Research Report on Hierarchical Text-Conditional Image Generation with CLIP Latents

1. Research Background

In recent years, artificial intelligence has made significant breakthroughs in the field of image generation. The traditional Generative Adversarial Network (GAN) performs well in image synthesis, but in the task of text-to-image conversion, it often suffers from insufficient semantic consistency and limited generation diversity. Meanwhile, the CLIP (Contrastive Language–Image Pretraining) model proposed by OpenAI learns through contrastive training of a large number of image-text pairs, constructing a unified representation space for images and text, and demonstrating strong semantic understanding and zero-shot generalization capabilities. This study is based on the representational advantages of CLIP, combined with the generative capability of the Diffusion Model, and proposes a new hierarchical text-conditioned image generation framework, unCLIP. It achieves significant improvements in generation quality, diversity, and interpretability.

2. Research Objectives and Innovation Points

The core objective is to achieve controllable, high-fidelity, semantically consistent, and diverse text-to-image generation within the CLIP semantic latent space.

The main innovations include:

1. **Proposing a two-stage generation system (Hierarchical Generation)**
   Stage 1: The Prior model maps the text to CLIP image embeddings.
   Stage 2: The Decoder model decodes the embeddings back into images.

2. **Introducing the CLIP latent space as an intermediate layer**
   Effectively combines semantic understanding and image generation; provides visualization and semantic editing capabilities.

3. **Adopting diffusion Prior instead of autoregressive Prior**
   Significantly improves training efficiency and generation quality.

4. **Proposing a language-guided image manipulation method (Text Diffs)**
   Enables image modification and hybrid generation based on textual semantics.

3. Model Architecture and Method

The **unCLIP** model is a two-stage hierarchical text-to-image generation framework. Instead of directly generating pixel-level images from text, it first produces semantic representations in the CLIP latent space and then decodes them into images using a diffusion model. This decomposition simplifies the complex text-to-image task into two learnable subproblems:

1. **Text to Image Representation (Text → CLIP Embedding)**
2. **Image Representation to Pixel Generation (Embedding → Image)**

4. Model Overview

The overall architecture of unCLIP consists of two components:

- **Prior Model:** Responsible for generating a CLIP image embedding z_i from a text caption y, i.e. $P(z_i|y)$.
- **Decoder Model:** Uses a diffusion model to generate an image x from the CLIP image embedding, i.e. $P(x|z_i, y)$.

The overall conditional probability can therefore be expressed as:

$$P(x|y) = P(x|z_i, y)P(z_i|y)$$

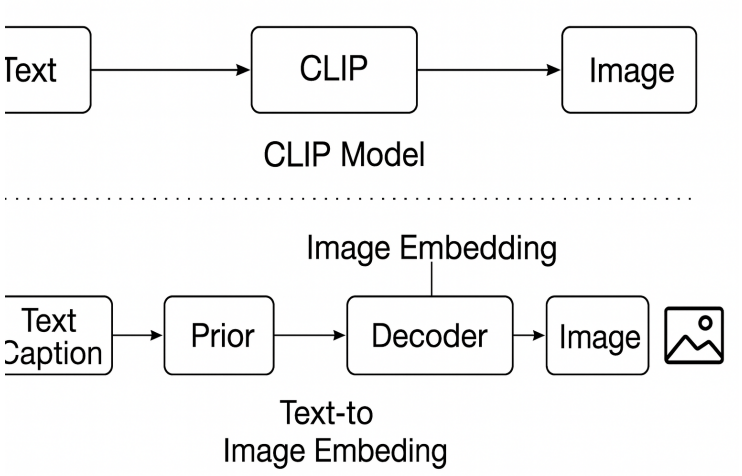


Figure 1: unCLIP Model Overview. The model first maps the text caption into a CLIP image embedding through the Prior model, then generates a realistic image via the Diffusion Decoder conditioned on that embedding.

5. Prior Model: Text-to-Image Embedding

The Prior model forms the first stage of unCLIP, transforming text embeddings into CLIP image embeddings. Two types of priors are proposed and compared:

- **Autoregressive Prior (AR):** Converts embeddings into discrete tokens and predicts them autoregressively. It is accurate but computationally expensive.
- **Diffusion Prior:** Models the embedding distribution in continuous space, offering better efficiency and generation quality.

During training, a classifier-free guidance mechanism is applied by randomly dropping text conditioning to improve diversity and robustness.

6. Decoder Model: Diffusion-Based Image Generation

The decoder employs a diffusion model to iteratively denoise random noise into an image. Conditioned on the CLIP image embedding, the model reverses the diffusion process step by step, gradually transforming a noisy latent into a realistic image.

7. Training Highlights

- **Dataset:** 650M text–image pairs collected from the Internet.
- **Classifier-Free Guidance:** Randomly drops conditioning inputs during training to balance fidelity and diversity.
- **Hierarchical Upsampling:** The model generates images progressively from 64×64 to 256×256 and finally 1024×1024.
- **PCA Optimization:** Principal Component Analysis is applied to reduce the dimensionality of the CLIP embedding space while preserving semantic richness.

