

Vision, Language, Interaction and Generation

Qi Wu

Australian Institute for Machine Learning

Australia Centre for Robotic Vision

University of Adelaide



**AUSTRALIAN INSTITUTE
FOR MACHINE LEARNING**



Vision-and-Language

Computer Vision (CV)

- Image Classification



- Object Detection



- Segmentation



- Object Counting

Natural Language Processing (NLP)

- Language Generation



- Language Understanding
- Language Parsing
- Sentiment analysis
- Machine Translation

Bonjour -> Good Morning

- Question Answering (QA)

*Q: Who is the president of US?
A: Barack Obama*

Vision-and-Language

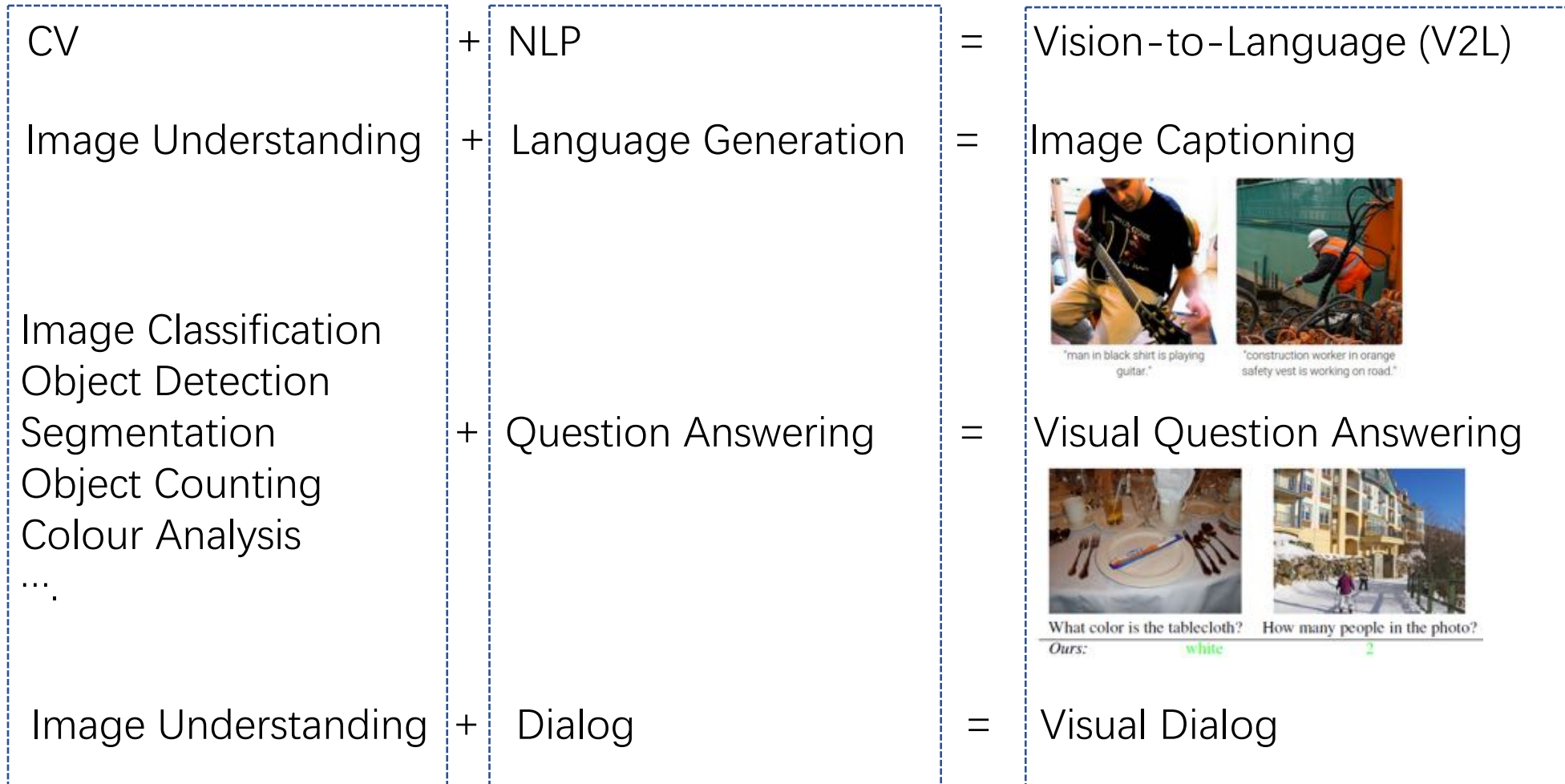


Image Captioning

- Definition
 - Automatic describe an image with natural language.



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

Visual Question Answering

Definition: An image and a free-form, open-ended question about the image are presented to the method which is required to produce a suitable answer.



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

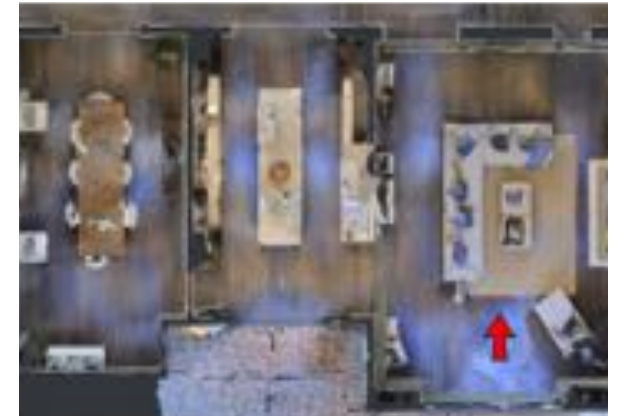


Does it appear to be rainy?
Does this person have 20/20 vision?

Connecting Vision and Language to Interaction



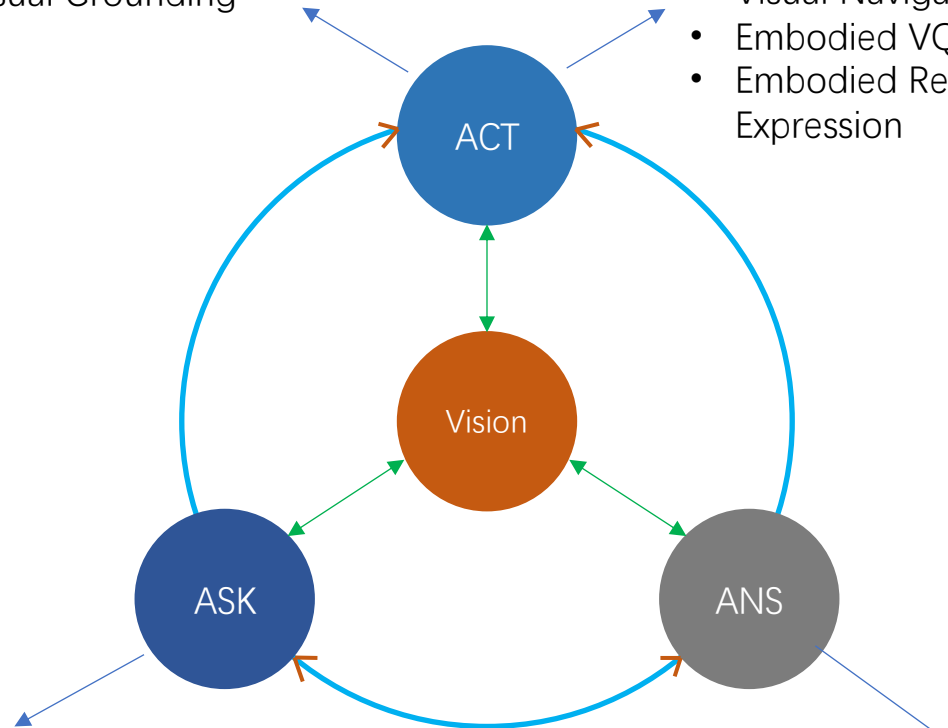
- Referring Expression
- Visual Grounding



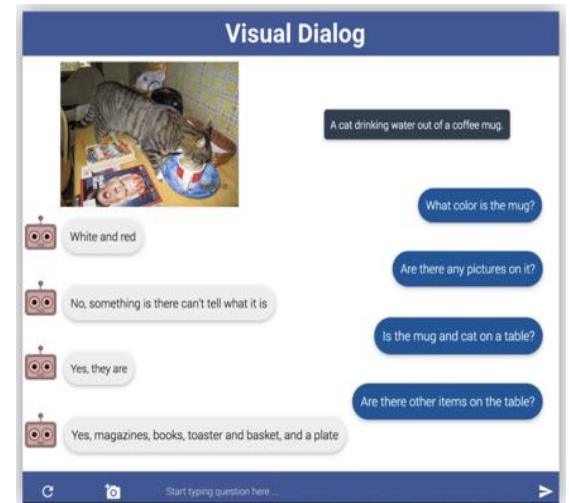
- Language-guided Visual Navigation
- Embodied VQA
- Embodied Referring Expression



- Visual Question Generation (VQG)
- Question2Query
- Image Captioning



- VQA
- VisDial



Our works

- Image Captioning

- Shizhe Chen, Qin Jin, Peng Wang, **Qi Wu**. Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs. **CVPR' 20**
- **Qi Wu**, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, Anthony Dick. *What Value Do Explicit High Level Concepts Have in Vision to Language Problems?* **CVPR' 16**
- **Qi Wu**, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel, *Image Captioning and Visual Question Answering Based on Attributes and Their Related External Knowledge.* **TPAMI**

- VQA

- **Qi Wu**, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick. *Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources.* **CVPR' 16**
- Peng Wang*, **Qi Wu***, Chunhua Shen, Anton van den Hengel. *The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions.* **CVPR' 17**
- Damien Teney, Lingqiao Liu, Anton van den Hengel, *Graph-Structured Representations for Visual Question Answering.* **CVPR' 17**
- Peng Wang*, **Qi Wu***, Chunhua Shen, Anton van den Hengel, Anthony Dick. *Explicit Knowledge-based Reasoning for Visual Question Answering.* **IJCAI' 17**
- Peng Wang*, **Qi Wu***, Chunhua Shen, Anton van den Hengel, Anthony Dick. *FVQA: Fact-based Visual Question Answering.* **TPAMI**
- **Qi Wu**, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. *Visual question answering: A survey of methods and datasets.* **CVIU**
- Damien Teney, **Qi Wu**, Anton van den Hengel. *Visual Question Answering: A Tutorial.* **IEEE Signal Processing Magazine.**
- Chao Ma, Chunhua Shen, Anthony Dick, **Qi Wu**, Peng Wang, Anton van den Hengel, Ian Reid. *Visual Question Answering with Memory-Augmented Networks.* **CVPR' 18**
- Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel, *Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge.* **CVPR' 18**

- Visual Dialog

- **Qi Wu**, Peng Wang, Chunhua Shen, Ian Reid, Anton van den Hengel. *Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning*. **CVPR' 18 [oral]**
- Jiang, X., Yu, J., Qin, Z., Zhuang, Y., Zhang, X., Hu, Y. and **Wu, Q.**, 2019. DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue. **AAAI 2020**.

