NEURAL INFORMATION PROCESSING SYSTEMS

东方理工高等研究院
EASTERN INSTITUTE FOR ADVANCED STUDY

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Scene Graph Disentanglement and Composition for Generalizable Complex Image Generation

Yunnan Wang[1,2] Ziqiang Li[1,2] Wenyao Zhang[1,2] Zequn Zhang[2,3]
Baao Xie[2] Xihui Liu[4] Wenjun Zeng[2] Xin Jin[2]

[1]MoE Key Laboratory of Artificial Intelligence, Shanghai Jiao Tong University
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology
[3]School of Information Science and Technology, University of Science and Technology of China
[4]Department of Electrical and Electronic Engineering, The University of Hong Kong

Introduction

Relationship and Quantity Confusion of T2I

(a) DALL·E 3 (middle) and Ours (right).

Non-Spatial Interaction Dilemma of L2I

(b) LayoutDiffusion (middle) and Ours (right).

Independent Nodes Missing of Semantics-based SG2I

(c) R3CD (middle) and Ours (right).

1. Challenges of Text-to-Image (T2I)

- Relationships and quantities of objects

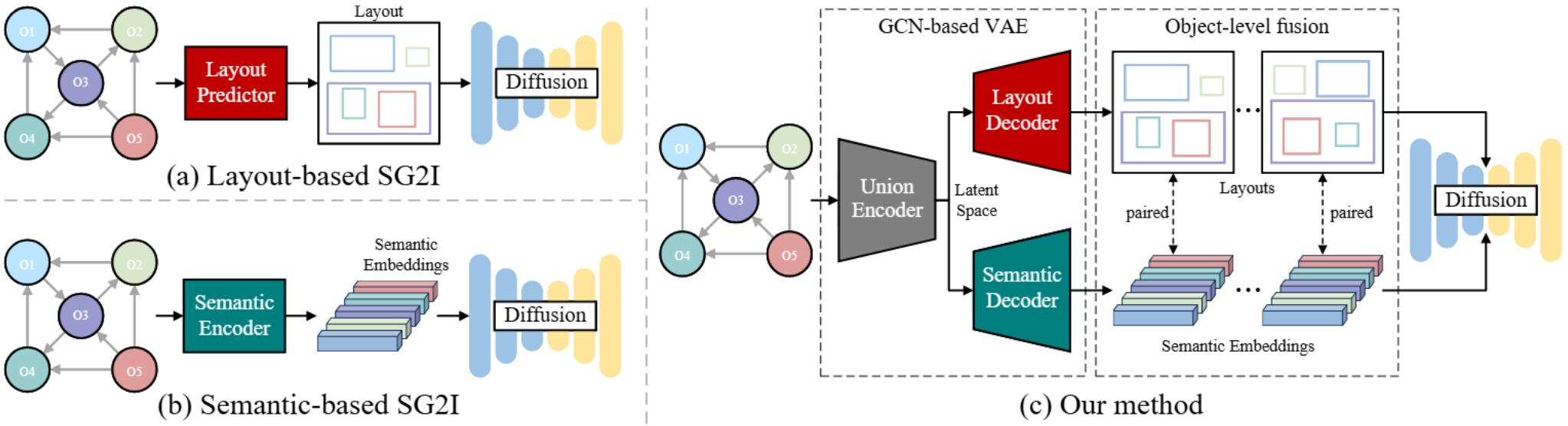
2. Challenges of Layout-to-Image (L2I)

- non-spatial interactions between objects

3. Challenges of Scene-Graph-to-Image (SG2I)

- absence of independent

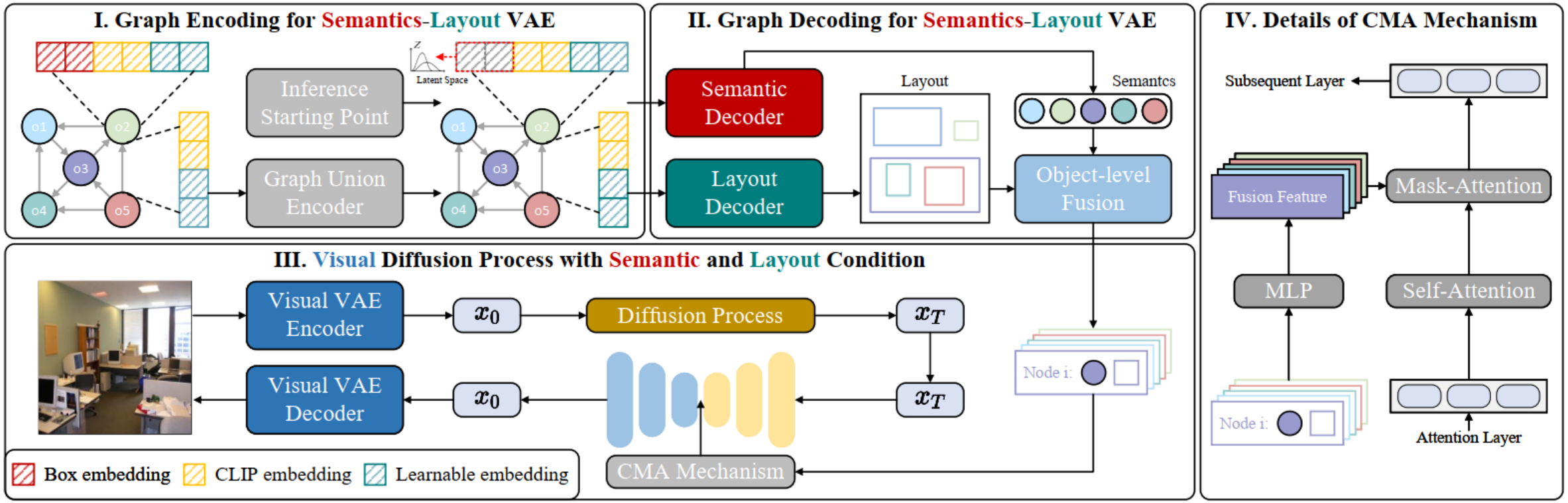
Introduction



Key contributions of our DisCo:

