Synergistic Dual Spatial-aware Generation of Image-to-text and Text-to-image

Yu Zhao, Hao Fei, Xiangtai Li, Libo Qin, Jiayi Ji,
Hongyuan Zhu, Meishan Zhang, Min Zhang, Jianguo Wei

Visual Spatial Understanding

- **Definition:**

  Extracting Spatial Information, including position, relation, pose, layout, etc., reasoning, and applying to special tasks (SLAM, navigation, etc.)

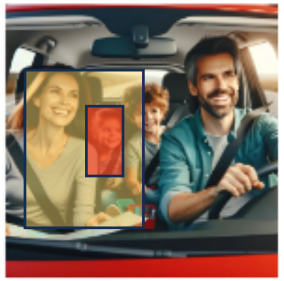
- **Forms of Tasks:**

  Relation Extraction
  Question Answering
  Image-to-Text Generation, Captioning
  Image Synthesis
  3D Reconstruction
  …

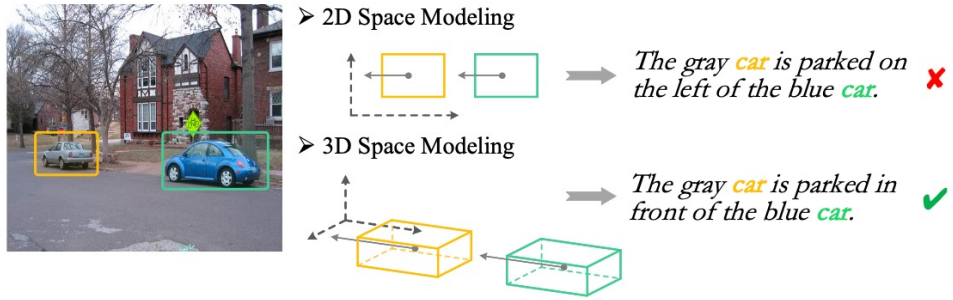
We study the representative spatial image-to-text (SI2T) and spatial text-to-image (ST2I)