- Visual Question Generation

- Junjie Zhang*, **Qi Wu***, Chunhua Shen, Jian Zhang, Anton van den Hengel. *Asking the Difficult Questions: Goal-Oriented Visual Question Generation via Intermediate Rewards*. **ECCV' 18**
- Ehsan Abbasnejad, **Qi Wu**, Javen Shi, Anton van den Hengel. *What's to know? Uncertainty as a Guide to Asking Goal-oriented Questions*. **CVPR' 19**

- Referring Expression/Visual Grounding

- Bohan Zhuang*, **Qi Wu***, Chunhua Shen, Ian Reid, Anton van den Hengel. *Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries*. **CVPR' 18**
- Chaorui Deng*, **Qi Wu***, Fuyuan Hu, Fan Lv, Mingkui Tan, Qingyao Wu. *Visual Grounding via Accumulated Attention*. **CVPR' 18**
- Peng Wang, **Qi Wu**, Jiewei Cao, Chunhua Shen, Lianli Gao, Anton van den Hengel. *Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks*. **CVPR' 19**

- Image-Sentence Matching

- Yan Huang, **Qi Wu**, Liang Wang. *Learning Semantic Concepts and Order for Image and Sentence Matching*. **CVPR' 18**
- Yan Huang, **Qi Wu**, Wei Wang, Liang Wang. *Image and Sentence Matching via Semantic Concepts and Order Learning*. IEEE Transaction on Pattern Analysis and Machine Intelligence (**TPAMI**),

- Language-guided Navigation

- Peter Anderson, **Qi Wu**, Damien Teney, Jake Bruce, Mark Johnson, Niko Sanderhauf, Ian Reid, Stephen Gould, Anton van den Hengel. *Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments*. **CVPR' 18**

- Visual Relationship Detection

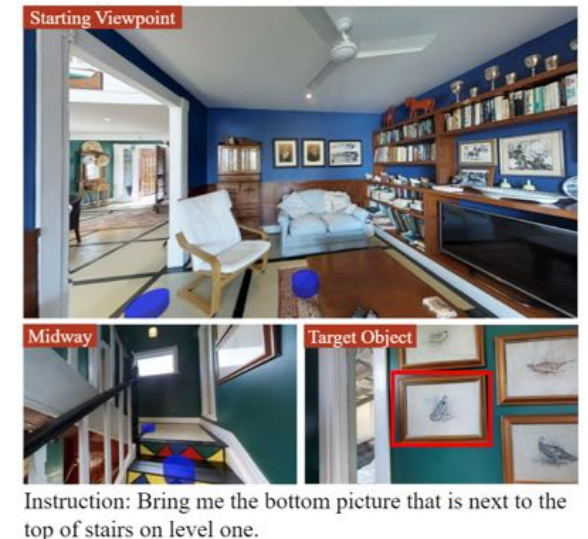
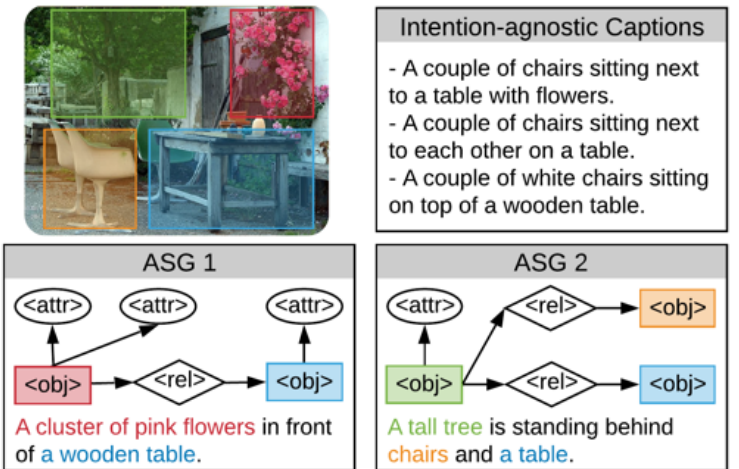
- Bohan Zhuang*, **Qi Wu***, Ian Reid, Chunhua Shen, Anton van den Hengel. *HCVRD: a benchmark for large-scale Human-Centered Visual Relationship Detection*. **AAAI' 18**

Interaction and Generation

- Controllable text generation
 - Novel object captioning
 - Captioning with styles
 - Describe different regions/objects/relationships
- Text-conditioned image/video generation
 - Text2image
 - Image editing with text
- Interact with environment with natural language
 - Vision-language navigation

Interaction and Generation

- Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs, CVPR 20, Oral
- Intelligent Home 3D: Automatic 3D-House Design from Linguistic Descriptions Only, CVPR 20
- REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments, CPVR 20, Oral



Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs

Shizhe Chen, Qin Jin, Peng Wang,

Qi Wu

CVPR2020

Image Caption Generation

- Aim to generate a sentence to describe image contents
 - One of the ultimate goal for holistic image understanding
- Most methods are intention-agnostic
 - Passively generate image descriptions
 - Fail to realize what a user wants to describe
 - Lack of diversity



Intention-agnostic Captions

- A couple of chairs sitting next to a table with flowers.
- A couple of chairs sitting next to each other on a table.
- A couple of white chairs sitting on top of a wooden table.

Controllable Image Caption Generation

- Generate sentence to describe designated image contents
 - Different image regions [1]
 - Single object [2]
 - A set / sequence of objects [3]
- None can control caption generation at **fine-grained** level
 - Whether (and how many) associative **attributes** should be used?
 - Any other objects (and its associated **relationships**) should be included?
 - What is the description **order**?

[1] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. CVPR 2016.

[2] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. CVPR 2019.

[3] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. CVPR 2019.

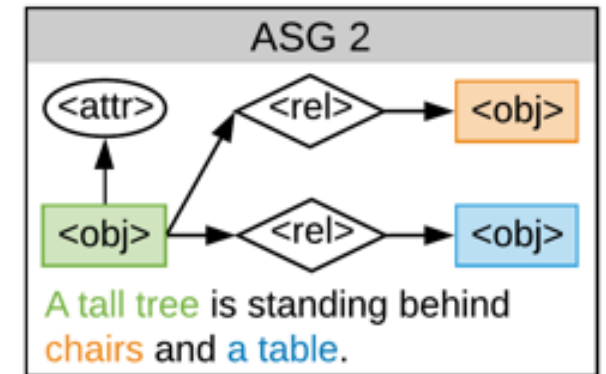
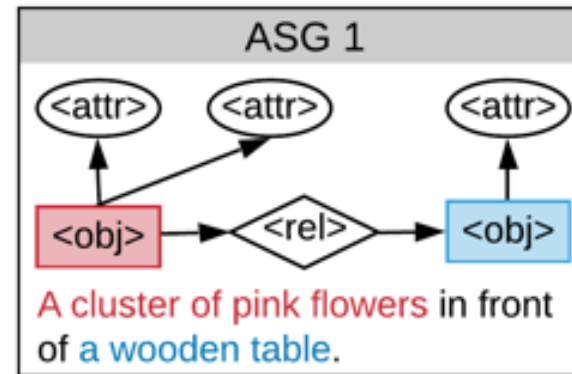
ASG: Fine-grained Controlling

- Abstract Scene Graph (ASG)
 - Directed graph consisting of **abstract nodes** (object, attribute, relationship)
 - Nodes are grounded but their semantic contents are unknown
 - Represent user desired contents at a fine-grained level
- Easy to construct
 - Designated by users
 - Created automatically



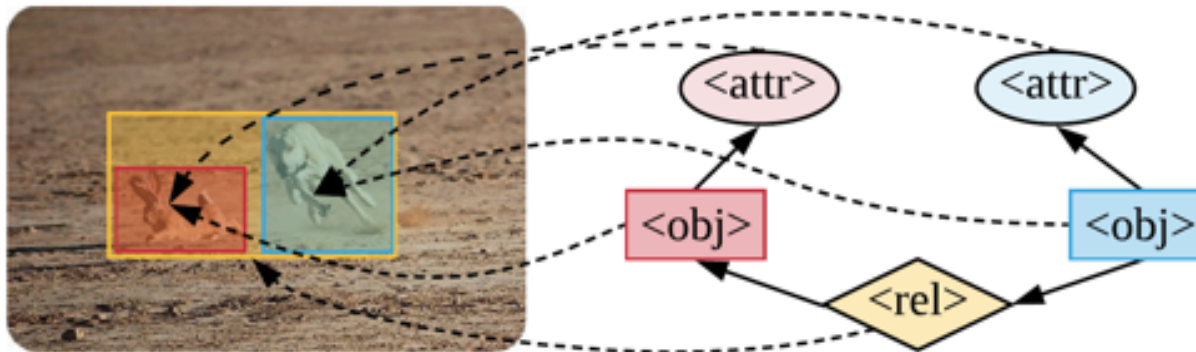
Intention-agnostic Captions

- A couple of chairs sitting next to a table with flowers.
- A couple of chairs sitting next to each other on a table.
- A couple of white chairs sitting on top of a wooden table.



