

Imagine That!!

Abstract-to-Intricate Text-to-Image Synthesis with Scene Graph Hallucination Diffusion

Shengqiong Wu¹, Hao Fei¹, Hanwang Zhang², Tat-Seng Chua¹

¹NExT++, School of Computing, National University of Singapore

² School of Computer Science and Engineering, Nanyang Technological University



Scan me!

Motivation

■ Text-to-image generation: generating images from natural language descriptions

✓ Existing issues:

- *Inherent gap between text and image*
 - *concise and abstract words can depict intricate and complex visual scenes*
- *Unwilling to take time to write detailed descriptions*

✓ Scene enrichment is needed

classroom



Motivation

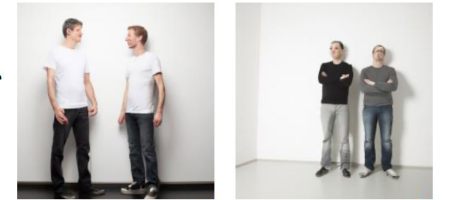
Text-to-image generation: generating images from natural language descriptions

✓ Issues:

- *Vision Distraction*
- *Wrong Binding*

► Original (*abstract*) Prompt:

Two young men give a presentation in the office.



Issue: Abstract-to-intricate Failure

► Enriched Prompt-I:

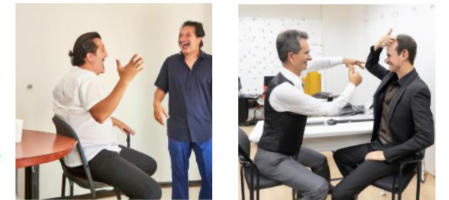
Two middle-aged, nice, enthusiastic, confident, man with polished shoes and sleek hair give a professional presentation in the spacious and modern conference room of the corporate blue office, room with chairs and tables.



Issue: Vision Distraction

► Enriched Prompt-II:

Two young man give a presentation in the office, old, nice, confident, enthusiastic, laughing man with polished hair, seek hair, room with chairs and tables, speaking to each other.



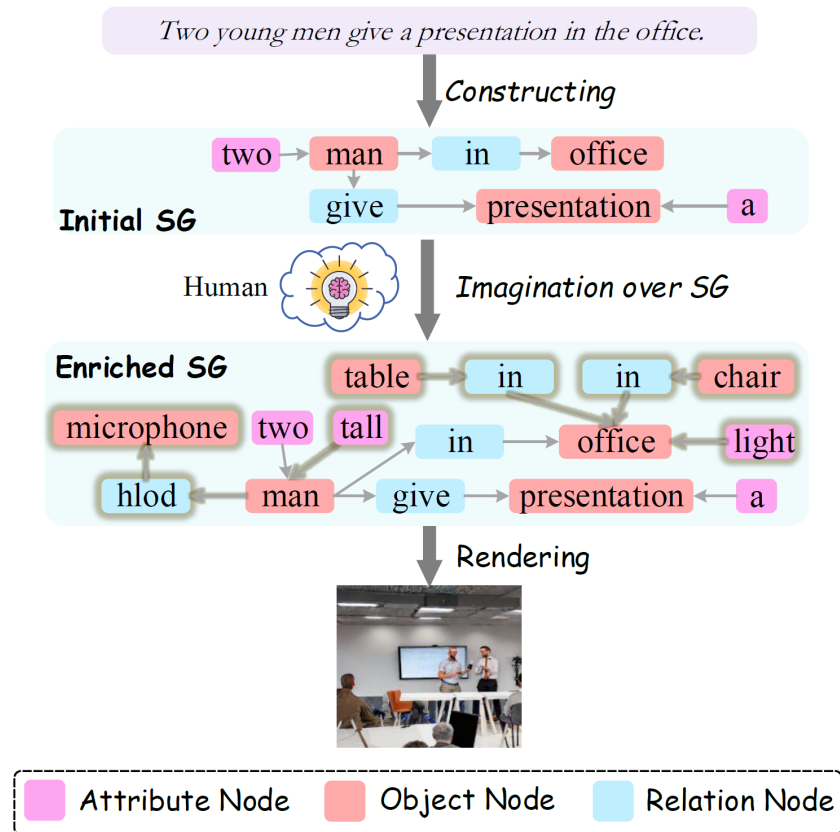
Issue: Wrong Binding

Motivation

■ **Text-to-image generation:** generating images from natural language descriptions

