InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models
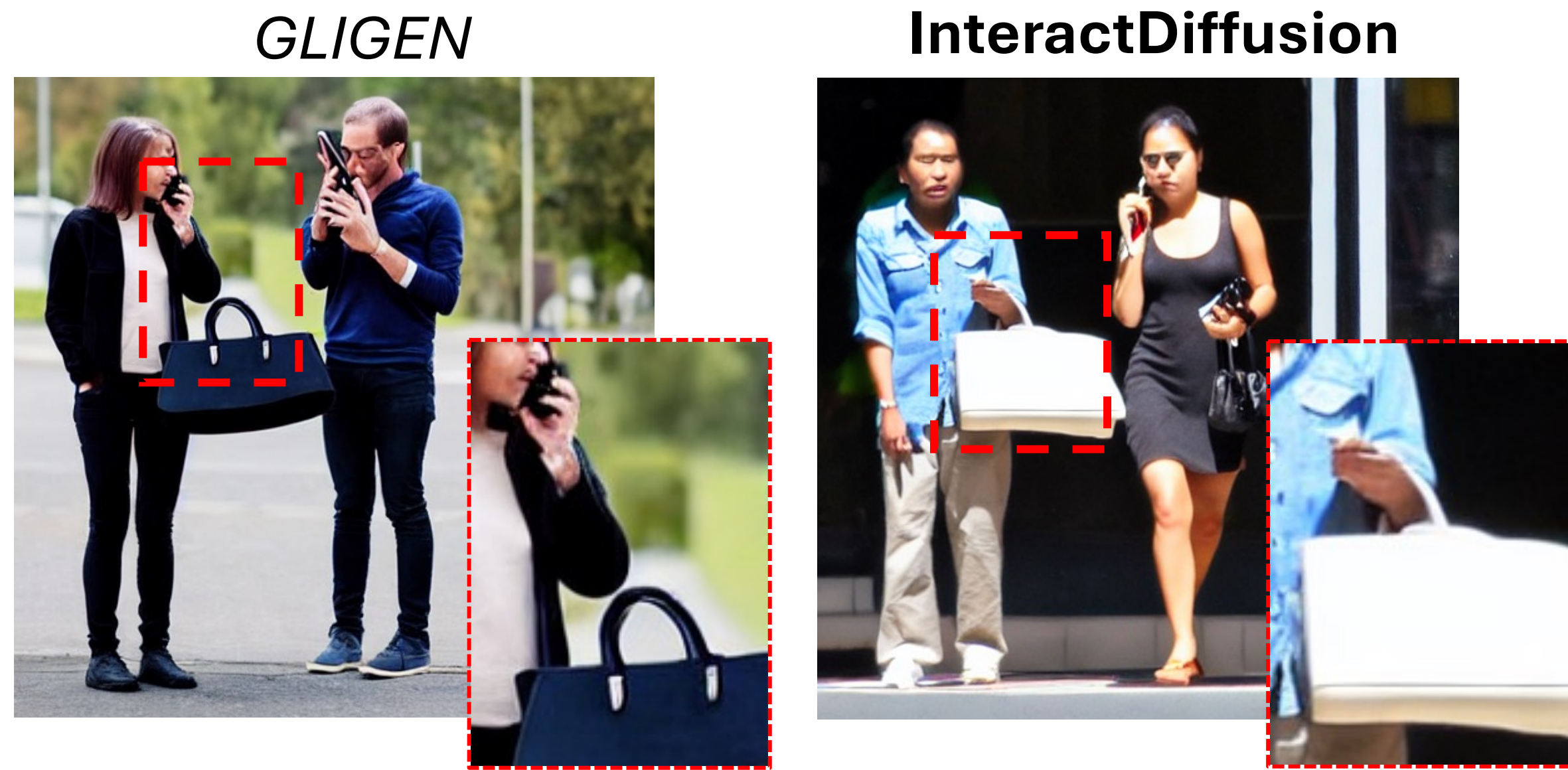
Jiun Tian Hoe · Xudong Jiang · Chee Seng Chan · Yap-Peng Tan · Weipeng Hu

