Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models

Kota Sueyoshi (Osaka University), Takashi Matsubara (Hokkaido University)

CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction

Background: Diffusion Models

- generate diverse, creative, high-quality images.
- struggle to capture the intended meaning of the text for generating images.

	Missing Objects	Object Mixture	Attribute Leakage	Possession Failure
Prompts	*a yellow car and a blue bird*	*a bird and a cat*	*a green balloon and a purple clock*	*a boy grasping a soccer ball*

Related Work

- Some studies developed *attention guidance* based on the attention maps of the cross-attention layers.
- Existing methods are specialized to limited types of user's intentions contained in the text.
  1. Attend-and-Excite only for missing objects[1]
  2. SynGen only for attribute leakage[2]
  3. Nothing has focused on possession failure.

References

[1] Chefer *et al*. "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models." In: SIGGRAPH2023

[2] Rassin *et al*. "Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment" In: NeurIPS2023

Method : Predicated Diffusion

Predicated Diffusion

- is a unified framework to effectively express users' intentions.
- represents the intended meaning as propositions using predicate logic.
- treats pixels $A_P[i]$ of attention map A_P for word P as fuzzy propositions $P(x)$.

Proposition	Attention Map
true	1
false	0
$P(x)$	$A_P[i]$
$\neg P(x)$	$1 - A_P[i]$
$P(x) \wedge Q(x)$	$A_P[i] \times A_Q[i]$
$P(x) \rightarrow Q(x)$	$1 - A_P[i] \times (1 - A_Q[i])$
$P(x) \vee Q(y)$	$1 - (1 - A_P[i]) \times (1 - A_Q[j])$
$\forall x. P(x)$	$\prod_i A_P[i]$
$\exists x. P(x)$	$1 - \prod_i (1 - A_P[i])$

Algorithm

(1) Deduce statements & represent them by propositions

- manually or using a syntactic dependency parser.

exist possess exist exist

A white teddy bear with a green shirt and a smiling girl

modify modify modify

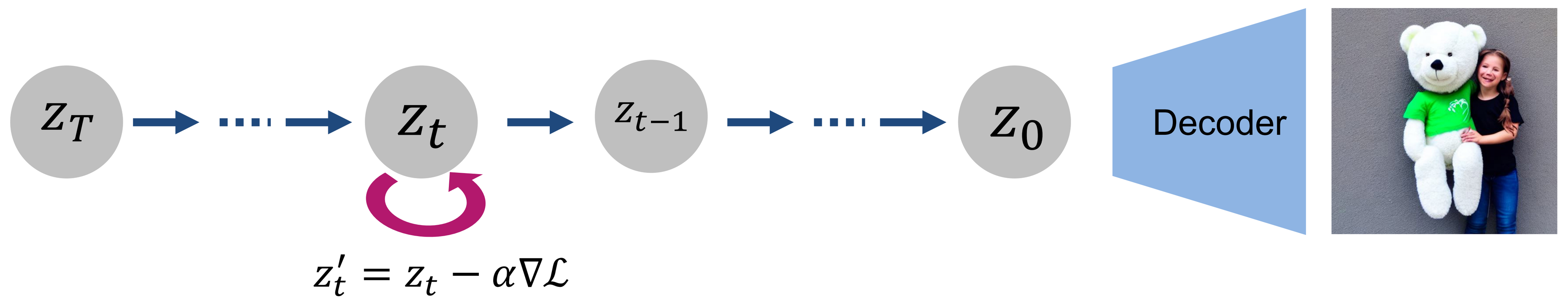
(2) Obtain loss functions on attention maps

- existence: $\mathcal{L}[\exists x.\ Girl(x)] = -\log\left(1 - \prod_i (1 - A_{Girl}[i])\right)$
- modification: $\mathcal{L}[\forall x.\ White(x) \leftrightarrow Bear(x)]$
  $= \mathcal{L}[\forall x.\ White(x) \rightarrow Bear(x)\ ] + \mathcal{L}[\forall x. Bear(x) \rightarrow White(x)]$
- possession: $\mathcal{L}[\forall x.\ Shirt(x) \rightarrow Bear(x)] = -\sum_i \log\left(1 - A_{Shirt}[i] \times (1 - A_{Bear}[i])\right)$

⋮

(3) Run a diffusion model & get an image faithful to the prompt

- The latent variable Z is updated to minimize the loss function, ensuring the generated images to be faithful to the intent of the text.



Experimental Results

Methods	Human Evaluation: Missing Objects ↓	Human Evaluation: Attribute Leakage ↓	Human Evaluation: Fidelity ↑	Automatic Evaluation: Similarity ↑	Automatic Evaluation: CLIP-IQA ↑
Stable Diffusion	64.8/73.5	88.5	6.0	0.345/0.744	0.756
Composable Diffusion	49.3/83.5	88.5	3.8	0.348/0.729	0.757
Structure Diffusion	64.3/69.5	86.5	5.8	0.346/0.741	0.760
Attend-and-Excite	28.0/35.8	64.5	19.3	0.367/0.792	0.761
SynGen	23.3/29.3	40.3	36.8	0.367/0.801	0.750
Predicated Diffusion	**10.0/16.5**	**33.0**	**44.8**	**0.379/0.811**	**0.769**

Methods	Human Evaluation: Missing Objects ↓	Human Evaluation: Possession Failure ↓	Human Evaluation: Fidelity ↑	Automatic Evaluation: Similarity ↑	Automatic Evaluation: CLIP-IQA ↑
Stable Diffusion	31.5/36.0	52.5	33.5	0.320/0.811	0.762
Attend-and-Excite	7.5/17.0	51.5	27.5	0.334/0.843	0.760
Predicated Diffusion	**4.0/7.0**	**29.5**	**52.0**	**0.345/0.855**	**0.769**

Missing Objects and Attribute Leakage

Possession Failure

Complicated Prompts