Challenges for ASG Controlled Captioning

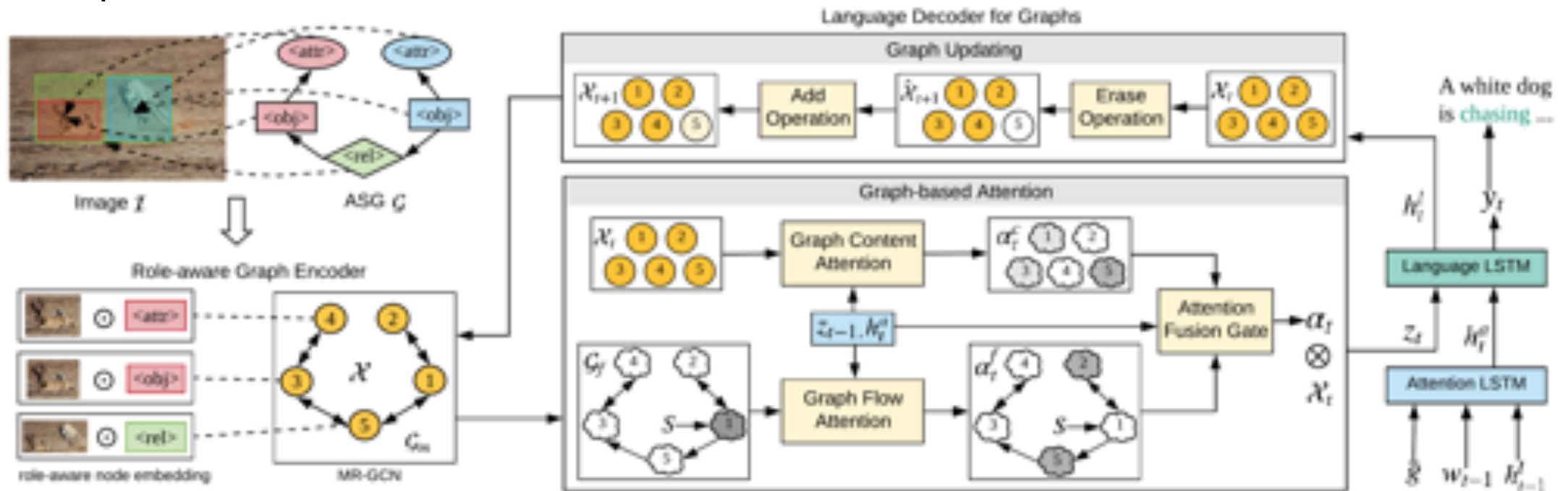
- Differentiate intentions of different types of abstract nodes
- Recognize semantic meanings of abstract nodes
- Follow the graph structure order to generate desired descriptions
- Cover all nodes in the graph without missing or repetition



A white dog
is chasing
a brown rabbit.

Proposed ASG2Caption Model

- ASG \rightarrow Role-aware Graph Encoder \rightarrow Language Decoder for Graphs



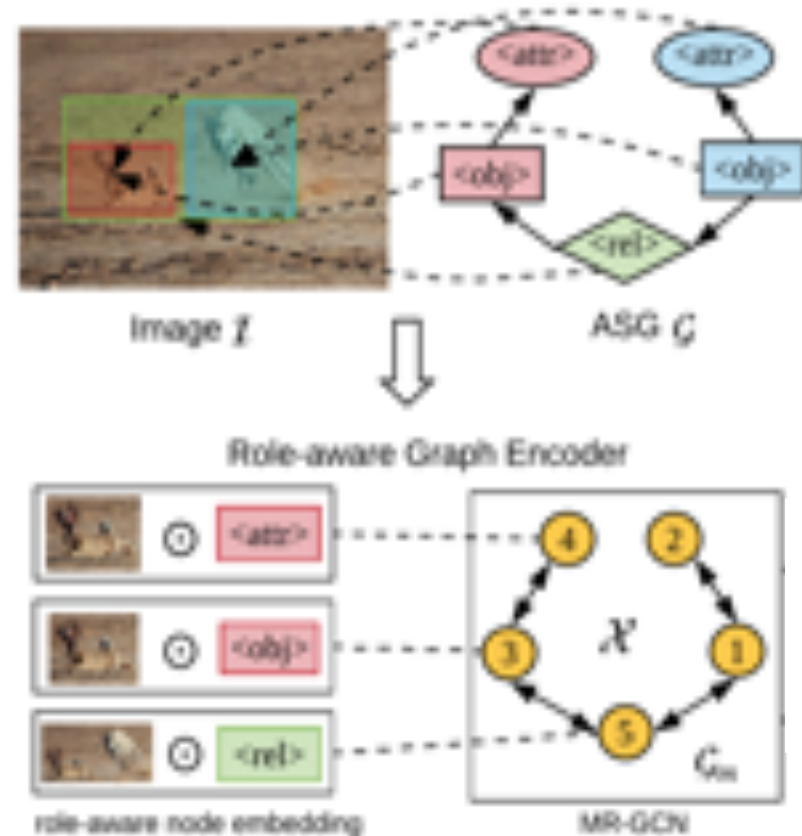
Role-aware Graph Encoder

- Role-aware Embedding
 - enhance visual grounded node with role embedding

$$x_i^{(0)} = \begin{cases} v_i \odot W_r[0], & \text{if } i \in o; \\ v_i \odot (W_r[1] + \text{pos}[i]), & \text{if } i \in a; \\ v_i \odot W_r[2], & \text{if } i \in r. \end{cases}$$

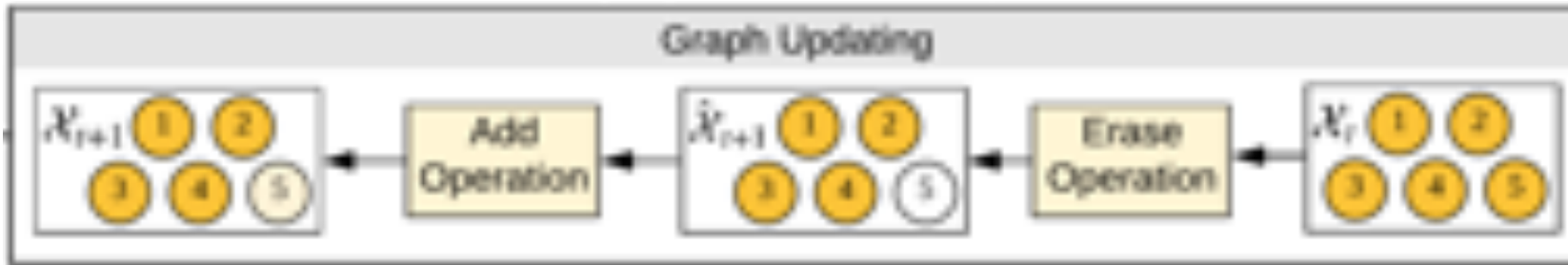
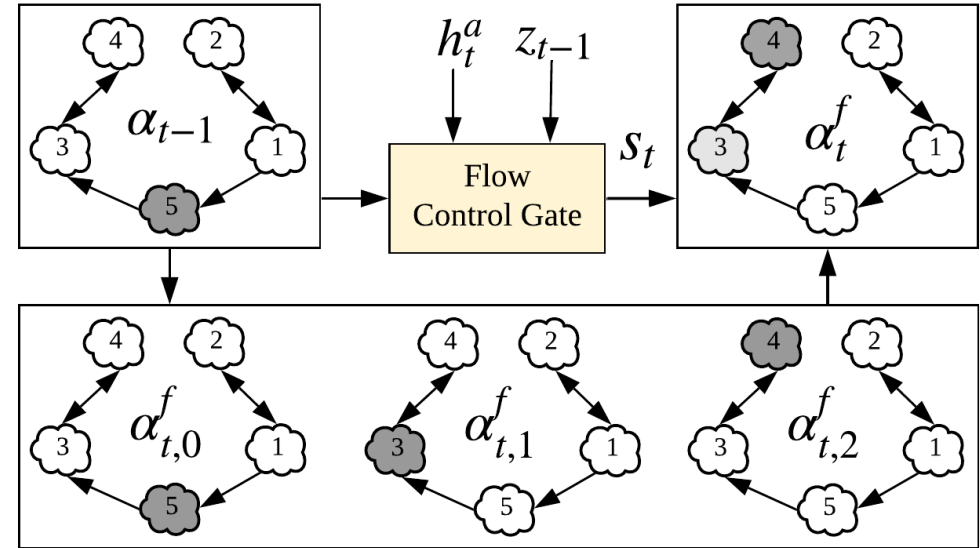
- Multi-relational Graph Convolution Network
 - Improve node representations with graph contexts

$$x_i^{(l+1)} = \sigma(W_0^{(l)} x_i^{(l)} + \sum_{\tilde{r} \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^{\tilde{r}}} \frac{1}{|\mathcal{N}_i^{\tilde{r}}|} W_{\tilde{r}}^{(l)} x_j^{(l)})$$



Language Decoder for Graphs

- Graph-based Attention
 - Graph Content Attention
 - Graph Flow Attention
 - Follow the graph structure order
- Graph Updating
 - Keep a record of accessed status
 - Erase + addition



Experiments

- Dataset Construction

- Utilize (image, caption) pairs on VisualGenome and MSCOCO datasets
- Automatically construct triplets of (image, ASG, caption)

dataset	#objs / sent	#rels / obj	#attrs / obj	#words / sent
VisualGenome	2.09	0.95	0.47	5.30
MSCOCO	2.93	1.56	0.51	10.28

- Evaluation Metrics

- Controllability
 - Structure-only: Graph structure difference (the lower the better)
 - Structure + semantic: BLEU4, METEOR, ROUGEL, CIDER, SPICE
- Diversity: Div-n, Self-cider

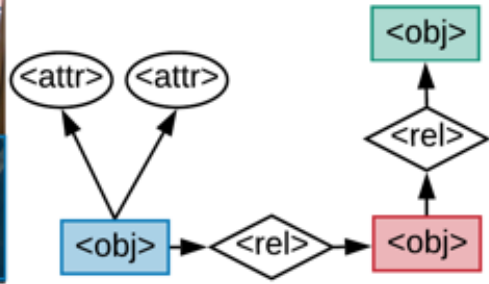
Evaluation on Controllability

- Comparison with State of the Arts
 - Baselines
 - Non-controllable: Show-Tell (ST), Bottom-up Top-down attention (BUTD)
 - Controllable: C-ST, C-BUTD
 - Our ASG2Caption > Controllable baselines > non-controllable baselines
 - Achieve significant improvements in terms of semantic quality and structure alignment

