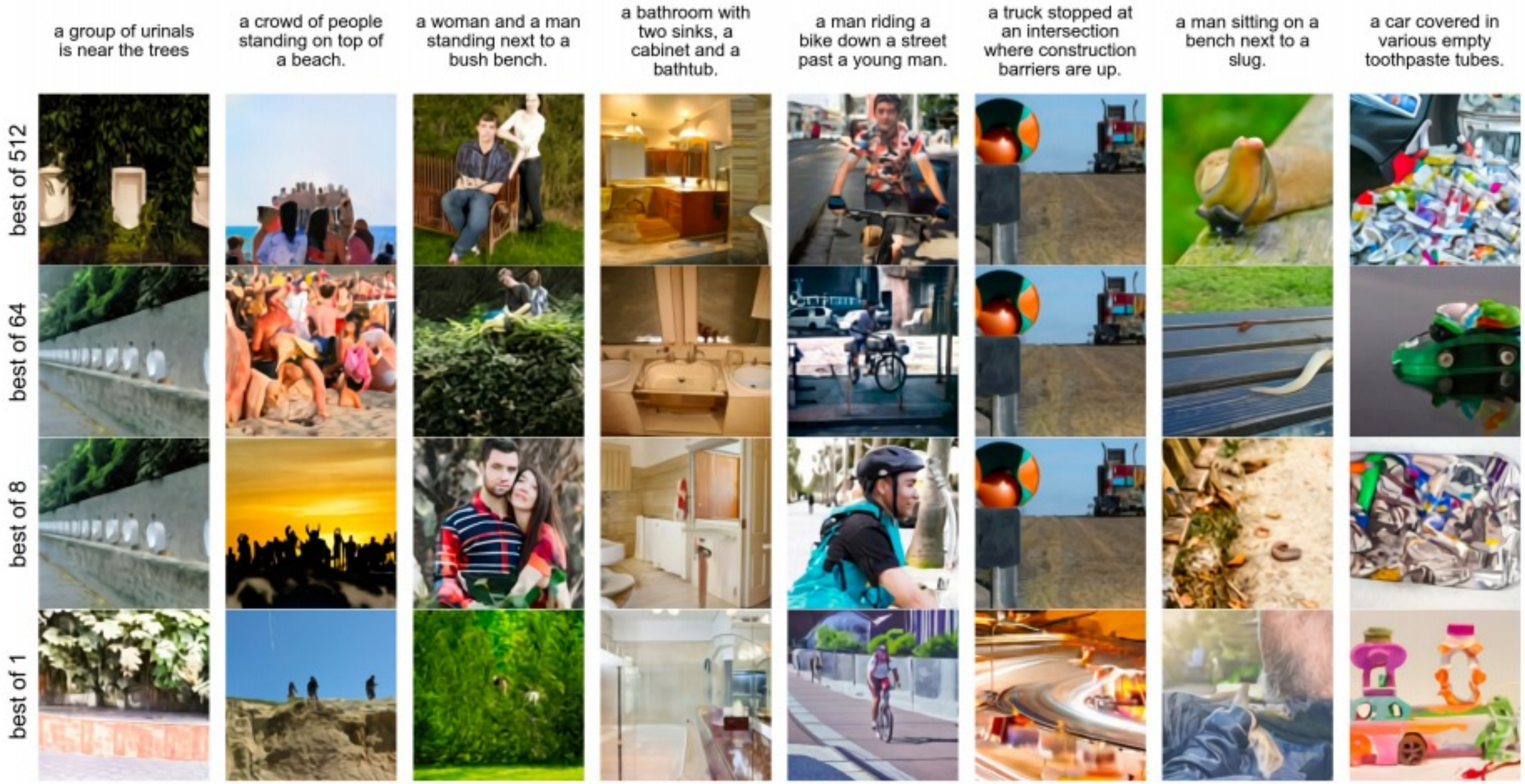


Figure 6. Effect of increasing the number of images for the contrastive reranking procedure on MS-COCO captions.