Instruct Scene on CF-GISS dataset

Peter HU, 18 Feb 2025

Project page: https://chenguolin.github.io/projects/InstructScene/
Github: https://github.com/chenguolin/InstructScene
arXiv: https://arxiv.org/abs/2402.04717

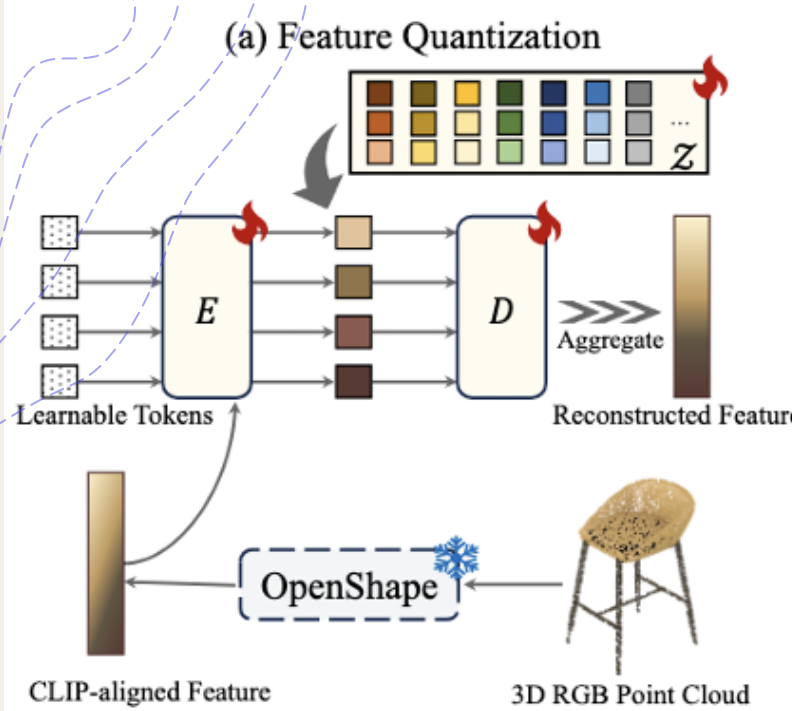


Table 5: Prompt for ChatGPT to refine raw object descriptions.

Given a description of furniture from a captioning model and its ground-truth category, please combine their information and generate a new short description in one line. The provided category must be the descriptive subject of the new description. The new description should be as short and concise as possible, encoded in ASCII. Do not describe the background and counting numbers. Do not describe size like 'small', 'large', etc. Do not include descriptions like 'a 3D model', 'a 3D image', 'a 3D printed', etc. Descriptions such as color, shape and material are very important, you should include them. If the old description is already good enough, you can just copy it. If the old description is meaningless, you can just only include the category. For example: Given 'a 3D image of a brown sofa with four wooden legs' and 'multi-seat sofa', you should return: a brown multi-seat sofa with wooden legs. Given 'a pendant lamp with six hanging balls on the white background' and 'pendant lamp', you should return: a pendant lamp with hanging balls. Given 'a black and brown chair with a floral pattern' and 'armchair', you should return: a black and brown floral armchair. The above examples indicate that you should delete the redundant words in the old description, such as '3D image', 'four', 'six' and 'white background', and you must include the category name as the subject in the new description. The old descriptions is '{`BLIP caption`}', its category is '{`ground-truth category`}', the new descriptions should be:

Instruction prompts for object features

fVQ-VAE

(0) Scene description text y => sematic-features

$$p_{\phi,\theta}(\mathcal{S}|\mathbf{y}) = p_{\phi,\theta}(\mathcal{S},\mathcal{G}|\mathbf{y}) = p_\phi(\mathcal{G}|\mathbf{y})p_\theta(\mathcal{S}|\mathcal{G}). \quad (2)$$

