LLM Blueprint: Enabling Text-to-Image Generation with Complex and Detailed Prompts

Hanan Gani[1], Shariq Farooq Bhat[2],
Muzammal Naseer[1], Salman Khan[1,3] , Peter Wonka[2]

ICLR 2024

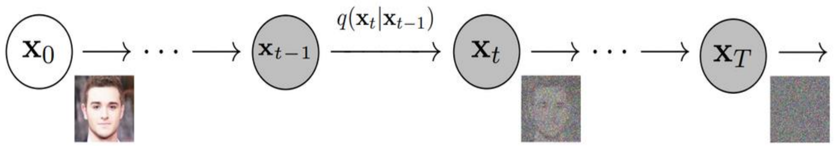
[1]Mohamed Bin Zayed University of Artificial Intelligence [2]KAUST [3]Autralian National University

Background: Diffusion models

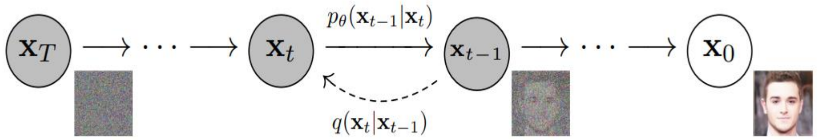
- Foundational Generative Model
    - Functions in two stages. The forward diffusion adds gaussian noise to the image and reverse diffusion iteratively learns to recover the data sample.
    - Stable diffusion - Trained on millions on image-text pairs. Works in latent space.

- Motivation
    - Generates high-quality and diverse images with fine details.
    - Avoids Mode collapse (common with GANs).
    - Diffusion models have been shown to be more robust in overfitting due to their use of likelihood-based training.
    - Flexible latent space

Forward diffusion



Reverse diffusion



(Ho et al. 2020)

Problem Statement

- Generate coherent images from long and complex textual prompts.
  - Training-free manner (zero-shot).
  - Generated images should encompass all objects present in the long textual paragraph such that each object follows its properties/description from the prompts.
  - Generated images should follow the spatial locations of the objects in the prompt.

Shortcomings of existing Diffusion based methods

- Often struggle to faithfully capture all the nuanced details within longer and elaborate textual inputs.
- Use CLIP text encoder which can only process first 77 text tokens potentially omitting critical details.
- Current diffusion models are not trained on long textual prompts and lack location-aware training.

Proposed Method: Scene Blueprint

ICLR

Long Textual Prompt

In a cozy living room, a heartwarming scene unfolds. A friendly and affectionate **Golden Retriever** with a soft, golden-furred coat rests contently on a plush rug, its warm eyes filled with joy. Nearby, a graceful and elegant **white cat** stretches leisurely, showcasing its pristine and fluffy fur. A sturdy **wooden table** with polished edges stands gracefully in the center, adorned with a **vase of vibrant flowers** adding a touch of freshness. On the wall, a **sleek modern television** stands ready to provide entertainment. The ambiance is warm, inviting, and filled with a sense of companionship and relaxation.

LLM

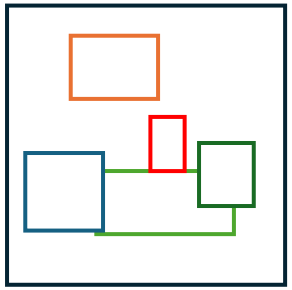
In-Context learning

Background prompt

A cozy living room filled with a sense of companionship and relaxation.

Layout Bounding boxes

White Cat: {x1, y1,w1, h1}, Golden retriever: {x2, y2,w2, h2}, Wooden table: {x3, y3,w3, h3}, T.V.: {x4, y4,w4, h4}, Flower vase: {x5, y5,w5, h5}



Object descriptions

White Cat: graceful and elegant white cat stretching leisurely, showcasing its pristine and fluffy fur; **Golden retriever**: soft, golden-furred coat; **Wooden table**: sturdy with polished edges; T.V.: sleek modern T.V. ; **Flower vase**: vase of vibrant flowers