Motivation

SI2T



Overlap



Perspective Illusion

ST2I

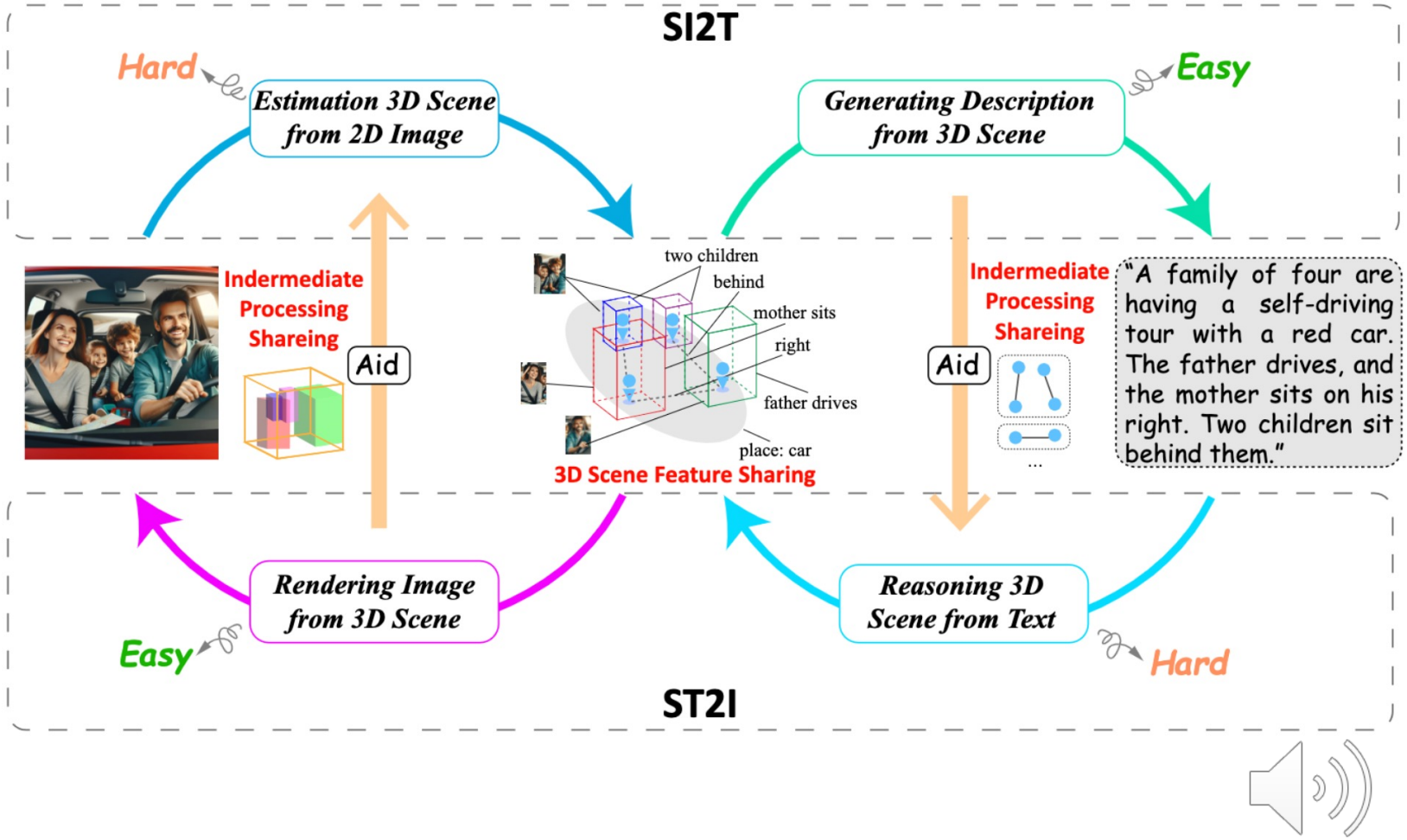
You

画一张图有一张桌子、一把椅子，一本书，椅子放在在桌子的左侧，书放在椅子上，椅背对着桌子

ChatGPT