- *Semantics-Layout Variational AutoEncoder* (SL-VAE) that disentangles diverse spatial layouts and interactive semantics from the scene graph;
- *Compositional Masked Attention* (CMA) that injects extracted object-level graph information with fine-grained attributes into the diffusion model;
- *Multi-Layered Sampler* (MLS) that leverages the diverse conditions produced by SL-VAE to implement object-level graph manipulation.

Framework Overview



Scene Graph Representation: Nodes (objects) + Edges (relationships)

$$G = (O, E) \quad = \quad O = \{o_i\}_{i=1}^{N_o} \quad + \quad E = \{e_{ij}\}_{1 \le i,j \le N_o, i \ne j}$$

Experiments

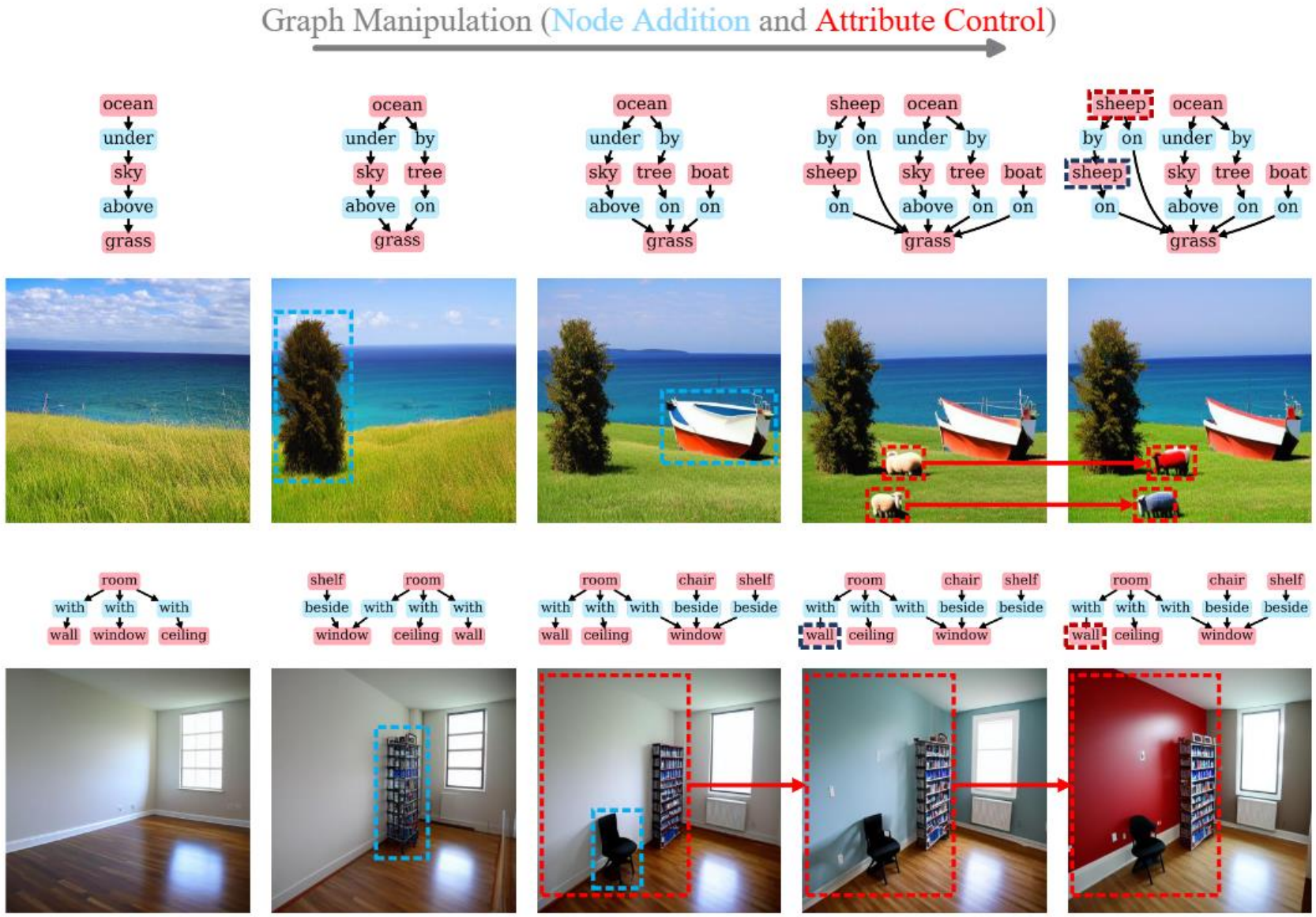
(a) Comparison with T2I methods in *spatial relationships* and *object quantities*.

(b) Comparison with L2I methods in *non-spatial interactions* and *rationality*.

Experiments

(c) Comparison with SG2I methods in *independent node inference* and *generation quality*.

Experiments



Thanks for listening!