

24

DO NOT PUT ANYTHING HERE

Qualitative Analysis

Text-to-Image Generation: Tackling Multiple Object Cardinality

DO NOT PUT ANYTHING HERE

Aiza Usman, Ramya Ganesh, Prahanya Sriram

DO NOT PUT ANYTHING HERE

Motivation & Problem Statement

Text-to-image models like Stable Diffusion struggle with accurately generating multiple objects with correct counts.

Example Prompt: "Forest with 7 deers with 7 rabbits"



✗ 10+ deers

✗ No rabbits

Impact: Limits use in precision-required applications (e.g. design, education)

Our Goal

Develop a text-to-image diffusion model that consistently honors object counts and spatial references in text prompts.



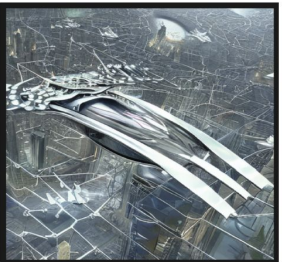







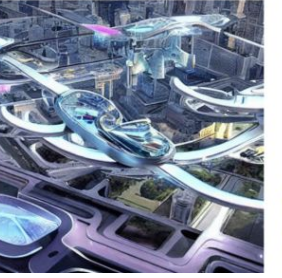









We hope to combine different techniques that eliminate semantic leakage between tokens so that every described element in the prompt appears exactly as such in the image



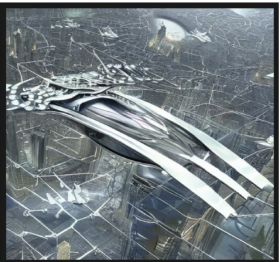





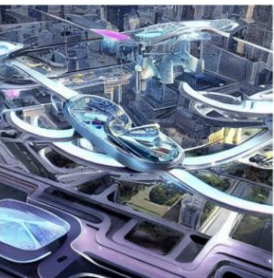









Baseline, Dataset & Metrics

Baseline Model	Stable Diffusion v1.5
Pre-trained Dataset	LAION-5B
Fine-tuning Dataset	“Conceptual Captions” dataset
Metrics	CLIP Score, Precision, F1, Recall

Proposed Techniques

- **Rotary Positional Embeddings (RoPE) + LoRA:**
 - **What this is:** Inject rotary positional embeddings into Text Encoder's CLIP model and fine-tune lightweight LoRA adapters on attention + MLP projections.
 - **Why this helps:** Enables better modeling of relative token positions for counting tasks while keeping fine-tuning memory-efficient and fast.
- **ALiBi (Attention Linear Biases) + LoRA :**
 - **What this is:** Add ALiBi linear biases to Text Encoder's CLIP model and fine-tune lightweight LoRA adapters on attention and MLP projections.
 - **Why this helps:** Supports longer prompt understanding by biasing attention distances, while LoRA ensures efficient and modular fine-tuning.
- **Bounded Cross-Attention:**
 - **What this is:** Limits each token's attention to a predefined range or region, reducing interference between unrelated tokens during image generation.
 - **Why this helps:** Ensures that each object token influences only its relevant part of the image, improving accuracy in generating correct object counts.
- **Slot-Guided Attention:**
 - **What this is:** Assigns fixed attention "slots" to object tokens, guiding the model to generate distinct regions for each object in the prompt.
 - **Why this helps:** Ensures each object in the prompt is represented separately, reducing overlap and improving accuracy in generating the correct number of objects.

Prompts	"A photo of 5 cups and 5 plates"	"a painting of 3 cats and 3 dogs"	"futuristic city with 10 flying cars"	"forest with 7 deer and 7 rabbits"
Baseline (Stable Diffusion Model)				
RoPE + LoRA				
ALiBi + LoRA				
Bounded Attention				
Slot Attention				

Prompts	"A photo of 5 cups and 5 plates"	"a painting of 3 cats and 3 dogs"	"futuristic city with 10 flying cars"	"forest with 7 deer and 7 rabbits"
Baseline (Stable Diffusion Model)				
RoPE + LoRA				
ALiBi + LoRA				
Bounded Attention				
Slot Attention				

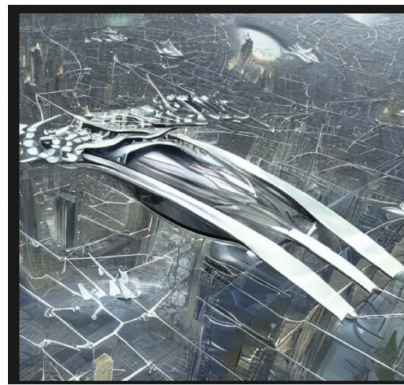
Baseline (Stable Diffusion Model)



'A photo of 5 cups and 5 plates'