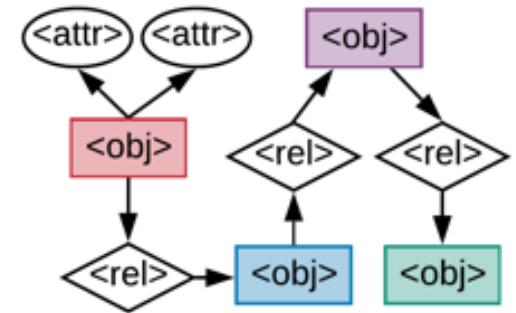
Method	VisualGenome									MSCOCO								
	B4	M	R	C	S	G	G _o	G _a	G _r	B4	M	R	C	S	G	G _o	G _a	G _r
ST [37]	11.1	17.0	34.5	139.9	31.1	1.2	0.5	0.7	0.5	10.5	16.8	36.2	100.6	24.1	1.8	0.8	1.1	1.0
BUTD [3]	10.9	16.9	34.5	139.4	31.4	1.2	0.5	0.7	0.5	11.5	17.9	37.9	111.2	26.4	1.8	0.8	1.1	1.0
C-ST	12.8	19.0	37.6	157.6	36.6	1.1	0.4	0.7	0.4	14.4	20.1	41.4	135.6	32.9	1.6	0.6	1.0	0.8
C-BUTD	12.7	19.0	37.9	159.5	36.8	1.1	0.4	0.7	0.4	15.5	20.9	42.6	143.8	34.9	1.5	0.6	1.0	0.8
Ours	17.6	22.1	44.7	202.4	40.6	0.7	0.3	0.3	0.3	23.0	24.5	50.1	204.2	42.1	0.7	0.4	0.3	0.3

Evaluation on Controllability

- Qualitative Examples
 - Given ASGs corresponding to ground-truth image descriptions
 - Faithfully follow the ASG structure



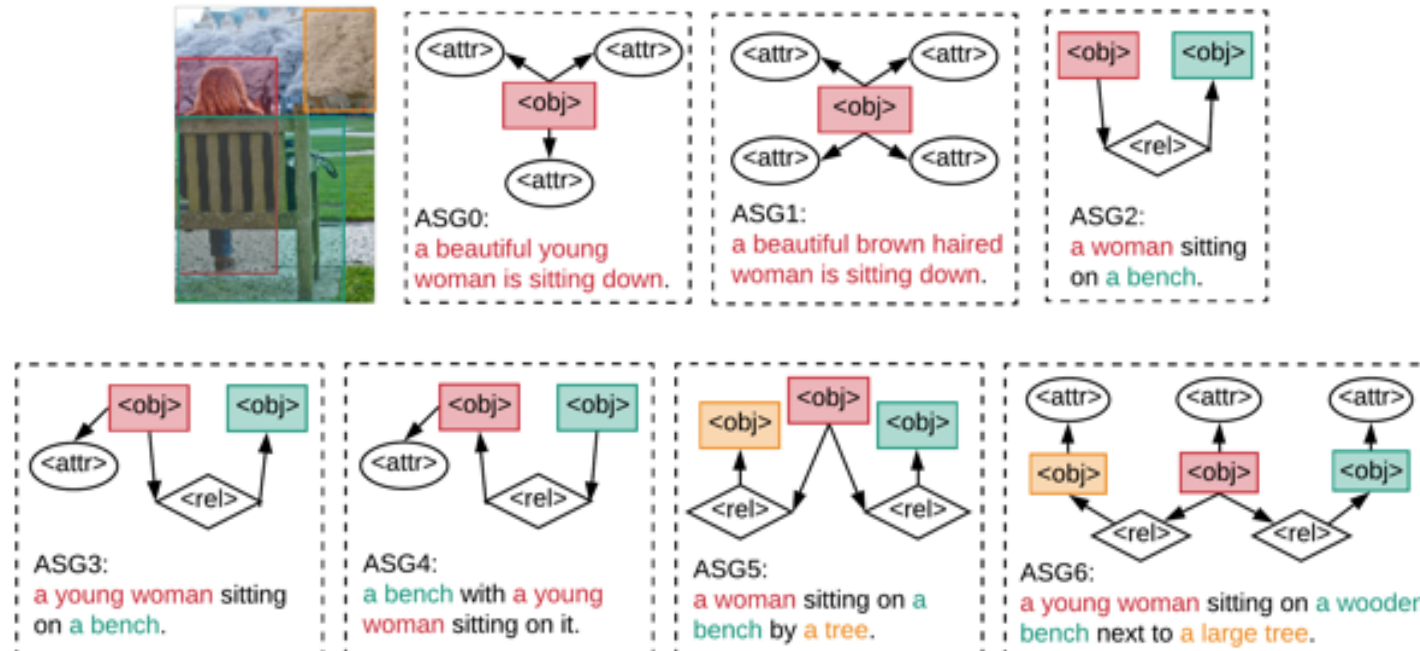
GT: large office desk with computers near a window.
 C-BUTD: a desk with two monitors and a window.
 Ours: a wooden computer desk with computers sitting next to a window.
 <---attr---> <---attr---> <obj> <rel> <---obj---> <---rel---> <---obj--->



GT: two small children with umbrellas in a field near the shore.
 C-BUTD: a couple of kids holding a kite on a beach.
 Ours: two small children playing with kites in a field near the ocean.
 <attr> <attr> <obj> <---rel---> <obj> <rel> <obj> <rel> <---obj--->

Evaluation on Controllability

- Qualitative Examples
 - Given user designated ASGs
 - Generate different descriptions for ASGs with very subtle changes

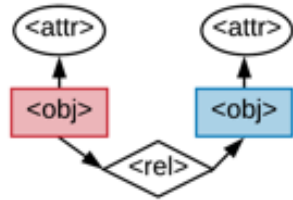


Evaluation on Diversity

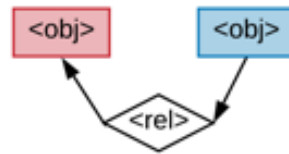
- Comparison with State of the Arts
 - Generate more diverse descriptions even compared with diversity-driven models

	Method	Div-1	Div-2	SelfCIDEr
Visual Genome	Region	0.41	0.43	0.47
	Ours	0.54	0.63	0.75
MS COCO	BS [4]	0.21	0.29	-
	POS [7]	0.24	0.35	-
	SeqCVAE [4]	0.25	0.54	-
	BUTD-BS	0.29	0.39	0.58
	Ours	0.43	0.56	0.76

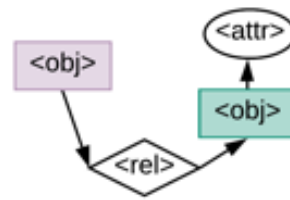
- Qualitative Examples



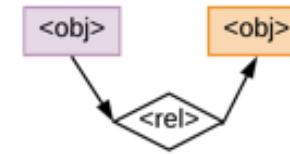
Dcap: the skis are red
Ours: **red skis** under **red boots**



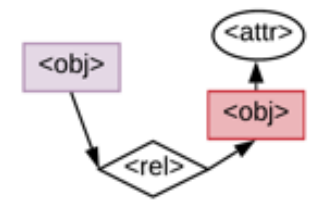
Dcap: the skis are red
Ours: **boots** on **the skis**



Dcap: a man is wearing a red jacket
Ours: **the man** holding a **ski pole**



Dcap: a man is wearing a red jacket
Ours: **the man** is wearing **gloves**



Dcap: a man is wearing a red jacket
Ours: **the man** is wearing **red skis**.

Ablation Studies

- Contributions from different components

Table 3: Ablation study to demonstrate contributions from different proposed components. (role: role-aware node embedding; rgcn: MR-GCN; ctn: graph content attention; flow: graph flow attention; gupdt: graph updating; bs: beam search)

#	Enc		Dec				VisualGenome					MSCOCO				
	role	rgcn	ctn	flow	gupdt	bs	B4	M	R	C	S	B4	M	R	C	S
1							11.2	18.3	36.7	146.9	35.6	13.6	19.7	41.3	130.2	32.6
2			✓				10.7	18.2	36.9	146.3	35.5	14.5	20.4	42.2	135.7	34.6
3	✓		✓				14.2	20.5	40.9	176.9	38.1	18.2	22.5	44.9	166.9	37.8
4	✓	✓	✓				15.7	21.4	43.6	191.7	40.0	21.6	23.7	48.6	190.5	40.9
5	✓	✓	✓	✓			15.9	21.5	44.0	193.1	40.1	22.3	24.0	49.4	196.2	41.5
6	✓	✓	✓		✓		15.8	21.4	43.5	191.6	39.9	21.8	24.1	49.1	194.2	41.4
7	✓	✓	✓	✓	✓		16.1	21.6	44.1	194.4	40.1	22.6	24.4	50.0	199.8	41.8
8	✓	✓	✓	✓	✓	✓	17.6	22.1	44.7	202.4	40.6	23.0	24.5	50.1	204.2	42.1

Conclusion

- Fine-grained control of image caption generation
 - control on what and how detailed to describe
 - need deep reasoning on regions and graphs of a given image
- Contributions
 - Design a novel control signal called Abstract Scene Graph (ASG)
 - Propose an ASG2Caption model with a role-aware graph encoder and a language decoder specifically for graphs for caption generation
 - Our model achieves state-of-the-art controllability and significantly improves diversity of captions given automatically sampled ASGs

Intelligent Home 3D: Automatic 3D-House Design from Linguistic Descriptions Only

Qi Chen^{1,2}, **Qi Wu**³, Rui Tang⁴, Yuhan Wang⁴, Shuai Wang⁴, Mingkui Tan¹

¹South China University of Technology, ²Pazhou Lab

³University of Adelaide, ⁴Kujiale Inc

Paper link: <https://arxiv.org/abs/2003.00397>



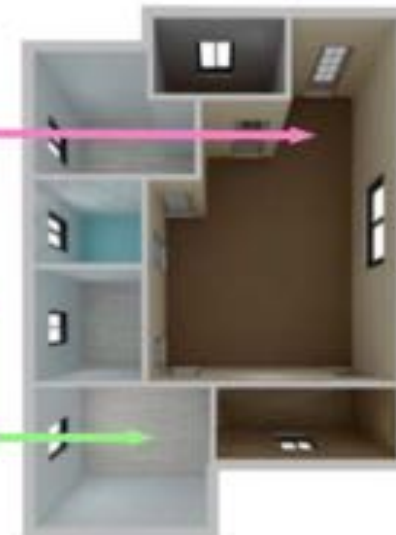
Task Description

- **What is 3D-house generation from requirements?**

3D-house generation from requirement seeks to design a 3D building **automatically** from given **linguistic descriptions**

- **An example of generated 3D-house with description:**

The building contains two bedrooms, one washroom, one balcony, one living room, and one kitchen. Bedroom2 is in southeast with 10 square meters. Bedroom2 floor is **White Wood Veneer** and wall is Blue Wall Cloth... Livingroom1 is next to bedroom1. Bedroom1 is adjacent to balcony1...



(a) User linguistic requirements

(b) 3D-House design

■ Why we try designing 3D house **automatically**?