8. Experiments and Results

The unCLIP model was trained on a large-scale dataset containing approximately 650 million text–image pairs collected from the Internet. The training data combines samples from both CLIP and DALL·E datasets.

Quantitative Results

On the MS-COCO 256×256 benchmark, unCLIP achieves a FID score of 10.39, outperforming GLIDE (12.24) and DALL·E (28). Both human evaluations and aesthetic metrics show that unCLIP achieves an excellent trade-off between fidelity and diversity, especially when using the diffusion prior.

Qualitative Results

Generated samples demonstrate strong semantic alignment and visual creativity. Prompts such as *"a shiba inu wearing a beret and black turtleneck"* or *"a teddy bear on a skateboard in Times Square"* produce photorealistic and diverse results. In addition, unCLIP supports:

- **Image Variations:** Generating stylistically consistent yet distinct versions of an image.
- **Interpolation:** Blending between different concepts or input images.
- **Text Diffs:** Performing language-guided semantic edits, e.g., changing "a lion" into "a lion cub".

Figure 2: Examples of unCLIP text-to-image generation and text-guided editing. Source: Ramesh et al., *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv:2204.06125, 2022.

9. Comparative Analysis

Compared with other text-to-image diffusion models, unCLIP demonstrates superior performance in diversity and interpretability.

Model	Photorealism	Caption Similarity	Diversity
GLIDE	52.9%	58.9%	37.4%
unCLIP (AR Prior)	47.1%	41.1%	62.6%
unCLIP (Diffusion Prior)	**48.9%**	**45.3%**	**70.5%**

Table 1: Human evaluation results comparing unCLIP and GLIDE across different metrics.

Humans preferred unCLIP for semantic diversity while maintaining comparable photorealism to GLIDE. Moreover, because unCLIP stores semantic content in CLIP embeddings, it avoids mode collapse under high guidance scales, unlike GLIDE which converges to fewer visual modes.

10. Summary of Innovations

Innovation	Technique	Effect
Hierarchical Generation	Text $\rightarrow$ Embedding $\rightarrow$ Image	Improves semantic consistency and diversity
Diffusion Prior	Continuous latent modeling	Outperforms AR prior in efficiency and qual
Latent Visualization	CLIP decoding and reconstruction	Enhances interpretability of semantic space
Classifier-Free Guidance	Random conditional dropout	Balances fidelity and diversity

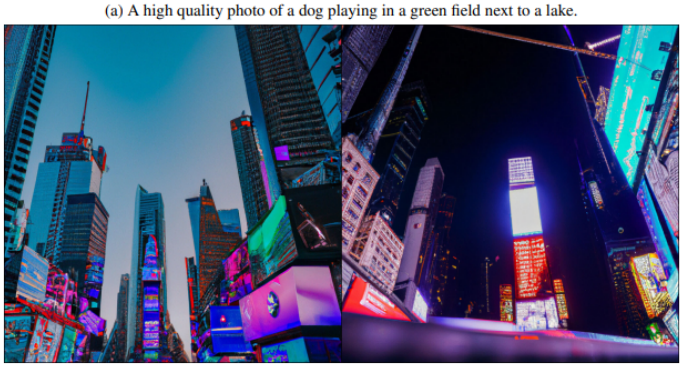
Table 2: Key innovations and corresponding benefits of unCLIP.

11. Limitations and Risks

Despite its strong performance, unCLIP has several limitations:

- **Attribute Binding Errors:** The model may confuse color–object pairs (e.g., "a red cube on a blue cube").
- **Text Rendering Issues:** Generated text often appears distorted or misspelled.
- **Detail Loss in Complex Scenes:** Fine details are sometimes lost in crowded or high-resolution environments.
- **Ethical Concerns:** The realism of generated content raises risks of misinformation and misuse.

(a) A high quality photo of a dog playing in a green field next to a lake.



(b) A high quality photo of Times Square.

Figure 3: Examples of unCLIP limitations: attribute confusion, distorted text, and missing details in complex scenes. Source: Ramesh et al., *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv:2204.06125, 2022.

12. Applications and Future Work

unCLIP has already been integrated into OpenAI's DALL·E 2 system, proving its value in artistic creation, media design, and digital content generation. Future directions include:

- Training at higher base resolutions to improve detail reconstruction.
- Enhancing attribute–object binding through structured attention mechanisms.
- Developing controllable latent editing for more precise text–image alignment.
- Establishing ethical safeguards and watermarking systems for responsible deployment.

Conclusion

The unCLIP framework bridges the gap between semantic understanding and high-fidelity image generation. By leveraging CLIP latent representations and diffusion modeling, it achieves a new balance of accuracy, diversity, and interpretability—surpassing GLIDE and DALL·E in both creativity and visual quality. However, as generated content becomes increasingly realistic, continued efforts toward ethical transparency and technical refinement are essential for the responsible evolution of generative AI.