'a painting of 3 cats and 3 dogs'



'futuristic city with 10 flying cars'



'forest with 7 deer and 7 rabbits'

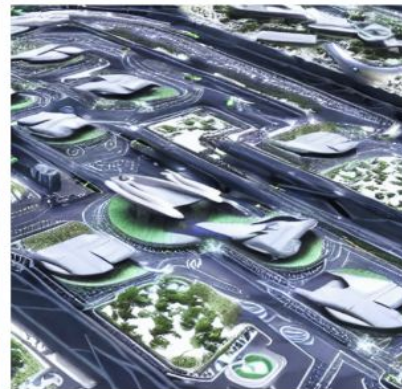
Our Best Model (ALiBi + Slot Attention + LoRA)



'A photo of 5 cups and 5 plates'



'a painting of 3 cats and 3 dogs'



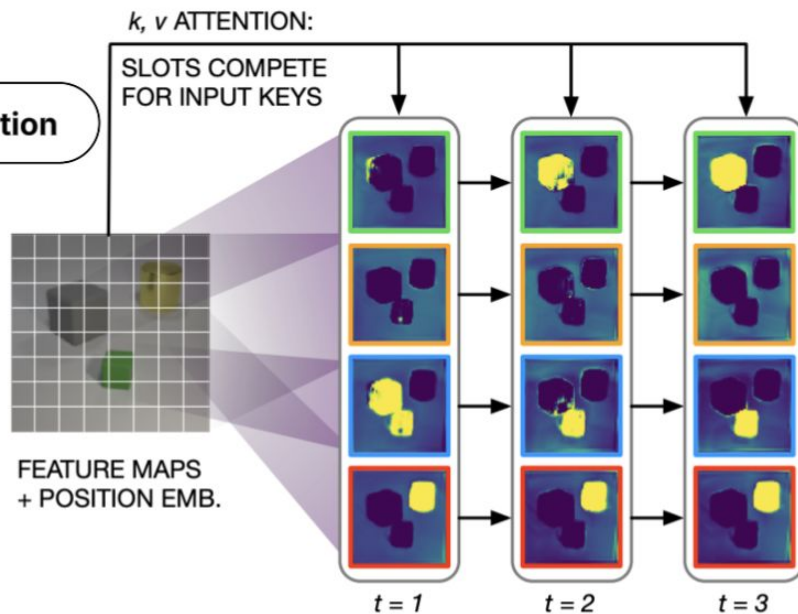
'futuristic city with 10 flying cars'



'forest with 7 deer and 7 rabbits'