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · UNIVERSITY OF MALAYA · CVPR SEATTLE, WA JUNE 17-21, 2024

Problem

GLIGEN   **InteractDiffusion**

- **Interaction** is an integral part in describing our daily activities
- Previous controlled image generation methods such as *GLIGEN* did not consider the interaction relationship
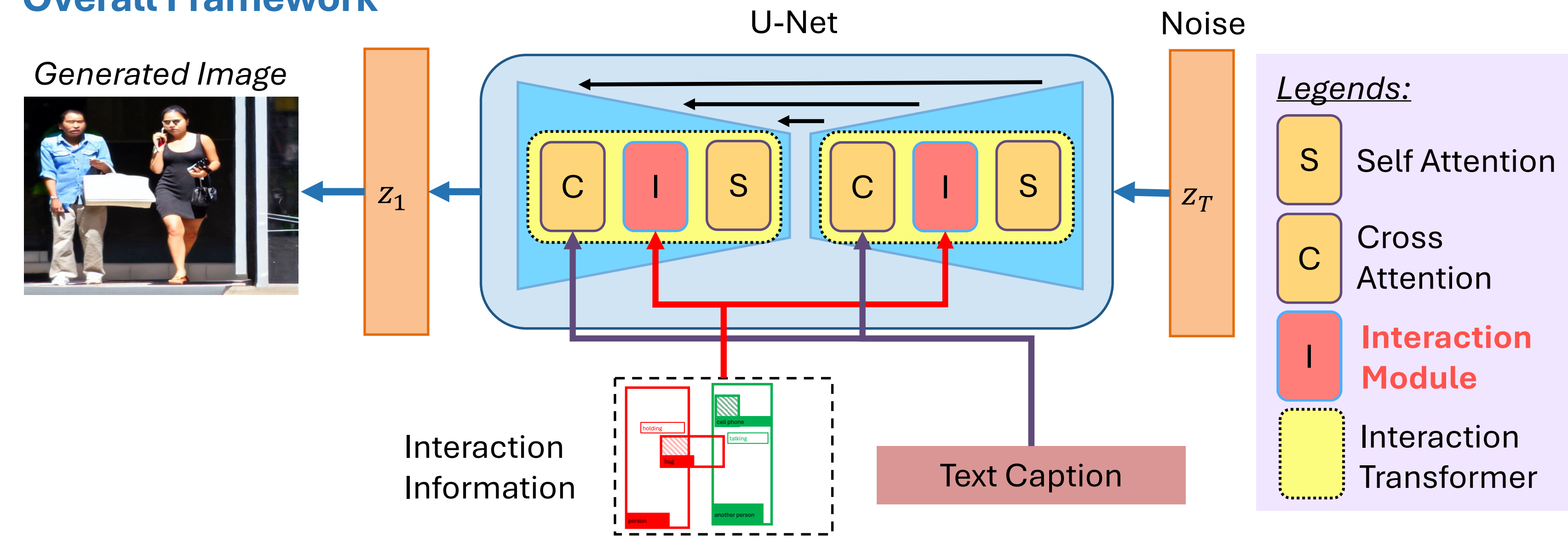
IDEA: Interaction as a Condition in T2I Models

Interaction Input   Caption   *Generated Image*

"a person is feeding a cat"