Design by humans has many limitations:

- High requirements for **professional skills and tools**
- High **time-consumption** (from a couple of days to several weeks)

■ Why we use the **linguistic descriptions** as inputs?

- People do **not** have design knowledge and experience of using designing tools
- People have **strong linguistic ability** to express our interests and desire

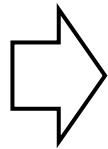
Task Description

We divide 3D-house generation process into two sub-tasks:

- **Building layout generation**
- **Texture synthesis**



(a) 3D-house generation

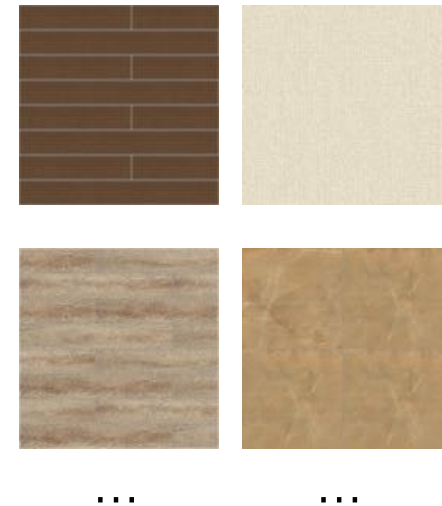


Input: *The building contains two bedrooms, one washroom, one balcony, one living room, and one kitchen. Bedroom2 is in southeast with 10 square meters. Bedroom2 floor is White Wood Veneer and wall is Blue Wall Cloth... Livingroom1 is next to bedroom1. Bedroom1 is adjacent to balcony1 ...*



(b) Layout generation

+



(c) Texture synthesis

Challenges

■ What is the challenges of our 3D-house generation task?

- A floor plan is a structured layout which require the **correctness** of size, direction and connection of different blocks

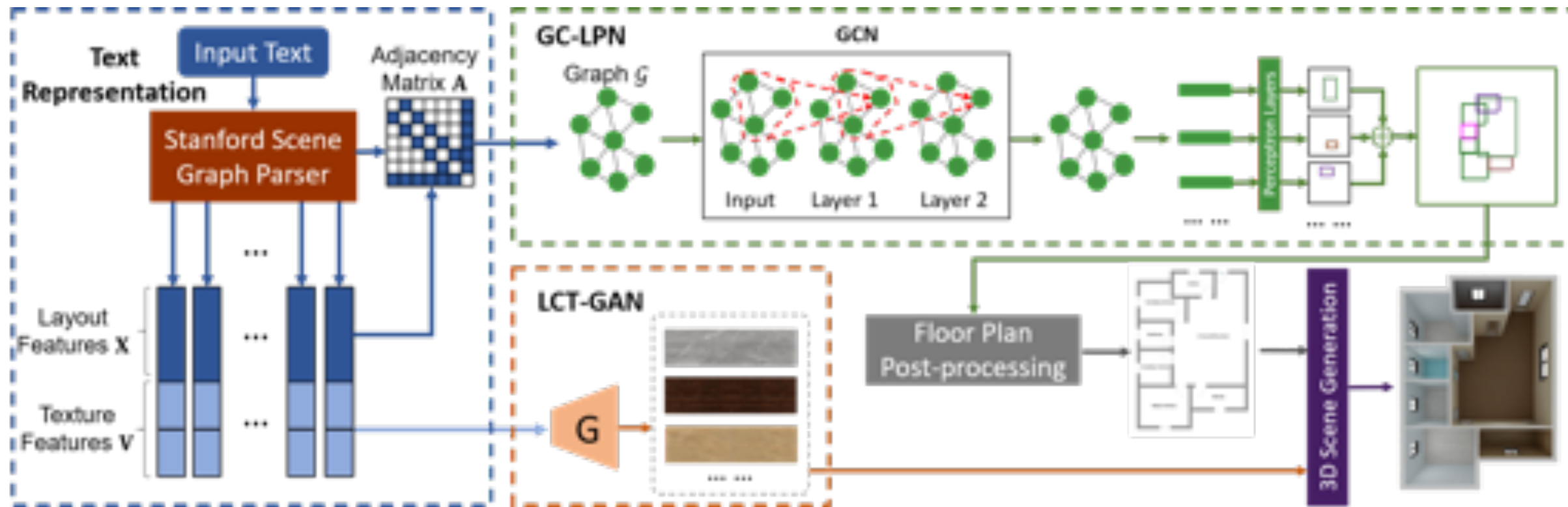


- The interior texture such as floor and wall needs **neater pixel** generation



- The generated 3D-house should be well **aligned with the given descriptions**

House Plan Generative Model (HPGM)



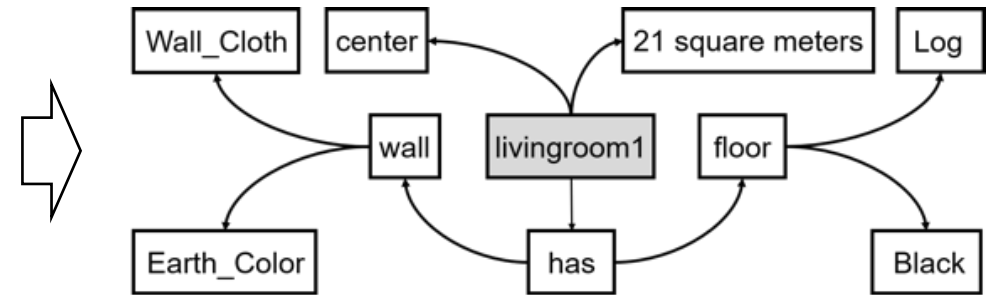
Architecture of HPGM

- Text representation block
- Graph conditioned layout prediction network (GC-LPN)
- Floor plan post-processing
- Language conditioned texture GAN (LCT-GAN)
- 3D scene generation and rendering

Text Representation

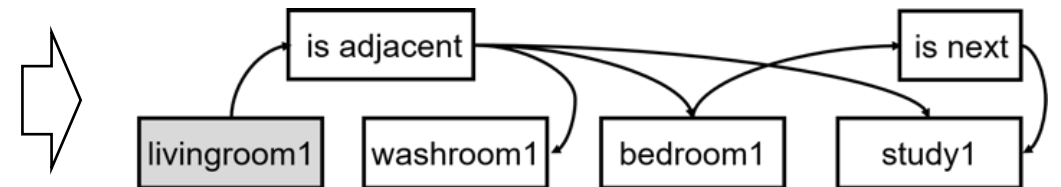
(1) Scene graph of each room

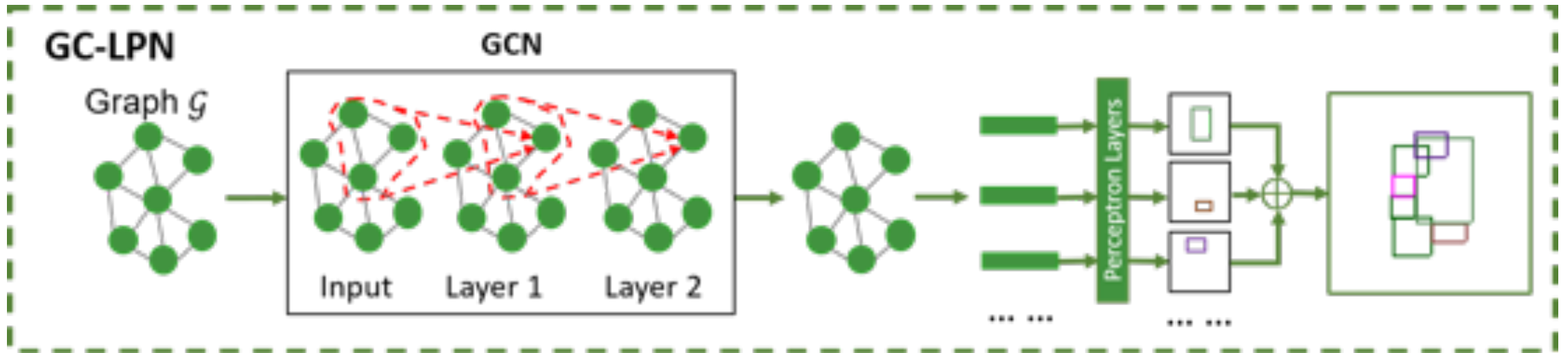
- S1 = “livingroom1 is in center with 21 square meters”
- S2 = “livingroom1 has Earth_Color Wall_Cloth for wall while black log for floor”



(2) Scene graph of adjacency between rooms

- S3 = “livingroom1 is adjacent to washroom1, bedroom1, study1”
- S4 = “bedroom1 is next to study1”





■ **A two-layer GCN model:**

$$\mathbf{Y} = g(\mathbf{X}, \mathbf{A}) = \text{Softmax}(\mathbf{A} \text{ReLU}(\mathbf{A} \mathbf{X} \mathbf{W}_0) \mathbf{W}_1)$$

$$\mathbf{S} = \mathbf{X} \oplus \mathbf{Y}$$

where

\mathbf{W}_0 First layer parameters \mathbf{W}_1 Second layer parameters \mathbf{A} Adjacency matrix \mathbf{X} Node and edge attributes \oplus Element-wise addition

■ **Bounding box regression:**

$$\mathcal{L}_B = \frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{b}}_i - \mathbf{b}_i\|_2^2$$

where Ground truth $\mathbf{b}_i = (x_0, y_0, x_1, y_1)$ Prediction $\hat{\mathbf{b}}_i = h(\mathbf{S}_i) = (\hat{x}_0, \hat{y}_0, \hat{x}_1, \hat{y}_1)$

Floor Plan Post-processing

■ Step (a)

Extract boundary lines of all generated bounding boxes

■ Step (b)