(0) Scene text y
=> sematic-features

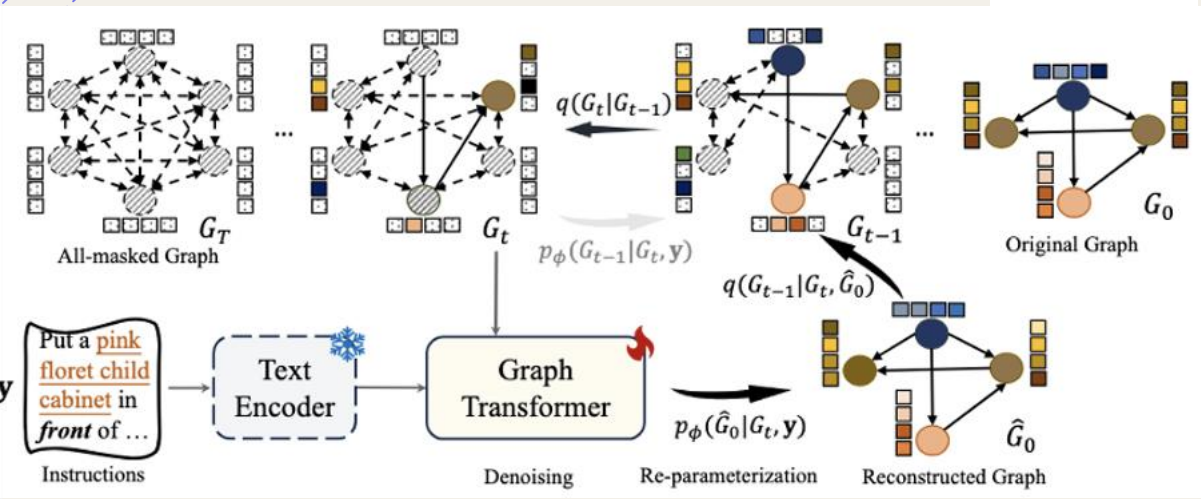
I

Graph:{ [
semantic-features: None,
objs_id: [2,3,5,13,13],
objs_mask: [1,1,1,0,0], edges: [N^2]],
(->) boxes: [postion^3, scale^3, cos\theta, sin \theta]
}

II

Final Scene

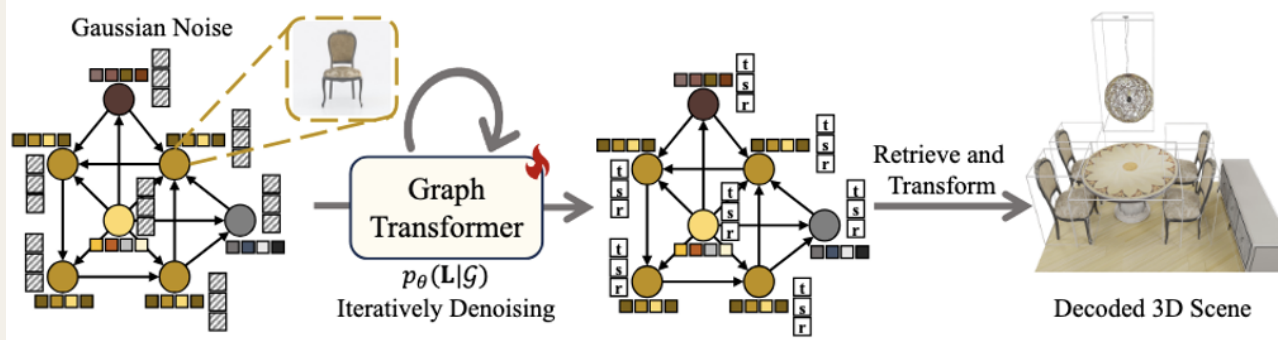
(I) Semantic graph prior $p_\phi(\mathcal{G}|\mathbf{y})$



(G_0, y) … … (G_T, y) … … (G_0*, y)
G: Graph without boxes

{train/gen}_sg.py

(II) 3D layout decoder $p_\theta(\mathcal{S}|\mathcal{G})$



G_0* => boxes

{train/gen}_sc2sg.py

Methods

edge_{i,j}: geometric relationship

Relationship	Rule
Left of	$(\theta_{so} \geq \frac{3\pi}{4}$ or $\theta_{so} < -\frac{3\pi}{4})$ and $1 < d(s,o) \leq 3$
Right of	$-\frac{\pi}{4} \leq \theta_{so} < \frac{\pi}{4}$ and $1 < d(s,o) \leq 3$
In front of	$\frac{\pi}{4} \leq \theta_{so} < \frac{3\pi}{4}$ and $1 < d(s,o) \leq 3$
Behind	$-\frac{3\pi}{4} \leq \theta_{so} < -\frac{\pi}{4}$ and $1 < d(s,o) \leq 3$
Closely left of	$(\theta_{so} \geq \frac{3\pi}{4}$ or $\theta_{so} < -\frac{3\pi}{4})$ and $d(s,o) \leq 1$
Closely right of	$-\frac{\pi}{4} \leq \theta_{so} < \frac{\pi}{4}$ and $d(s,o) \leq 1$
Closely in front of	$\frac{\pi}{4} \leq \theta_{so} < \frac{3\pi}{4}$ and $d(s,o) \leq 1$
Closely bebind	$-\frac{3\pi}{4} \leq \theta_{so} < -\frac{\pi}{4}$ and $d(s,o) \leq 1$
Above	$(\texttt{Center}_{Z_s} - \texttt{Center}_{Z_o}) > (\texttt{Height}_s + \texttt{Height}_o)/2$ and (Inside(s,o) or Inside(o,s))
Below	$(\texttt{Center}_{Z_o} - \texttt{Center}_{Z_s}) > (\texttt{Height}_s + \texttt{Height}_o)/2$ and (Inside(s,o) or Inside(o,s))
None	$d(s,o) > 3$

We set the threshold from 1~3 to **150~450** after observation, leaving the rest same as theirs.

Bedroom inference results

Livingroom inference results

Thank you~

Peter HU, 18 Feb 2025

Project page: https://chenguolin.github.io/projects/InstructScene/
Github: https://github.com/chenguolin/InstructScene
arXiv: https://arxiv.org/abs/2402.04717