InteractDiffusion
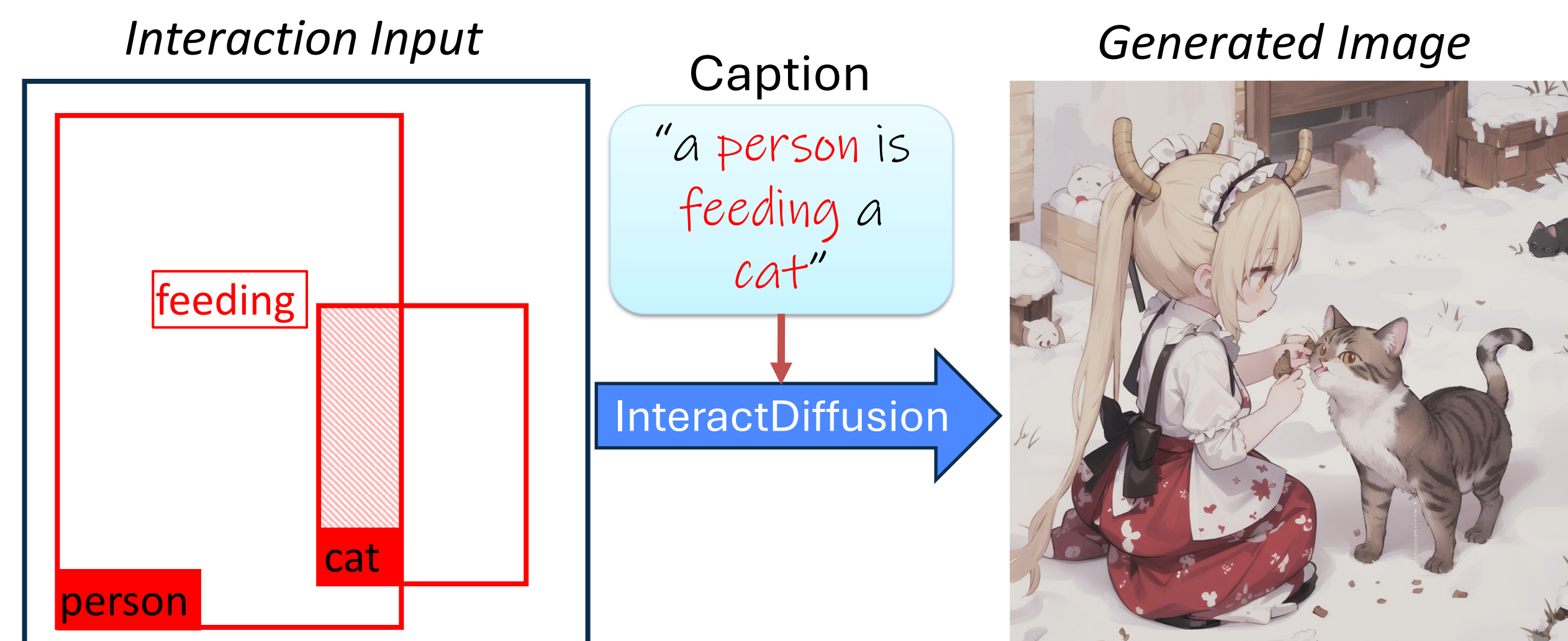
Specifying **interaction** as extra conditions faces