Scene Graph-based Scene enrichment

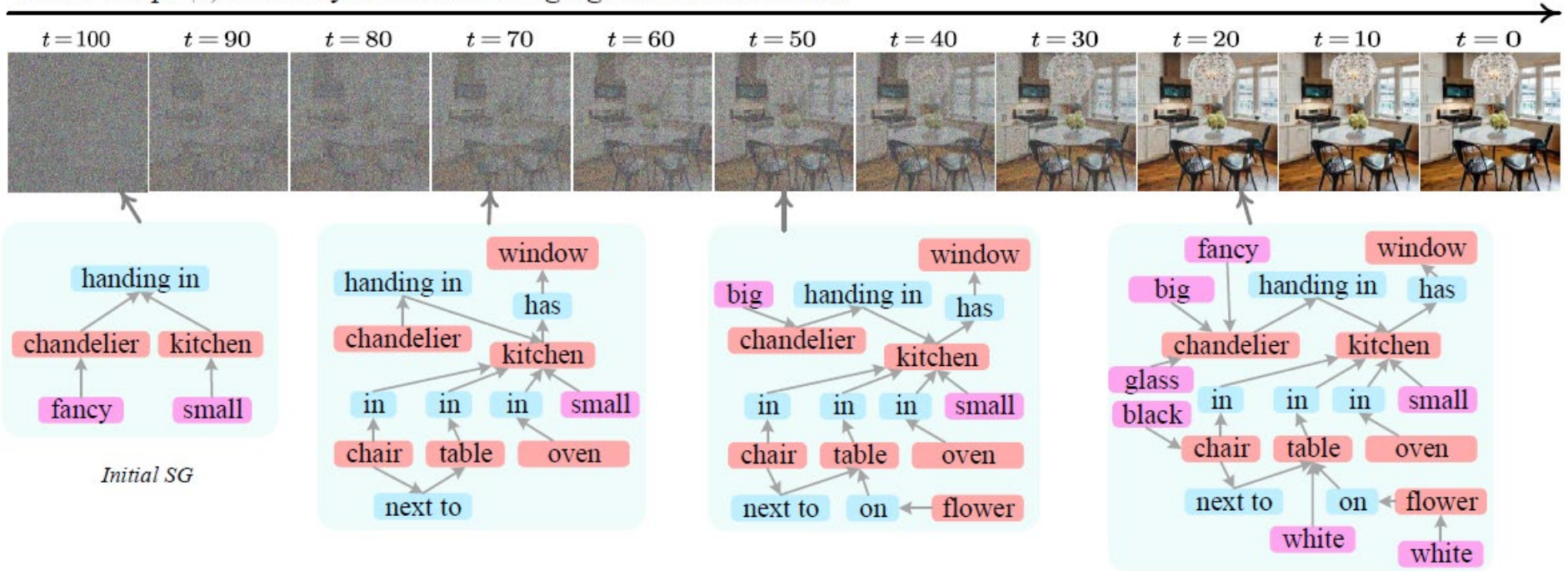
- ✓ Advantages:
- *SG advances in depicting the intrinsic semantics of texts (or vision) with structured representations*
 - *the enrichment process can be much more accurate and controllable*



Method

SG-based hallucination diffusion system

Text Prompt (a): A fancy chandelier hanging in a small kitchen.