Consider the Two Tasks Together

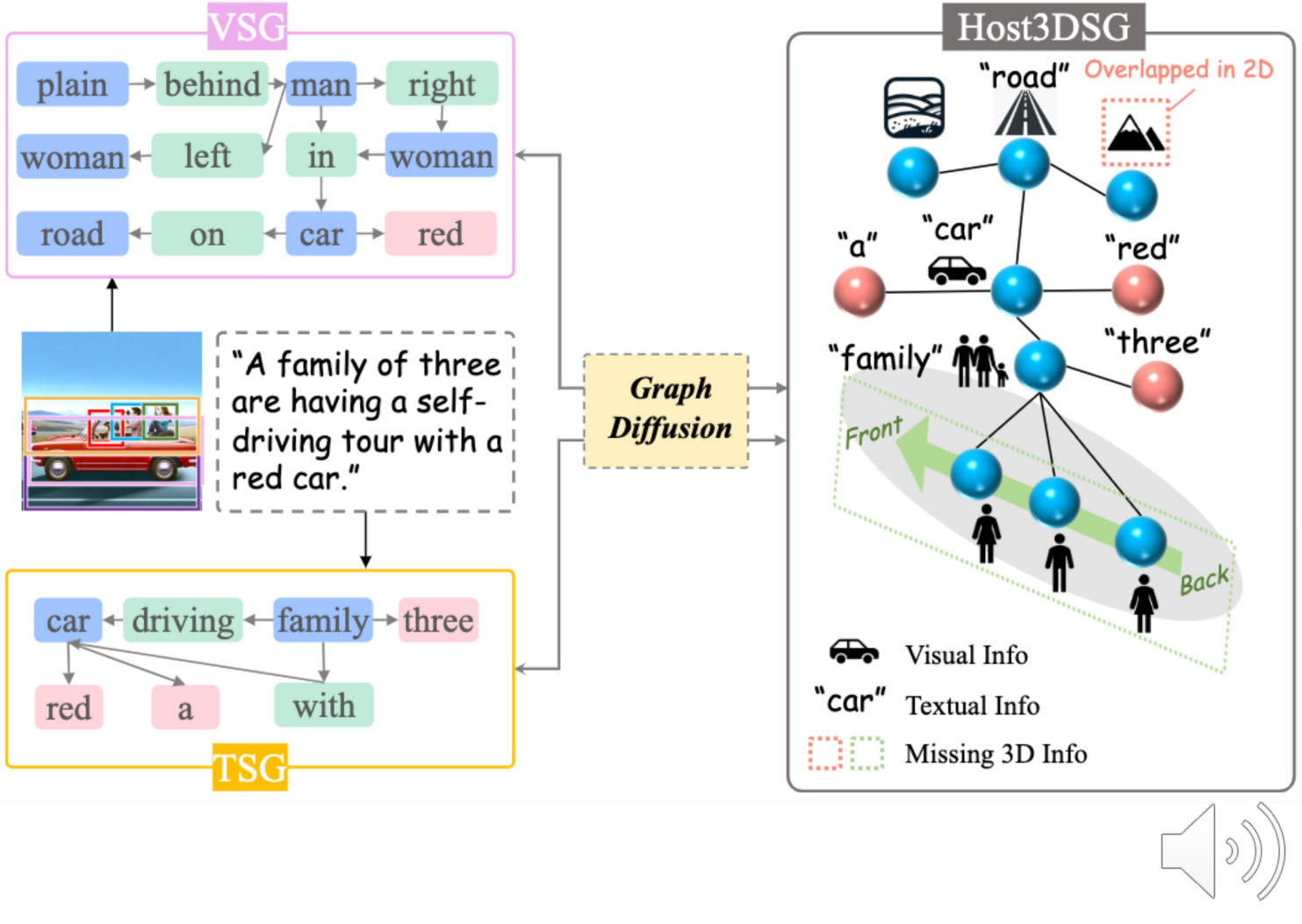
- **Dual for each other**
- **Share the 3D scene feature:**
- **The dual processes help each other**