multiple challenges:
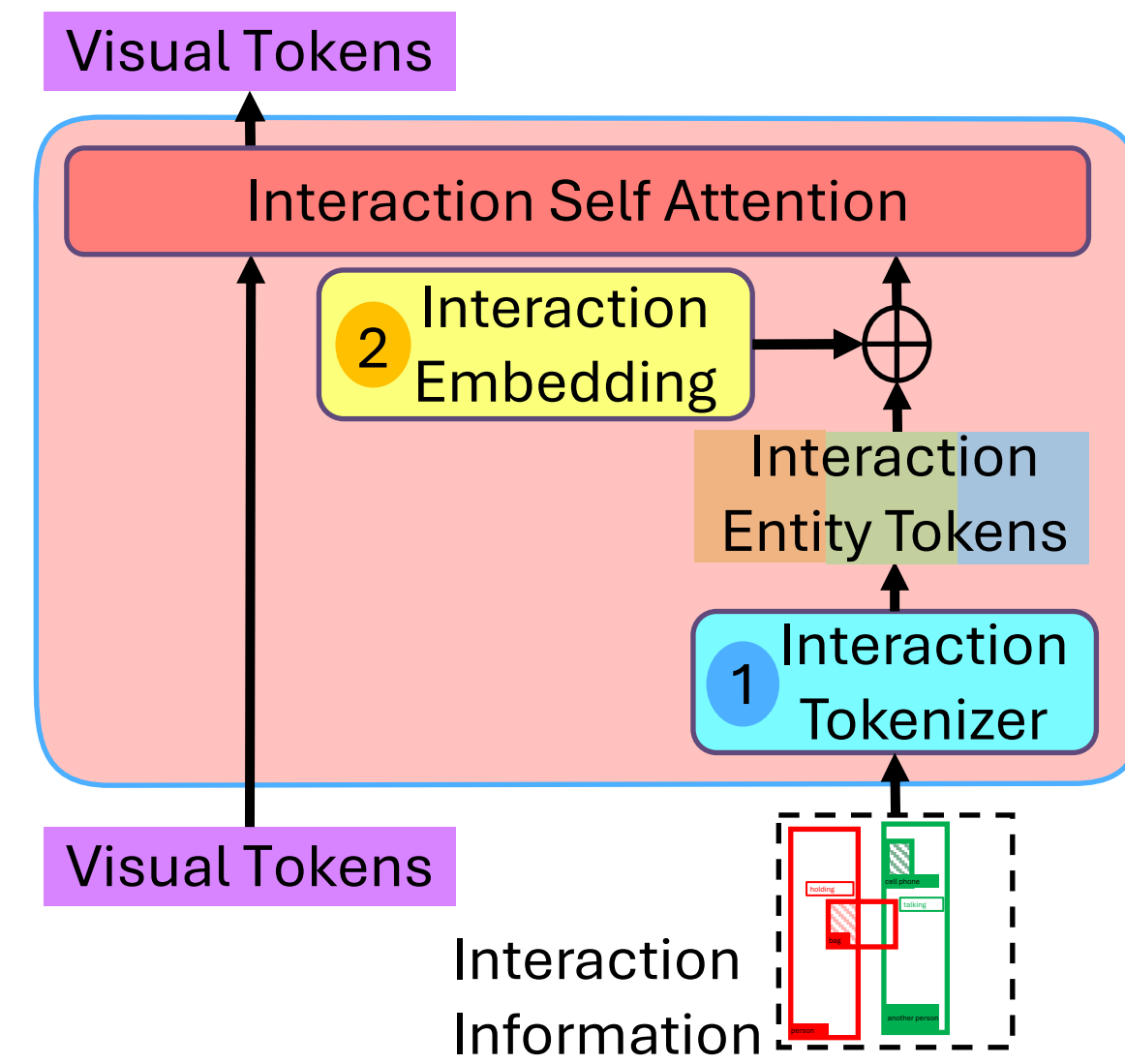- i. Interaction representation