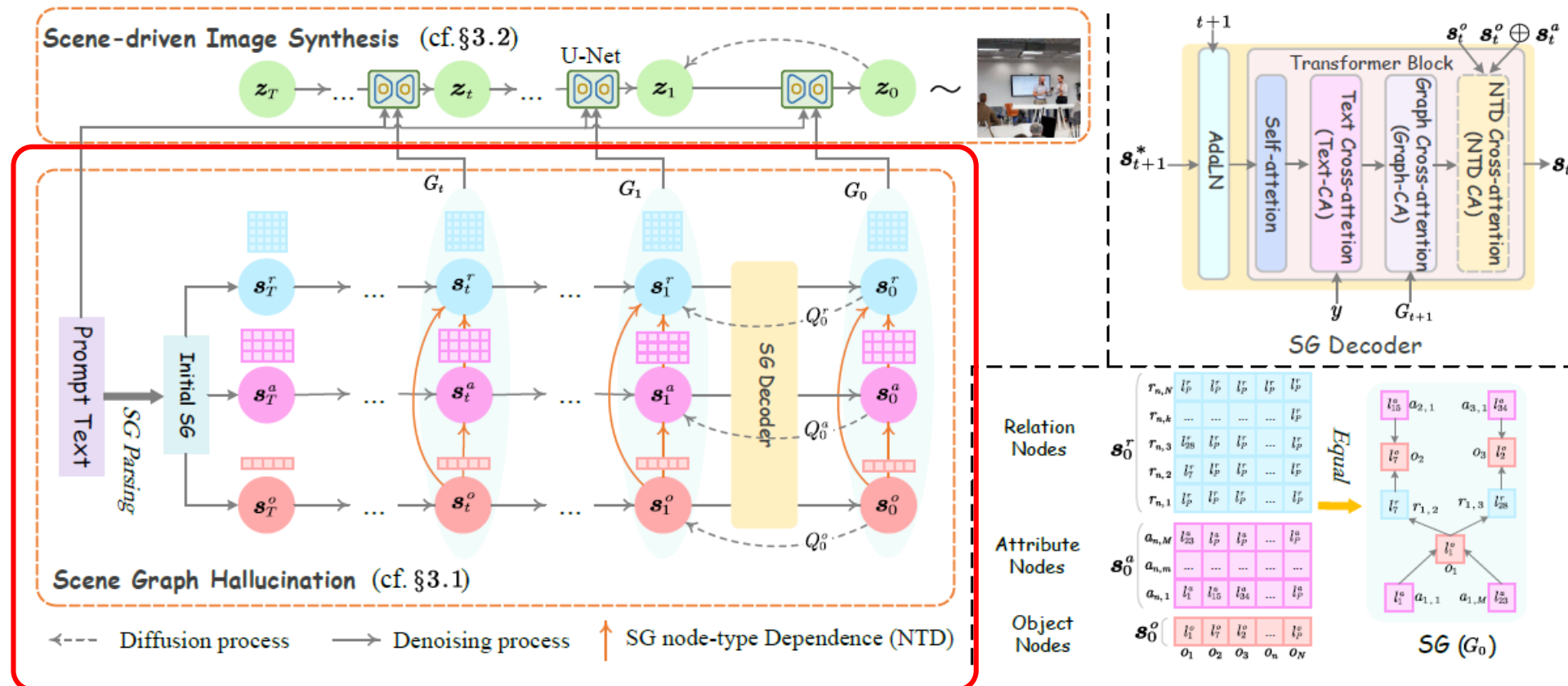
T2I

scene-graph hallucination

Method

SG-based hallucination diffusion system