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

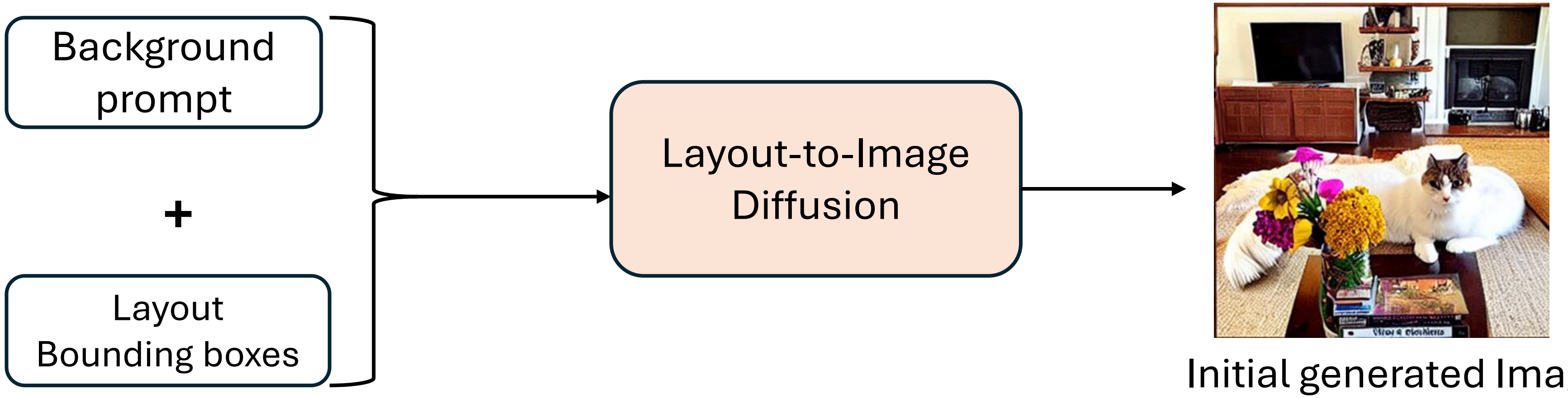
KAUST

Australian National University

Proposed Method: Global Scene Generation

- Leverage LLM to generate:
  - K bounding boxes for each object - `(x,y,width,height)`
  - A succinct background prompt
  - Object-specific descriptions present in the textual prompt - `{object: description}`

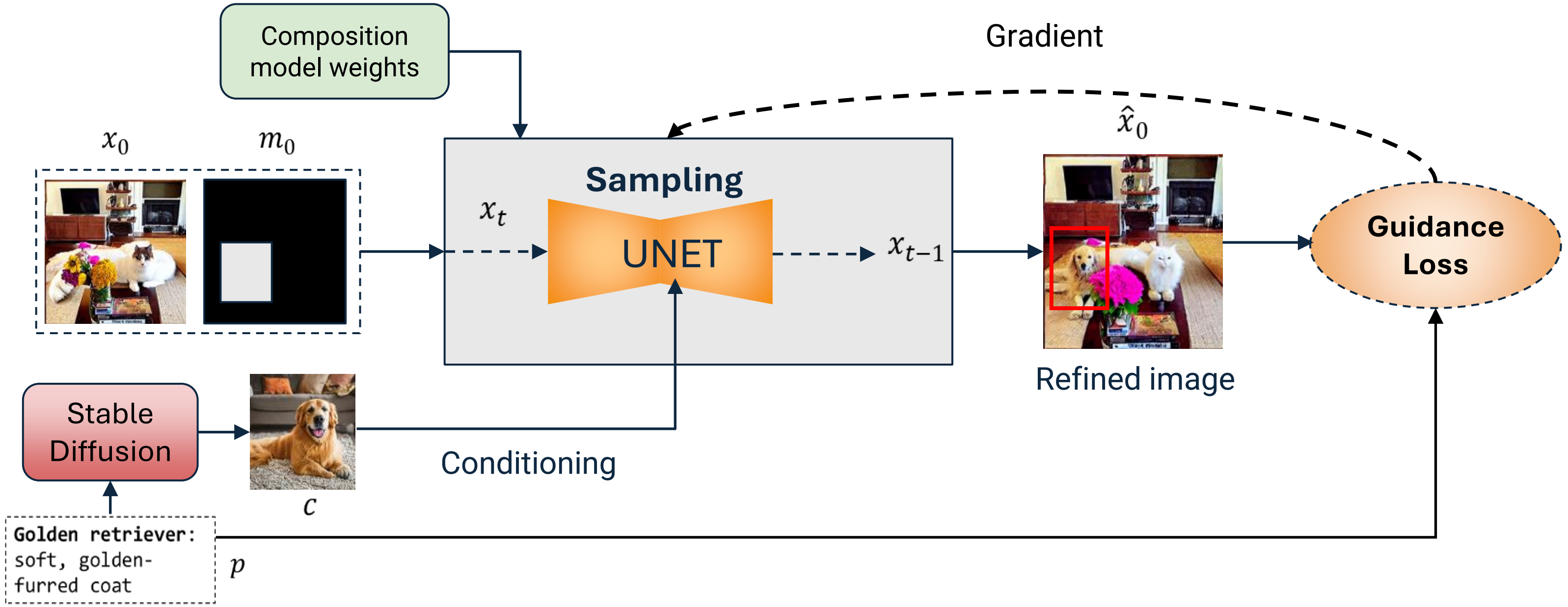
- Obtain initial image using Layout-to-Image generator by feeding bboxes and background prompt to a diffusion model. The initial image has missing objects or inaccurate misrepresentations.