- ii. Intricate interaction relationship
- iii. Integrate into existing models
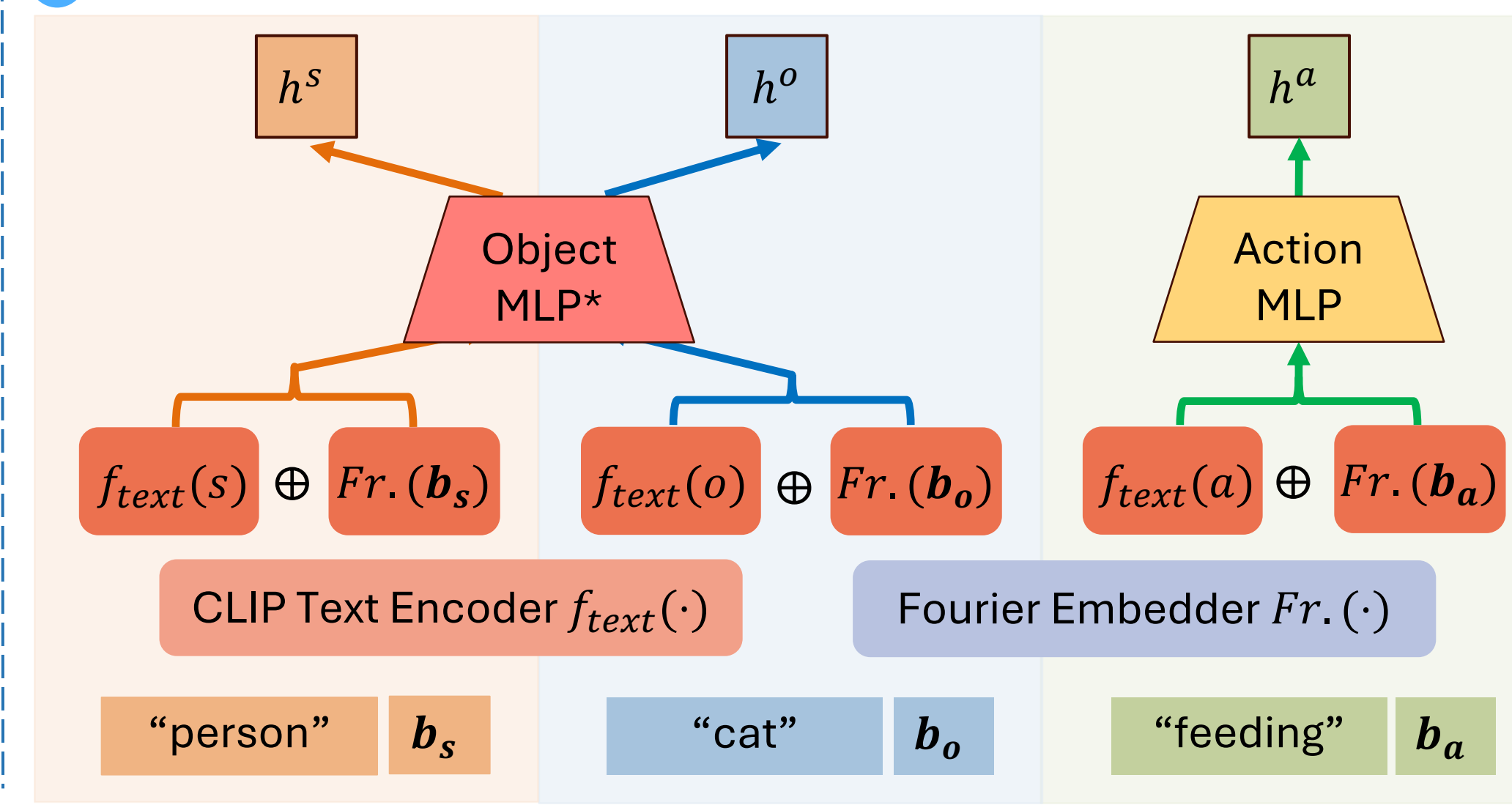