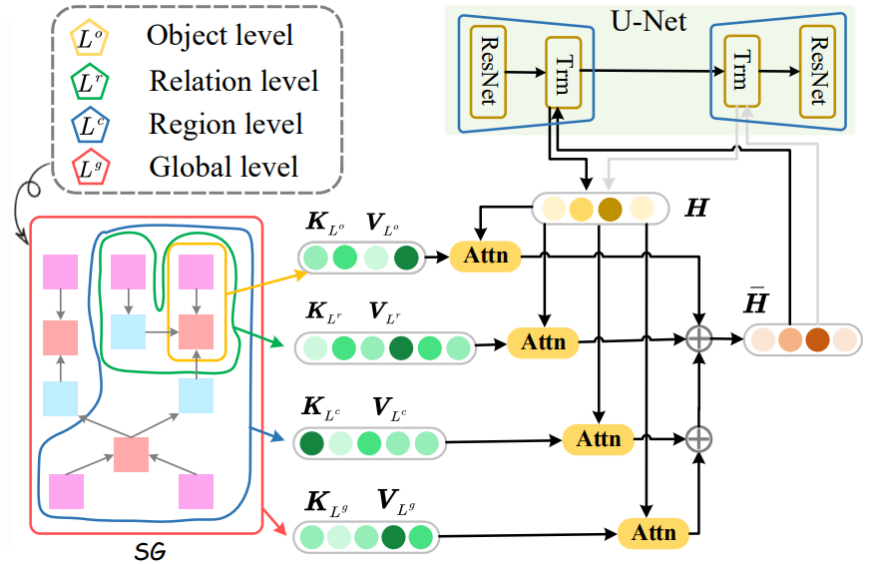
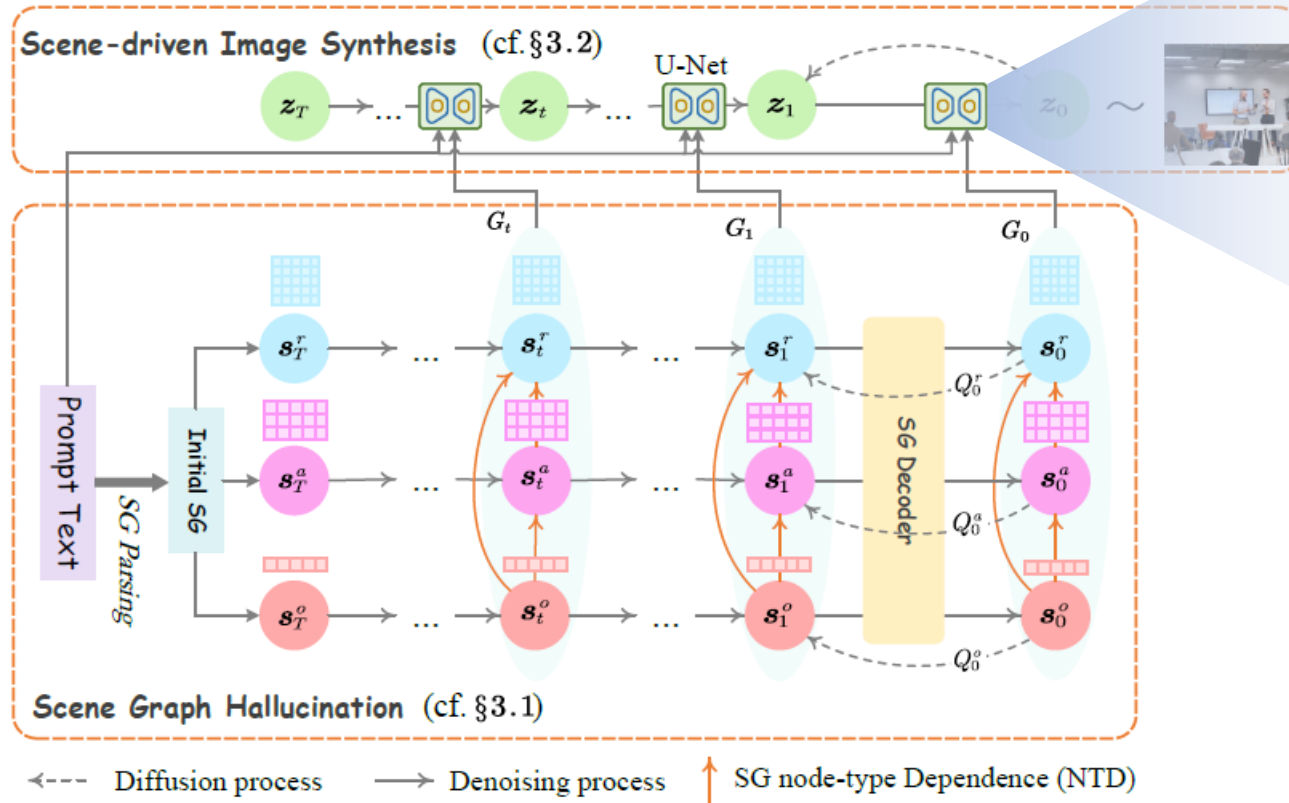
➤ Scene Graph Hallucination