天津大学 Tianjin University 新媒体与传播学院 School of New Media and Communication

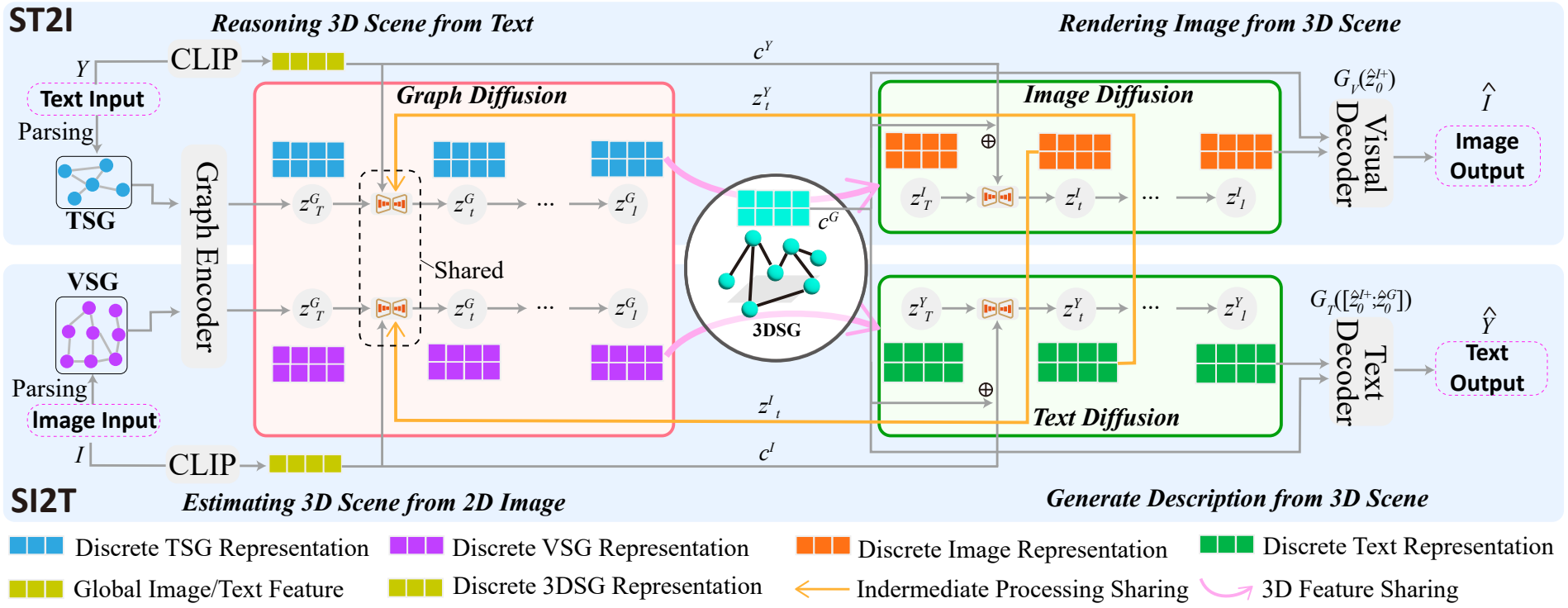
Model the Shared 3D Feature

- **Share a Holistic 3D Scene Graph**
- **Generating via a Graph Diffusion**
- **Initialized from Visual Scene Graph or Textual Scene Graph**



Dual Learning Framework

- **Three Diffusion Module**
- **Dual Training**



Training Objectives

- **Diffusion**

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \quad p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \quad L = \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right]$$

- **Spatial Feature Alignment**

  Train the Image Decoder and Text Decoder

$$\mathcal{L}_{v-rec} = ||I - G_V(z_0^{I+})||^2 \qquad \mathcal{L}_{t-dec} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, [z_0^Y; z_0^G])$$

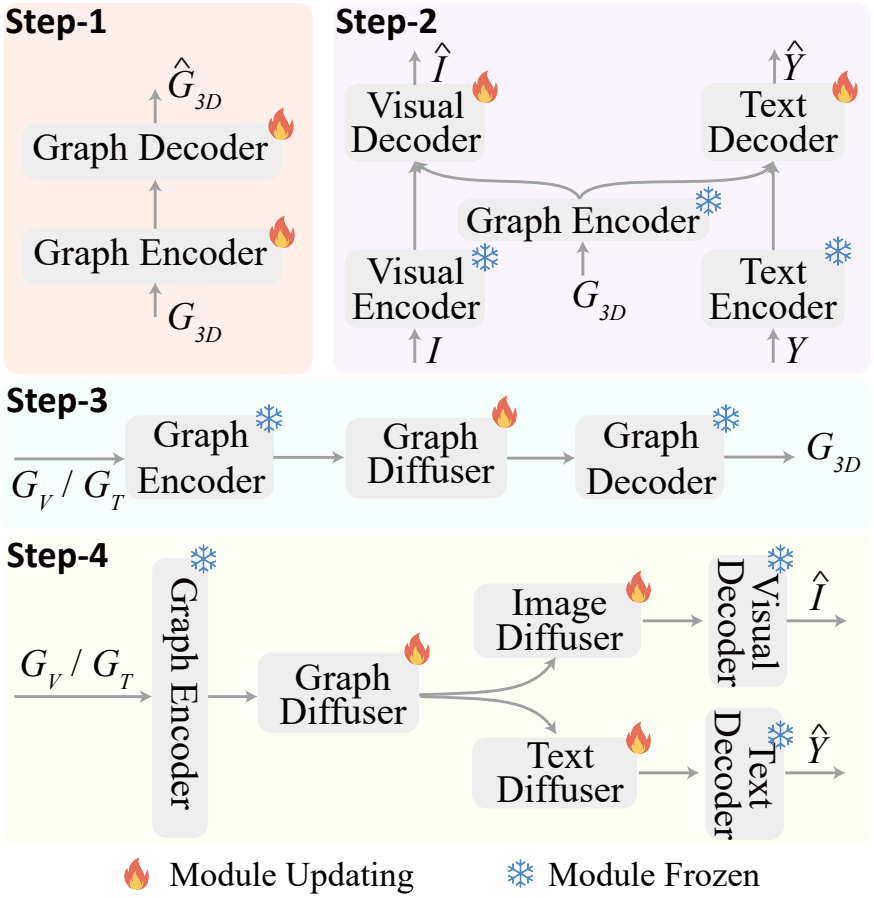
- **Host3DSG Reconstruction Training**

$$\mathcal{L}_{DGAE} = -\mathbb{E}_{\hat{\mathcal{Z}}^G} \ln(p(G_{host}|\hat{\mathcal{Z}}^G)),$$