Initial generated Image

Proposed Method: Iterative Refinement

- **Iterative Refinement:**
  - At each box location, refine the object through an image composition strategy.
  - Guide the object generation sampling process through a multi-modal loss function such that the generated objects align well with their respective descriptions.



Forward diffusion

- Add noise:

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1 - \alpha_t}\epsilon, \qquad \epsilon \sim N(0, I)$$

Reverse Diffusion

- Predict clean sample: $\hat{x}_0 = \dfrac{x_t - \sqrt{1 - \alpha_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}$
- DDIM Sampling: $x_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\,\epsilon_\theta(x_t, t)$
- Classifier guidance: $\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + s(t).\nabla_{x_t} l(p, \hat{x}_0)$

Proposed Method: CLIP Guidance

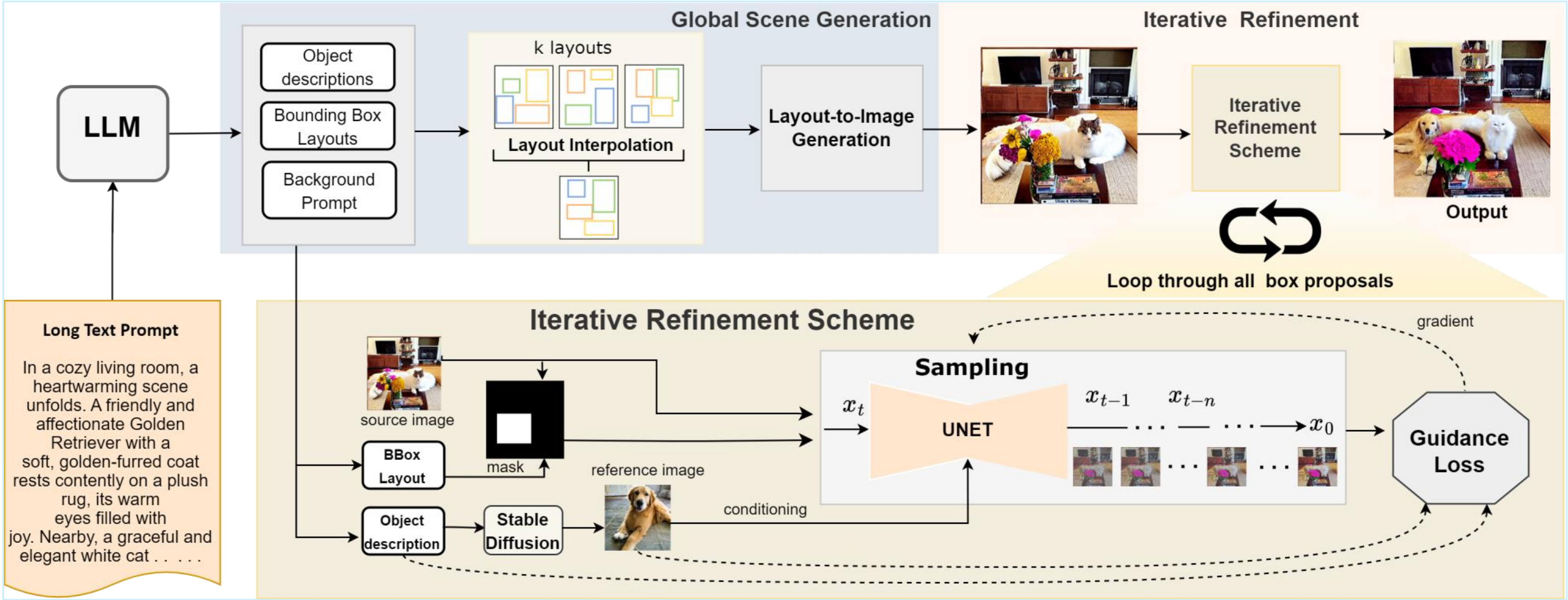
- CLIP based guidance: $\mathcal{L}_{CLIP}(p, \hat{x}_0) = \mathcal{L}_{cosine}\left(\mathrm{CLIP}_{image}(\hat{x}_0 \cdot m_j), \mathrm{CLIP}_{text}(p_j)\right)$