Merge the adjacent segments together

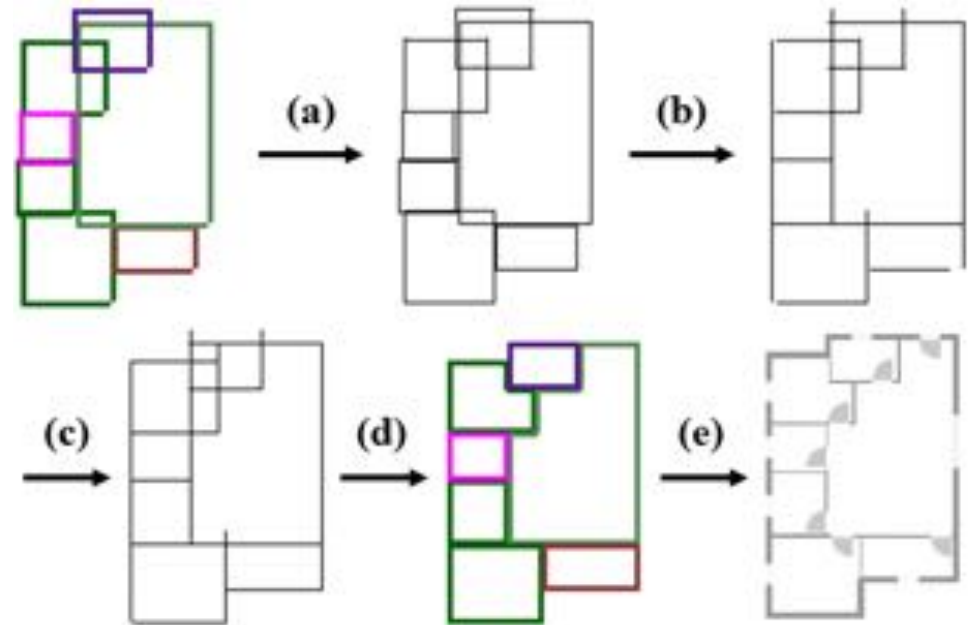
■ Step (c)

Align the line segments with each other to obtain the closed polygon

■ Step (d)

Judge the belonging of each closed polygon based on a weight function:

$$W_{ij} = \iint \frac{1}{w_i h_i} \exp\left(-\left(\frac{x_j - c_{x_i}}{w_i}\right)^2 - \left(\frac{y_j - c_{y_i}}{h_i}\right)^2\right) dx_j dy_j$$



■ Step (e)

Apply a simple rule-based method to add doors and windows in rooms

LCT-GAN

Modules

■ Generator G

Generate image $G(\mathbf{Z})$ using tensor \mathbf{Z} (including random noise \mathbf{Z}' , material vector \mathbf{p} and color vector \mathbf{q})

■ Discriminator D

- Ensure the generated images are **natural and realistic**
- Preserve the **semantic alignment** between texts and texture images

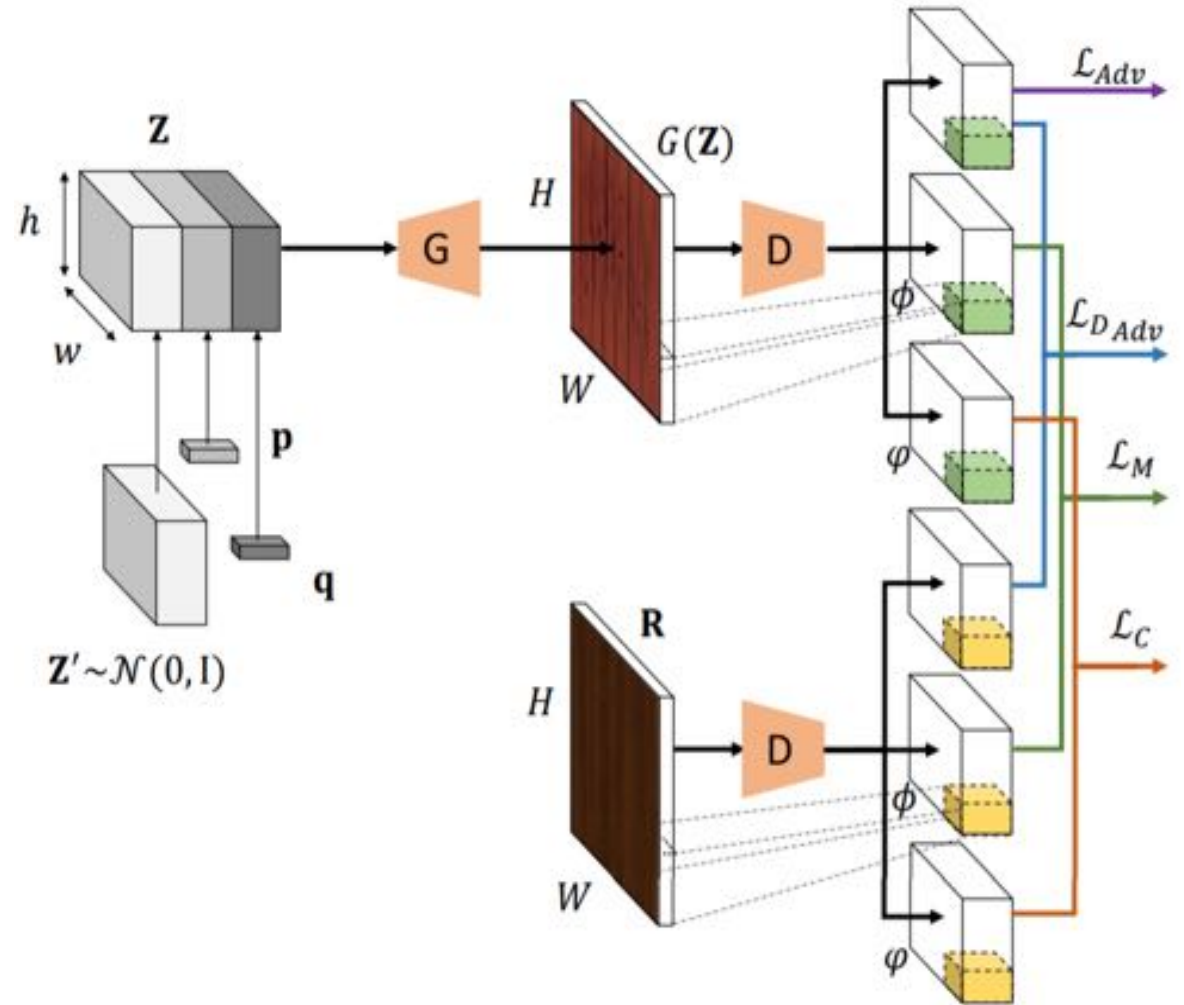
Losses

■ Adversarial loss

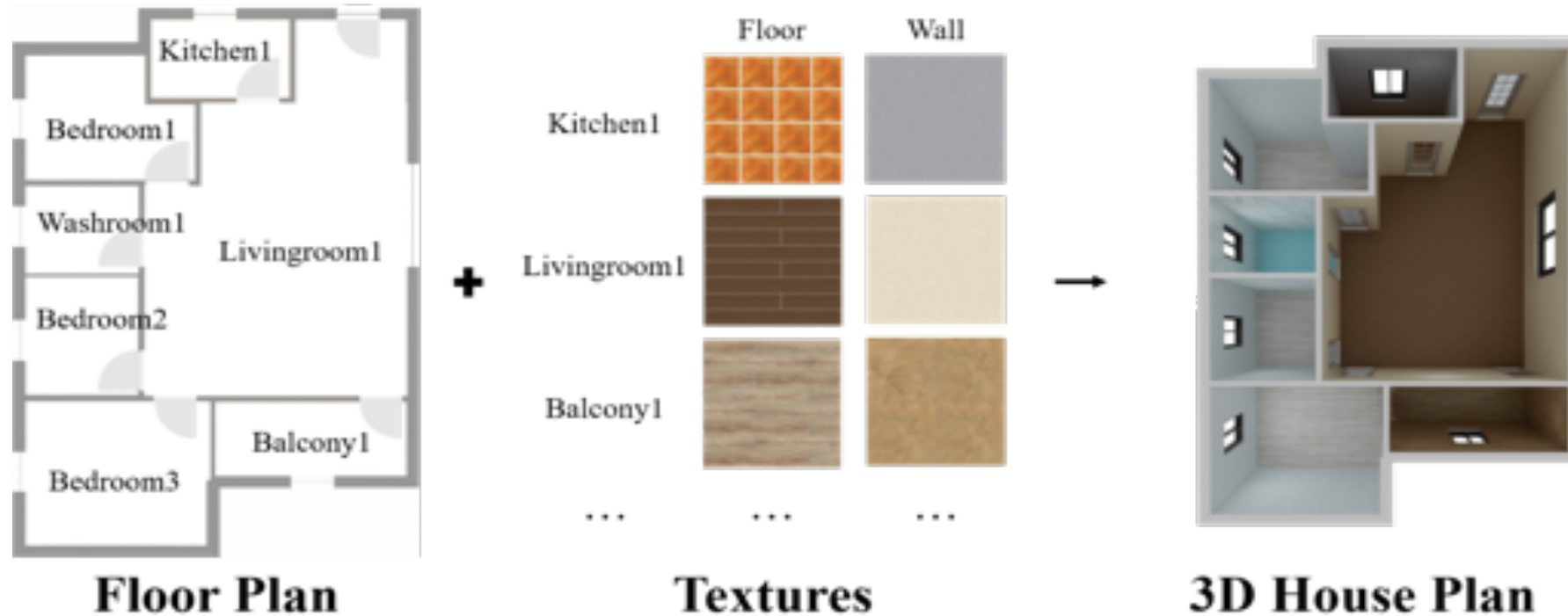
Synthesize the **natural** images

■ Material-aware and color-aware loss

Preserve the **semantic alignment** between generated textures and given texts



3D Scene Generation and Rendering



Rule-based Processing

- Generate walls from boundaries of rooms with **fixed** height and thickness
- Set the length of the window to thirty percent of the length of the wall it belongs to

Photo-realistic Rendering

- Simulates real-world effects such as indirect lighting and global illumination
- Capture a **top-view** render image

Evaluation Metrics and Baselines

Building layout generation

■ Evaluation metric

- Intersection-over-Union (IoU)

■ Baselines

- Manually Layout Generation (MLG): draw layouts directly with the **predefined rules**
- Conditional Layout Prediction Network (C-LPN): **remove GCN**
- Recurrent Conditional Layout Prediction Network (RC-LPN): **replace GCN** with an **LSTM**

Texture synthesis

■ Evaluation metric

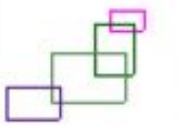
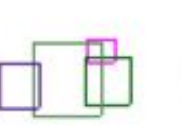
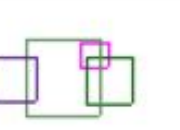
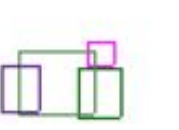
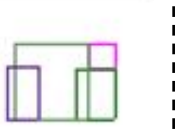

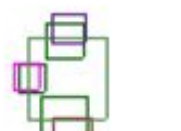



- Fréchet Inception Distance (FID)
- Multi-scale Structural Similarity (MS-SSIM)