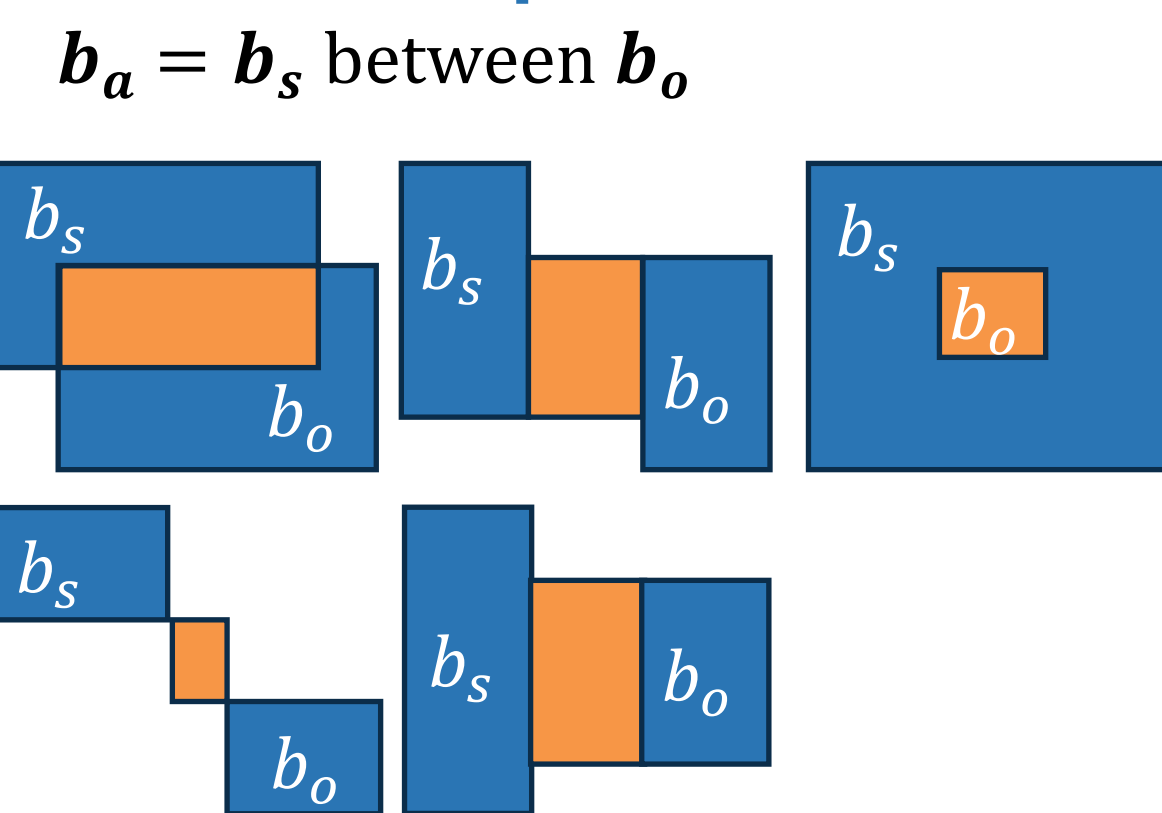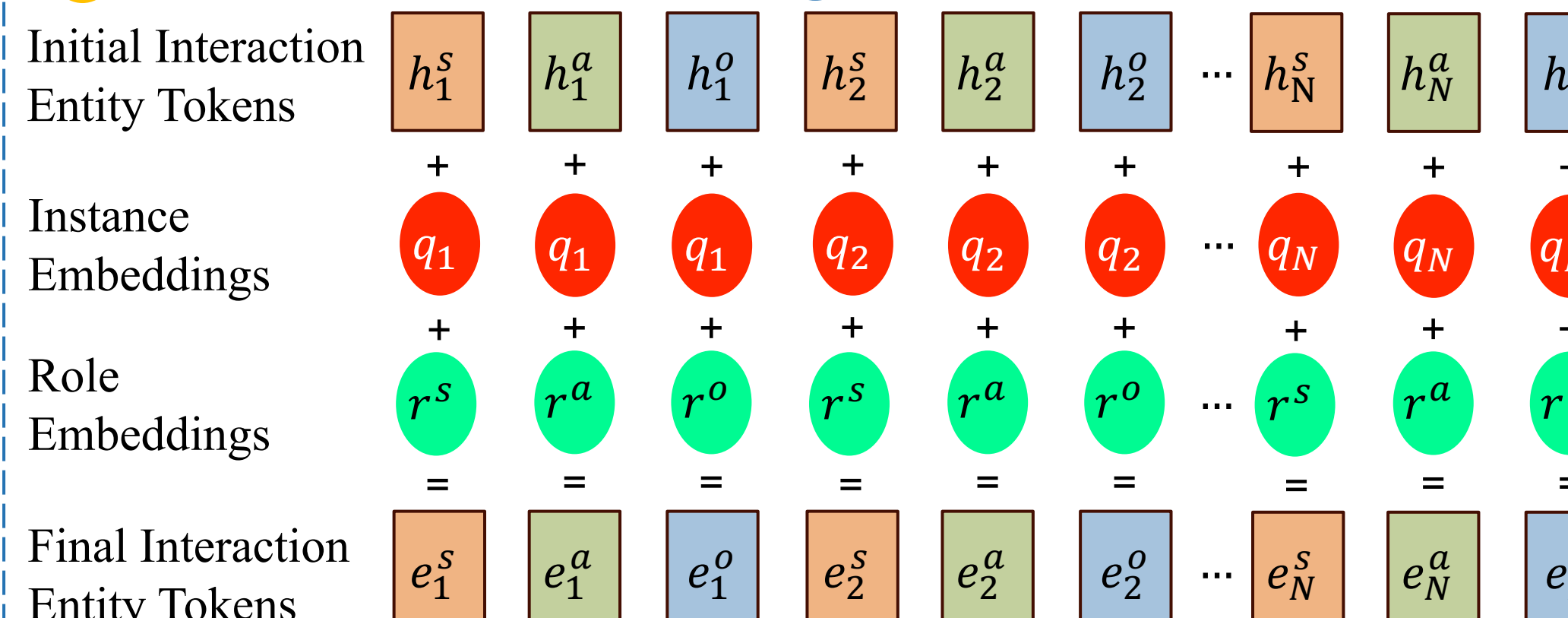## Our Proposed: InteractDiffusion