- Additional background preservation loss: $\mathcal{L}_{bg} = \mathcal{L}_2\left((\hat{x}_0 \cdot (1 - m_j)),\ x_{initial} \cdot (1 - m_j)\right)$

- Final guidance Loss: $\boxed{\mathcal{L}_{guidance} = \mathcal{L}_{CLIP} + \gamma . \mathcal{L}_{bg}}$

- Updated DDIM Noise: $\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + s(t). \nabla_{x_t} \mathcal{L}_{guidance}$

Proposed Method: Overall pipeline



Quantitative Comparisons: Prompt Adherence Recall (PAR)

- To the best of our knowledge, there is currently no established metric for assessing the performance of diffusion models in handling lengthy text descriptions.
- We propose Prompt Adherence Recall (PAR) score which uses an off-the-shelf grounded object detector (GLIP) to detect the objects in the image.

$$\text{PAR Score} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbb{1}_{i,j}(o, p)$$

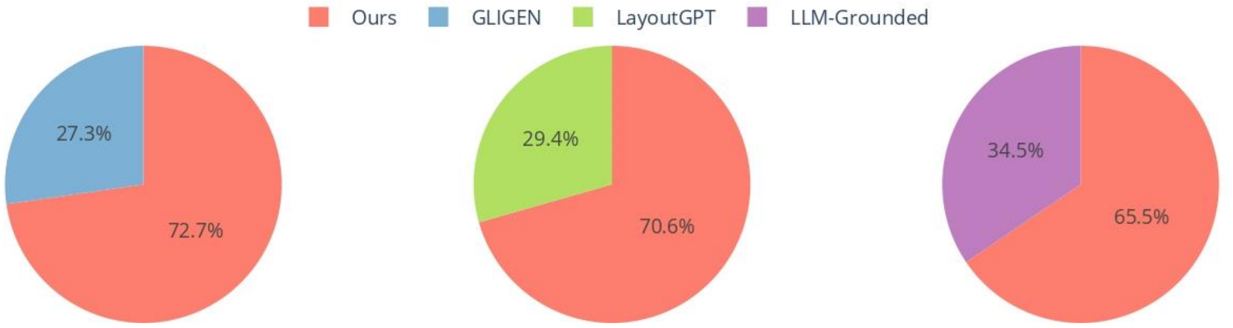
$$\mathbb{1}_{i,j}(o, I) = \begin{cases} 1, & \textbf{detect}(o_{i,j}, p_i) \\ 0, & \text{otherwise} \end{cases}$$

Method	PAR score (%) ↑
Stable Diffusion (CVPR'22)	49
GLIGEN (CVPR'23)	57
LayoutGPT (NeurIPS'23)	69
Ours	**85**