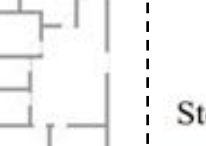
■ Baselines

- ACGAN, StackGAN-v2 and PSGAN

Experimental Results

	MLG	C-LPN	RC-LPN	GC-LPN (ours)
IoU	0.7208	0.8037	0.7918	0.8348

	MLG	C-LPN	RC-LPN	GC-LPN (ours)	GT
Text1					
Text2					

	Ours	GT		Ours	GT
Text1			Text2		

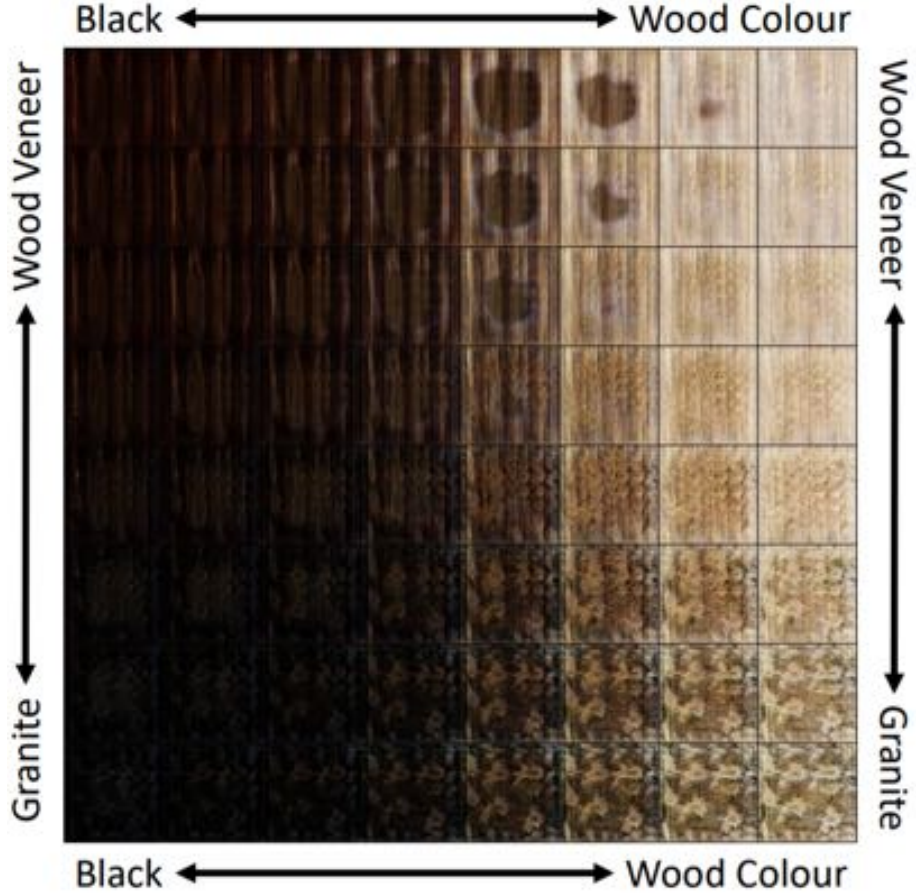
(a) Building Layouts

Methods	Train Set		Test Set	
	FID	MS-SSIM	FID	MS-SSIM
ACGAN [26]	198.07	0.4584	220.18	0.4601
StackGAN-v2 [46]	182.96	0.6356	188.15	0.6225
PSGAN [2]	195.29	0.4162	217.12	0.4187
LCT-GAN (ours)	119.33	0.3944	145.16	0.3859



(b) Generated Textures

Generalization Ability







(c) Interpolation results



(d) Novel material-color scenarios

3D House Design

Linguistic Requirements	Ours		Ground-truth	
	2D Floor Plan	3D House Plan	2D Floor Plan	3D House Plan
<p>The house has one washroom, one livingroom, one storage, two bedrooms, one kitchen, and one balcony. In practice, washroom1 is in west with 5 square meters. wall of washroom1 is Coating and Yellow, and has White Wood_Veneer floor. Moreover, livingroom1 has Yellow Marble floor as well as wall is Wall_Cloth and Black. livingroom1 is in center with 26 square meters. Besides, storage1 is in northwest with 9 square meters. storage1 has Wood_color Wood_Grain floor while wall is White Wall_Cloth. bedroom1 has 13 squares in southwest. bedroom1 uses Black Log for floor, and wall is White Wall_Cloth. bedroom2 uses Black Log for floor while has White Wall_Cloth wall. bedroom2 has 7 squares in northeast. Moreover, kitchen1 uses White Wood Veneer for floor, and wall is Coating and Yellow. kitchen1 has 5 squares in north. balcony1 has 5 squares in south. balcony1 has Black Wall_Cloth wall as well as has Yellow Marble floor. livingroom1 is adjacent to bedroom1, storage1, bedroom2, washroom1, balcony1, kitchen1. washroom1, balcony1 and bedroom1 are connected. storage1 is next to washroom1, kitchen1. bedroom2 is adjacent to kitchen1.</p>	 <p>A 2D floor plan showing the layout of the house. The rooms are labeled: Storage, Kitchen, Secondary Bedroom, Bathroom, Living & Dining Room, Master Bedroom, and Balcony. The layout is consistent with the linguistic requirements.</p>	 <p>A 3D perspective view of the house generated by 'Ours'. The rooms are color-coded: Storage (green), Kitchen (yellow), Secondary Bedroom (purple), Bathroom (blue), Living & Dining Room (yellow), Master Bedroom (purple), and Balcony (black). The layout is consistent with the linguistic requirements.</p>	 <p>A 2D floor plan showing the layout of the house, identical to the 'Ours' version. The rooms are labeled: Storage, Kitchen, Secondary Bedroom, Bathroom, Living & Dining Room, Master Bedroom, and Balcony.</p>	 <p>A 3D perspective view of the house generated by 'Ground-truth'. The rooms are color-coded: Storage (green), Kitchen (yellow), Secondary Bedroom (purple), Bathroom (blue), Living & Dining Room (yellow), Master Bedroom (purple), and Balcony (black). The layout is consistent with the linguistic requirements.</p>

Conclusion

Contributions

- We propose a **novel architecture**, called House Plan Generative Model (HPGM), which generates 3D house models with given linguistic expressions. To reduce the difficulty, we divide the generation task into two sub-tasks to generate floor plans and interior textures, separately.
- To achieve the goal of synthesizing 3D building model from the text, we collect **a new dataset** consisting of the building layouts, texture images, and their corresponding natural language expressions.
- **Extensive experiments** show the effectiveness of our method on both qualitative and quantitative metrics. We also study the generalization ability of the proposed method by generating unseen data with the given new texts.

REVERIE: Remote Embodied Visual Referring Expressions in Real Indoor Environments

Yuankai Qi^{1,2}, Qi Wu¹, Peter Anderson³, Xin Wang⁴, William Yang Wang⁴,
Chunhua Shen¹, Anton van den Hengel¹

¹Australian Centre for Robotic Vision, The University of Adelaide, Australia

²Harbin Institute of Technology, Weihai, China

³Georgia Institute of Technology, USA ⁴University of California, Santa Barbara, USA

A Long-hold Goal



Build intelligent robots that can **perceive** the environment, **execute** commands, and **communicate** with human.

A New Task

- **However**, many of the most appealing uses of robots require communication about **remote objects**.

Examples:

“**Bring** me a blue cushion from the living room”

“**Clean** the round table in the dining room”

REVERIE: Remote **E**mbodied **V**isual Referring
Expressions in **R**eal **I**ndoor **E**nvironments

The REVERIE Task



R2R vs. REVERIE

Two key difference:

- Fine-grained instructions vs. High-level instruction
 - R2R**: 'Go to the top of the stairs then turn left and walk along the hallway and stop at the first bedroom on your right'
 - REVERIE**: 'the cold tap in the first bedroom on level two'
- Point navigation vs. Remote object grounding

RefExp vs. REVERIE

Three key difference

- Visible target object vs. Invisible target object
- Single candidate image vs. Panoramas of all possible viewpoints
- Front view vs. Various Views

RefExp



REVERIE



Instruction Examples

1 Fold the towel in the bathroom **with the fishing theme**

2. Push in the bar chair, in the kitchen, **by the oven.**

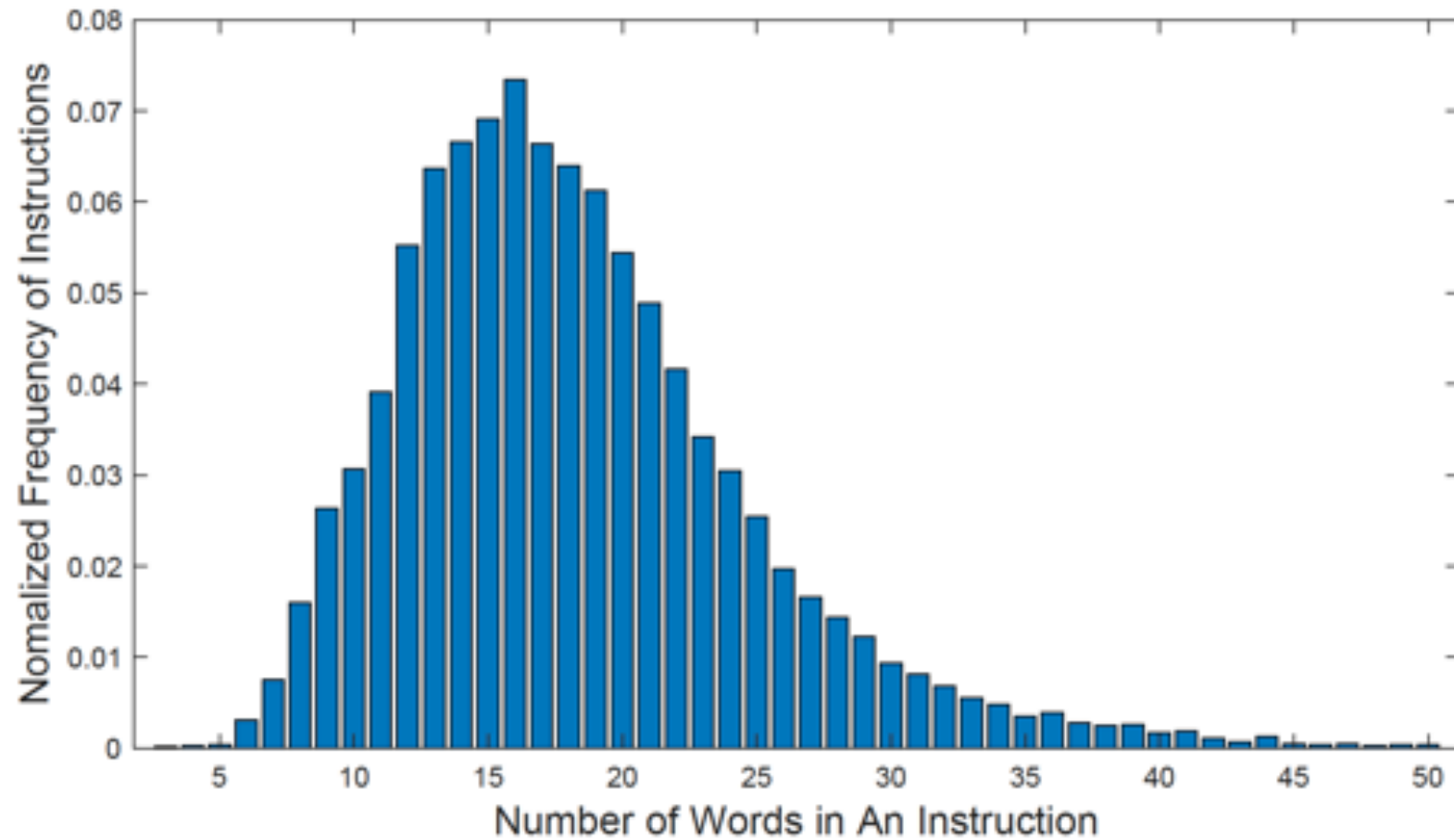
3. Go to the blue family room and bring the framed picture of a person on a horse **at the top left corner above the TV.**