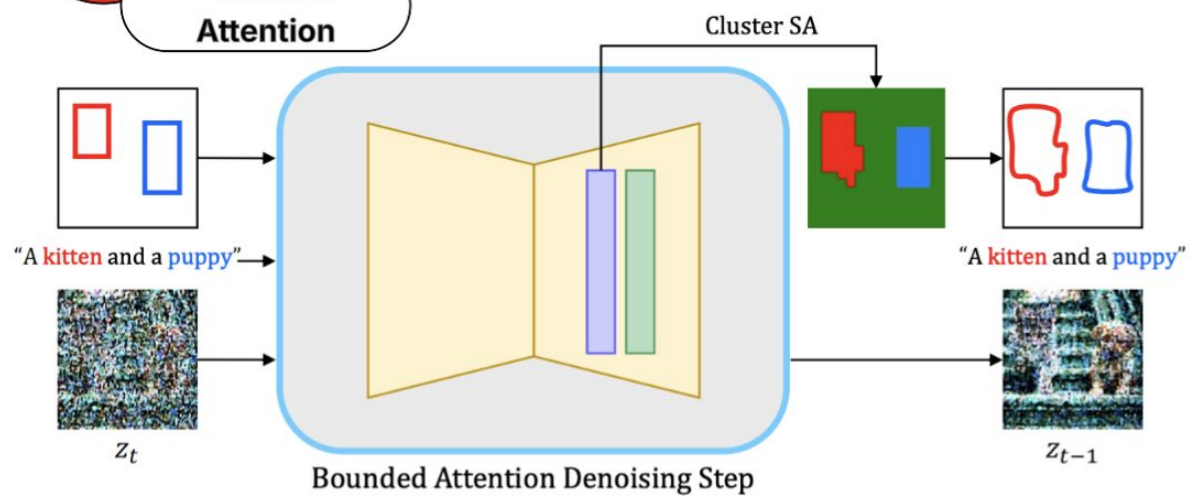
4

Slot Attention

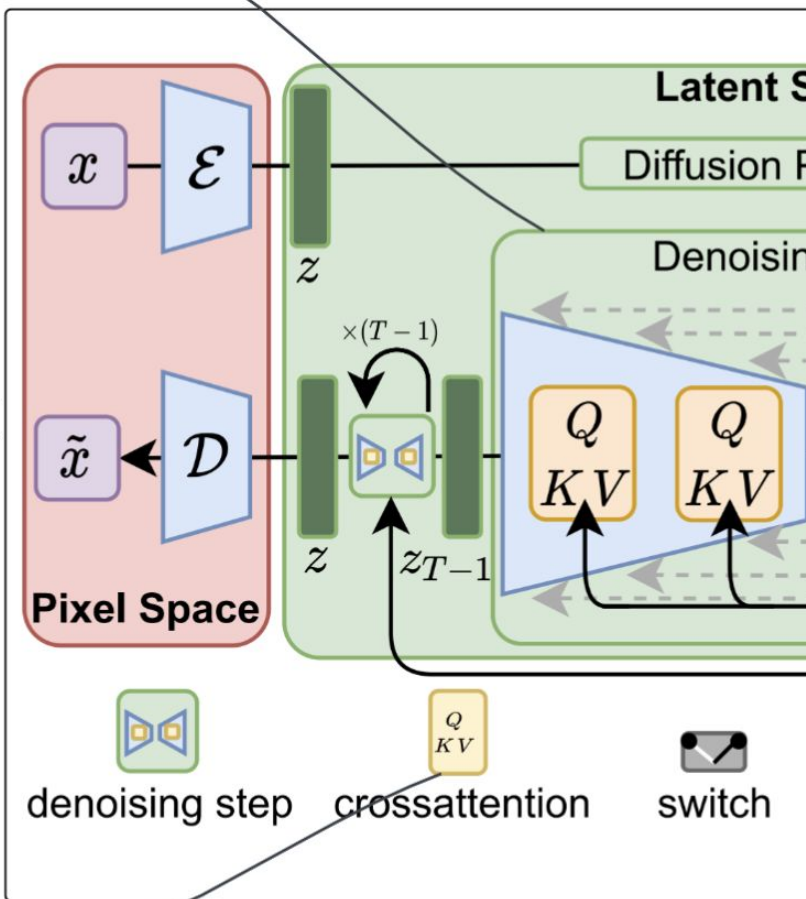


5

Bounded Attention



Method

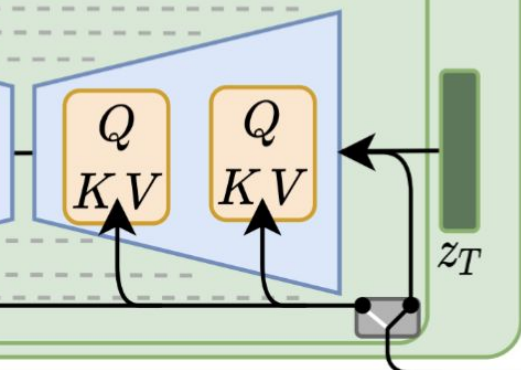


Methodology

Space

Process

ing U-Net ϵ_θ



skip connection

concat

Conditioning

Semantic Map

Text

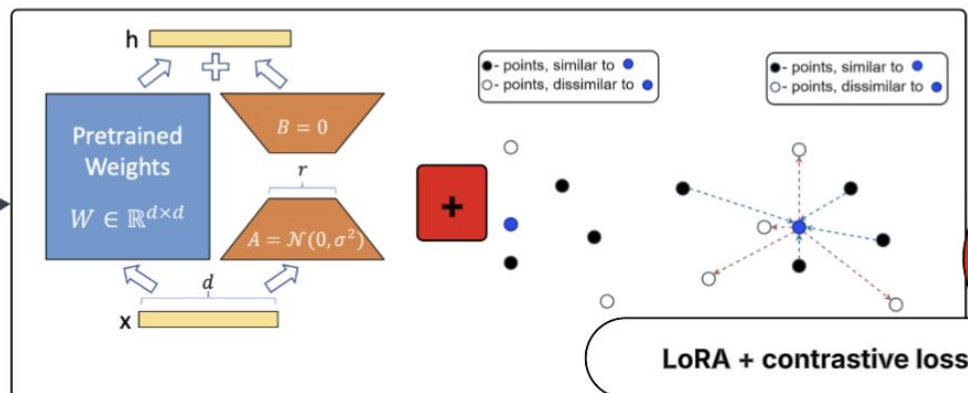
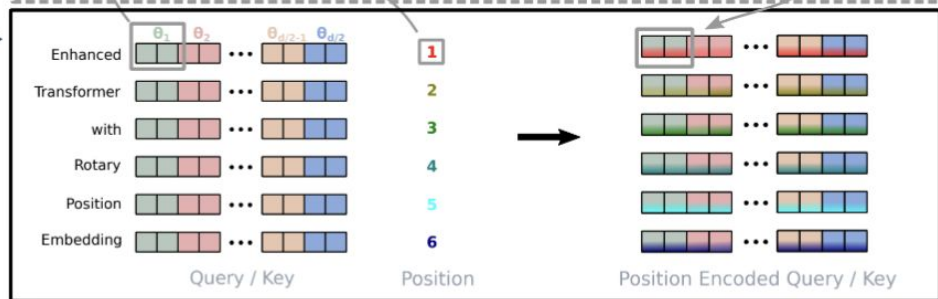
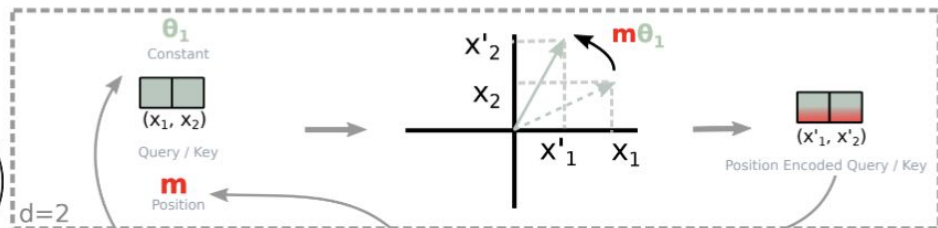
Representations