天津大学 Tianjin University 新媒体与传播学院 School of New Media and Communication

Training Strategy

- **Step-1: DGAE pre-training**
- **Step-2: Spatial Alignment**
- **Step-3: 2DSG→3DSG Diffusion Training**
- **Step-4: Overall Training**



Main Results

	VSDv1					VSDv2				
	ST2I			SI2T		ST2I			SI2T	
	FID↓	IS↑	CLIP↑	BLEU4↑	SPICE↑	FID↓	IS↑	CLIP↑	BLEU4↑	SPICE↑
• **T2I Baselines**										
DALLE [62]	32.55	17.01	62.16	-	-	28.52	21.18	64.58	-	-
Cogview [13]	32.30	17.07	61.85	-	-	28.17	21.74	64.76	-	-
LAFITE [98]	30.73	24.39		-	-	25.73	25.47		-	-
VQ-Diffusion [28]	18.34	20.58	63.42	-	-	15.66	24.75	66.30	-	-
Friodo [15]	12.86	25.92	64.65	-	-	11.41	26.02	67.01	-	-
• **I2T Baselines**										
3DVSD [97]	-	-	-	54.85	68.76	-	-	-	26.40	46.97
MNIC [25]	-	-	-	34.21	66.87	-	-	-	20.01	43.88
FNIC [23]	-	-	-	37.03	66.50	-	-	-	22.62	43.52
DiffCap [29]	-	-	-	34.75	66.39	-	-	-	20.27	43.30
DDCap [99]	-	-	-	37.93	67.10	-	-	-	23.14	44.07
Singleton	18.05	20.42	63.51	48.77	66.59	14.70	24.62	66.41	23.51	43.70
Singleton + 3D	12.56	26.92	65.62	50.05	67.20	10.43	25.62	67.29	25.37	45.13
Vanilla Dual Learning	11.80	27.85	67.18	51.59	67.79	11.67	27.80	68.46	26.10	46.72
SD^3 (Ours)	**11.04**	**29.20**	**68.31**	**56.23**	**68.02**	**10.09**	**29.76**	**71.10**	**27.63**	**48.03**

Results

Cases

GT Image

GT Text

A man is on the couch next a dog laying down.

VQ-Diffusion

DDCap

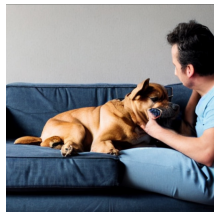
A man sitting on a chair with a dog on his lap.

Frido

3DVSD

A man is sitting on the sofa.

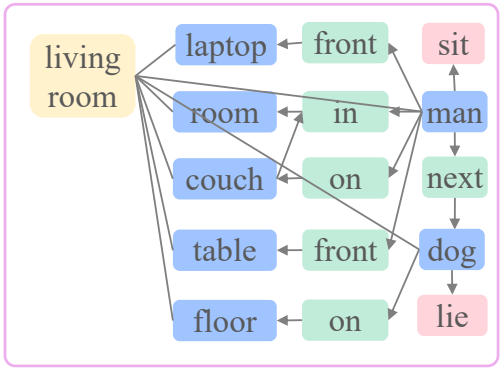
SD³



SD³

A man is sitting on the couch with a dog aside.

Generated 3DSG



GT Image

GT Text

Some weeds in the middle are in front of the railway tracks.

VQ-Diffusion

DDCap

A train traveling down tracks next to a forest.

Frido

3DVSD

Some weeds lie on the right of the railway tracks.

SD³

SD³

Some weeds grow in front of the railway tracks.

Generated 3DSG