4. **Could** you please dust the light above the toilet in the bathroom that is near the entry way?

5 There is a bottle in the office alcove next to the piano. **It** is on the shelf above the sink on the extreme right. Please bring **it** here.

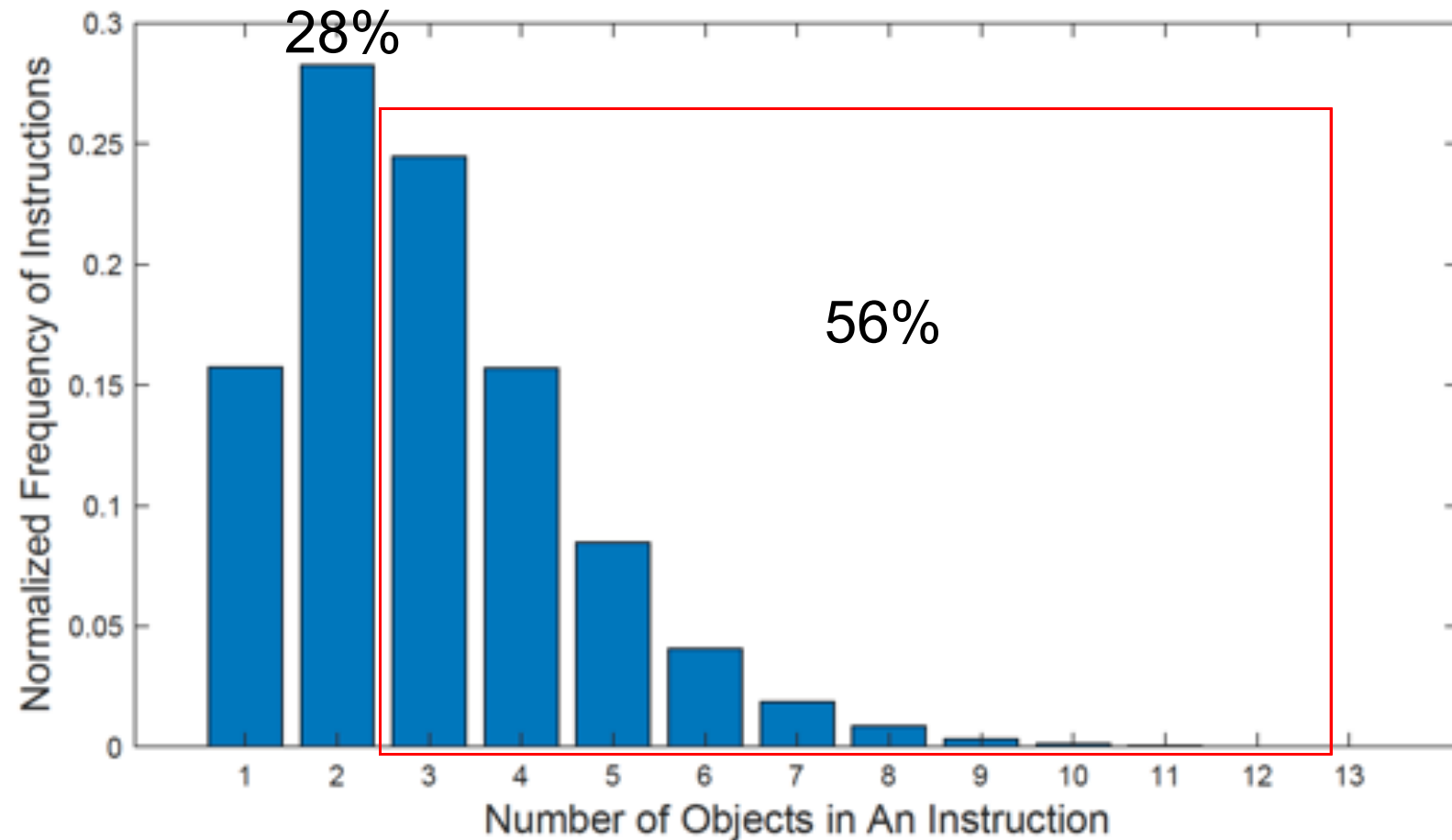
Statistics

- Instruction length



Statistics

- Number of objects in an instruction



Dataset Splits

	Buildings	Instructions	Objects
Train	60	10,466	2,353
Val Seen	46	1,423	440
Val Unseen	10	3,521	513
Test	16	6,292	834

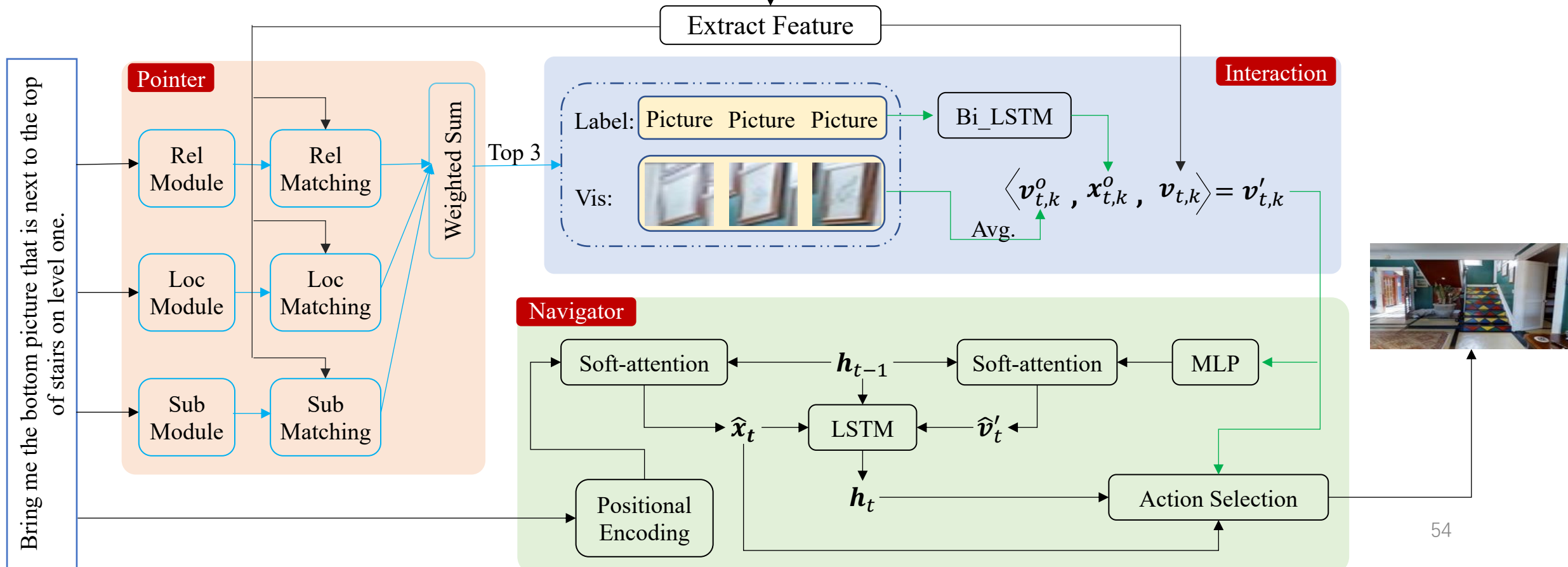
* The split follows the strategy of R2R dataset for research convenience.

Solution

- Navigation Model + Referring Expression Comprehension Model
 - 4 Baseline Navigation Model + 4 SoTA Navigation Model
 - Random
 - Shortest
 - R2R-TF
 - R2R-SF
 - SelfMonitor: Chih-Yao Ma, etal, ICLR 2019
 - RCM: Xin Wang, etal, CVPR 2019
 - FAST-Short: Liyinming Ke, etal, CVPR 2019
 - FAST-Lan-Only: a variant of FAST-Short
 - 1 Baseline RefExp Model + 2 SoTA RefExp Model
 - CNN-RNN
 - MAttNet: Licheng Yu, etal, CVPR 2018
 - CM-Erase: Xihui Liu, etal, CVPR 2019

Solution: Interactive Navigator-Pointer Model

Navigable views



Metrics

- A Successful Task
 - Select the correct object from a list of candidatesOr
 - IoU ≥ 0.5 between predicted bounding box and ground-truth
- Main Metric
 - RGS : Remote Grounding Success rate
- Auxiliary Metric for Navigation
 - Succ: Success rate
 - Osucc: Oracle success rate
 - SPL: Success rate weighted by path length
 - Length: Path length

Results

Success Rate on the REVERIE Task Using MAttNet as Pointer

Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	50.53	55.17	45.50	16.35	31.97	14.40	28.20	7.19	45.28	7.84	19.88	30.63	11.61	39.05	11.28
Human	-	-	-	-	-	-	-	-	-	-	81.51	86.83	53.66	21.18	77.84

Code on GitHub