Images

τ_θ

1

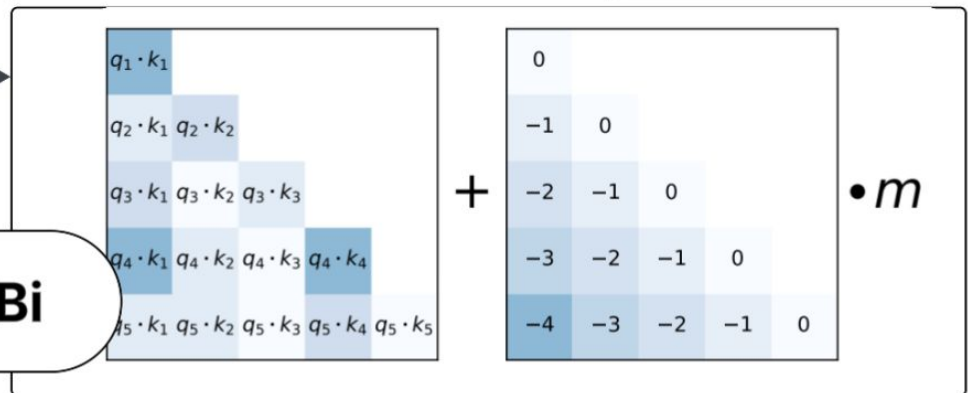
RoPE



LoRA + contrastive loss

3

ALiBi



Experiments & Results

Method	CLIP Score	Precision	Recall	F1 Score
Stable Diffusion (Baseline)	0.30	0.74	0.78	0.73
RoPE + LoRA	0.31	0.67	0.40	0.55
ALiBi + LoRA	0.32	0.84	0.60	0.71
Bounded Attention	0.32	0.71	0.30	0.42
Slot Guided Attention	0.33	1.00	0.4	0.57
Bounded + RoPE + LoRA	0.32	1.00	0.40	0.57
Slot + RoPE + LoRA	0.31	1.00	0.35	0.52
Bounded + ALiBi + LoRA	0.30	0.91	0.53	0.67
Slot + ALiBi + LoRA	0.32	0.93	0.63	0.75

Conclusion

- Stable Diffusion struggles with accurately generating multiple objects with correct counts.
- We employed and combined different techniques to improve the model's ability to generate images with better multiple object cardinality in relation to the text prompts.
- **Slot Guided Attention + ALiBi Text Embedding + Fine tuning with LoRA** outperforms the baseline with the following metrics:
 - CLIP score : 0.32
 - Precision: 0.93
 - Recall: 0.63
 - F1 score: 0.75

References (Related Work)

1. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low Rank Adaptation of Large Language Models", *arXiv preprint* arXiv:2106.09685, Jun 2021
2. Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, Yunfeng Liu, "RoFormer: Enhanced Transformer with Rotary Position Embedding", *arXiv preprint* arXiv:2104.09864, Nov 2023
3. A. K. Akan and Y. Yemez, "Slot-Guided Adaptation of Pre-trained Diffusion Models for Object-Centric Learning and Compositional Generation," *arXiv preprint* arXiv:2501.15878, Mar. 2025.
4. O. Dahary, O. Patashnik, K. Aberman, and D. Cohen-Or, "Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation," *arXiv preprint* arXiv:2403.16990, Mar. 2024.
5. Ofir Press, Noah A. Smith, Mike Lewis, "Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation", *arXiv preprint* arXiv:2108.12409, Apr 2022