Method

SG-based hallucination diffusion system

➤ Scene-driven Image Synthesis



Experiments

Main Comparison

Table 1: The T2I results on the overall COCO dataset. ^b: taken from Fan et al. [15]; ^h: copied from Ding et al. [11], [‡]: taken from Hinz et al. [23]. The best score is in bold, and the second best is underlined.


Model	FID ↓	IS ↑	CLIP ↑
► GAN Model			
AttnGAN ^b	33.10	26.61	-
ObjGAN ^b	36.52	24.09	-
DFGAN ^b	21.42	-	-
OPGAN [‡]	24.70	<u>27.88</u>	-
► Auto-aggressive Model			
DALLE ^h	27.34	17.90	-
CogView ^h	27.10	18.20	-
► Diffusion Model			
LDM ^b	17.61	19.34	65.00
VQ-diffusion ^b	14.06	21.85	67.70
LDM-G ^b	12.27	27.86	69.27
Frido ^b	11.24	26.84	70.46
Salad	10.19	29.96	74.83

Table 2: Results on the COCO-A2I subset for the abstract-to-intricate T2I generation. ‘SPY[†]’ denotes enriched texts are parsed into SG, and then perform SG-to-image generation via Frido.

Model	FID ↓	IS ↑	CLIP ↑
► T2I Baseline			
AttnGAN	78.19	11.09	52.78
ObjGAN	75.33	13.16	55.20
DFGAN	71.24	15.56	56.91
DALLE	66.36	16.03	63.05
CogView	62.85	16.98	63.97
LDM	55.27	16.20	67.79
VQdiffusion	69.14	15.78	64.58
Frido	40.36	18.36	68.53
► Text-based Enrichment (+Frido)			
SD-PG	36.50	17.64	65.23
SPY	<u>35.59</u>	21.59	<u>67.86</u>
SPY [†]	39.41	<u>22.16</u>	66.93
Salad	31.25	28.63	71.29

Experiments

Qualitative Results

	Frido	SPY	Salad
a man sits in the bank			 <pre>graph LR; man[man] -- sit in --> bank[bank]; man -- sit on --> chair[chair]; bank -- in --> windows[windows]; windows -- in front of --> bank;</pre>
a man is travelling			 <pre>graph LR; man[man] -- wear --> Tshirt[Tshirt]; man -- wear --> sunglasses[sunglass]; man -- carrying on --> bag[luggage bag]; man -- in front of --> building[building]; Tshirt -- black --> sunglasses; bag -- young --> man; building -- old --> man;</pre>
a man sits in the kitchen			 <pre>graph LR; man[man] -- sit on --> chair[chair]; man -- next to --> table[table]; man -- next to --> fridge[fridge]; chair -- in --> kitchen[kitchen]; table -- in --> kitchen; fridge -- next to --> microwave[microwave];</pre>

Take-away

- This work focuses on investigating to **generate intricate images from abstract prompts**.
- We solve the abstract-to-intricate T2I with a **discrete diffusion-based SG hallucination mechanism**, enriching the scene with reasonable imagination.
- We propose a diffusion-based model with a **hierarchical scene integration** strategy for highly controllable and scalable image generation.
- Our framework achieves new SoTA results in the abstract-to-intricate T2I generation.