Quantitative Comparisons: Composition of Generated Images

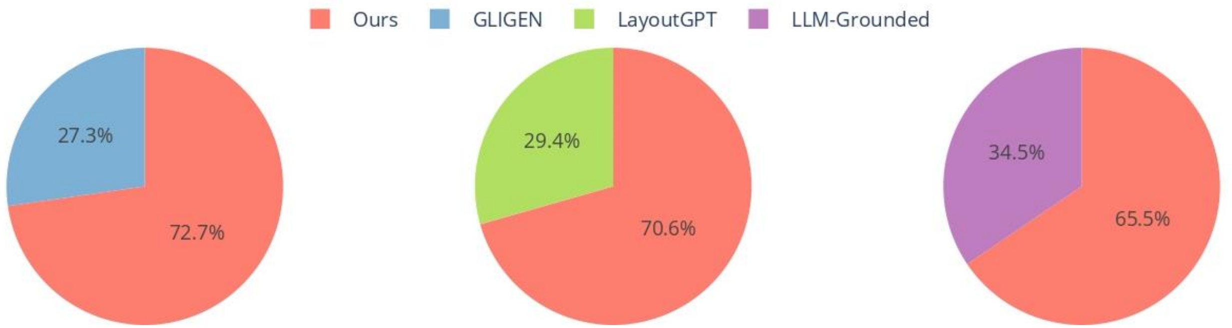
- Participants were instructed to select one image from a pair of images randomly selected from two distinct approaches following a 2-AFC task.

- Their goal was to choose the image that most accurately represented the provided textual descriptions regarding spatial arrangement, object characteristics, and overall scene dynamics.



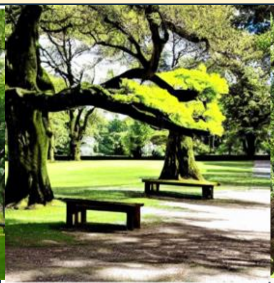
Quantitative Comparisons: Image Quality

- Participants were instructed to choose the image that is best in terms of quality from a pair of images randomly selected from two distinct approaches.



Qualitative Visualizations: Baseline comparisons

ICLR

Long textual prompt	Stable diffusion	GLIGEN	LayoutGPT	LLM-Grounded	Ours
The image shows a living room with a *christmas tree in the corner*. There is a *couch, a chair, and a coffee table* in the room. The *walls are painted* white and there is a *large window with curtains* on one side. There is a *bicycle parked in the corner* of the room. The room is well lit by a *large chandelier hanging* from the ceiling. The overall atmosphere of the room is cozy and festive.					
	chair, table, white painted walls, bicycle christmas tree	christmas tree, chair, chandelier, bicycle	bicycle christmas tree	Bicycle, coffee table christmas tree	
In a serene park area, nature's beauty flourishes. A sprawling *green lawn* stretches out, inviting visitors to revel in its vibrancy. *Towering trees provide a soothing canopy*, their branches offering refuge from the *sun's warm embrace*. Amidst this natural oasis, a *simple bench* beckons, a quiet sanctuary for those seeking respite. Nearby, a *quaint building* blends seamlessly with the park's landscape, adding a touch of architectural charm. The scene is a harmonious fusion of lush greenery and tranquil spaces, where the chair offers solace, the grass whispers secrets, and the park itself is a haven of serenity.					
	bench, sun's warm embrace, quaint building	sun's warm embrace, quaint building	sun's warm embrace	park bench, sun's warm embrace	

- Red text below refers to missing objects in the image
- Purple text shows incorrect object locations

Comparison with recent SOTA

- Our approach further outperforms recent SOTA approaches for image generation.
- While Deepfloyd used T5 LLM as text encoder, it still fails to generate coherent images.

Method	PAR score (%) ↑	Inference time (min.) ↓
DenseDiffusion	52	2.50
DeepFloyd	60	8.33
Ours	**85**	**3.16**

Conclusion

- We identify the limitations of prior text-to-image models in handling complex and lengthy text prompts.
- We introduce a framework involving a data structure (Scene Blueprint) and a multi-step procedure involving global scene generation followed by an iterative refinement scheme to generate images that faithfully adhere to the details in such lengthy prompts.
- Our framework offers a promising solution for accurate and diverse image synthesis from complex text inputs, bridging a critical gap in text-to-image synthesis capabilities.

Project page