Overall Framework

Generated Image — U-Net — Noise

z_1 C I S   C I S z_T

Legends:
- S — Self Attention
- C — Cross Attention
- I — **Interaction Module**
- — Interaction Transformer

Interaction Information   Text Caption

Interaction Module

Visual Tokens

Interaction Self Attention

② Interaction Embedding ⊕

Interaction Entity Tokens

① Interaction Tokenizer

Visual Tokens

Interaction Information

① Interaction Tokenizer

h^s h^o h^a

Object MLP*   Action MLP

$f_{text}(s) \oplus Fr.(b_s)$ $f_{text}(o) \oplus Fr.(b_o)$ $f_{text}(a) \oplus Fr.(b_a)$

CLIP Text Encoder $f_{text}(\cdot)$ Fourier Embedder $Fr.(\cdot)$

"person" b_s "cat" b_o "feeding" b_a

"*Between*" Operation

$b_a = b_s$ between b_o

b_s / b_o

② Interaction Embedding

Initial Interaction Entity Tokens	h_1^s	h_1^a	h_1^o	h_2^s	h_2^a	h_2^o	...	h_N^s	h_N^a	h_N^o
Instance Embeddings	q_1	q_1	q_1	q_2	q_2	q_2	...	q_N	q_N	q_N
Role Embeddings	r^s	r^a	r^o	r^s	r^a	r^o	...	r^s	r^a	r^o
Final Interaction Entity Tokens	e_1^s	e_1^a	e_1^o	e_2^s	e_2^a	e_2^o	...	e_N^s	e_N^a	e_N^o

Results

Comparison with Existing Methods

Model	Quality ↓		Interaction Controllability: FGAHOI (mAP) ↑	
	FID	KID	Swin-Tiny	Swin-Large
StableDiffusion	35.85	0.01297	0.63	0.64
GLIGEN	29.35	0.01275	21.73	23.99
GLIGEN*	18.82	0.00694	25.23	26.45
InteractDiffusion	**18.69**	**0.00676**	**29.53**	**31.56**
HICO-DET	-	-	29.94	37.18

- We generate images based on ground-truth in HICO-DET, and perform evaluation using FGAHOI, a state-of-the-art HOI (Human-Object Interaction) detector
- This quantifies the interaction controllability in interaction generation

Controlling with Different Actions

a person is <*> a bottle

licking   opening   pouring

drinking   holding   inspecting