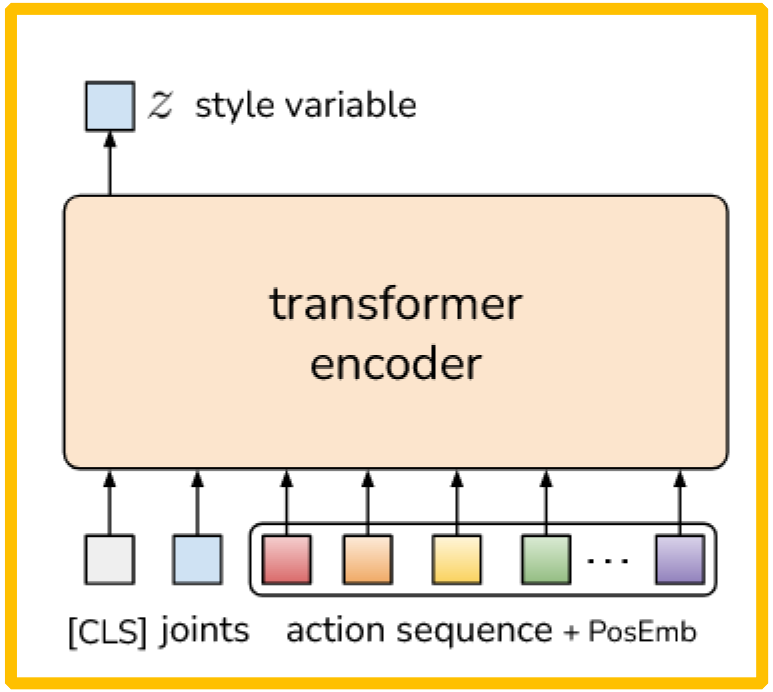
Action Chunking with Transformer(2023.4)

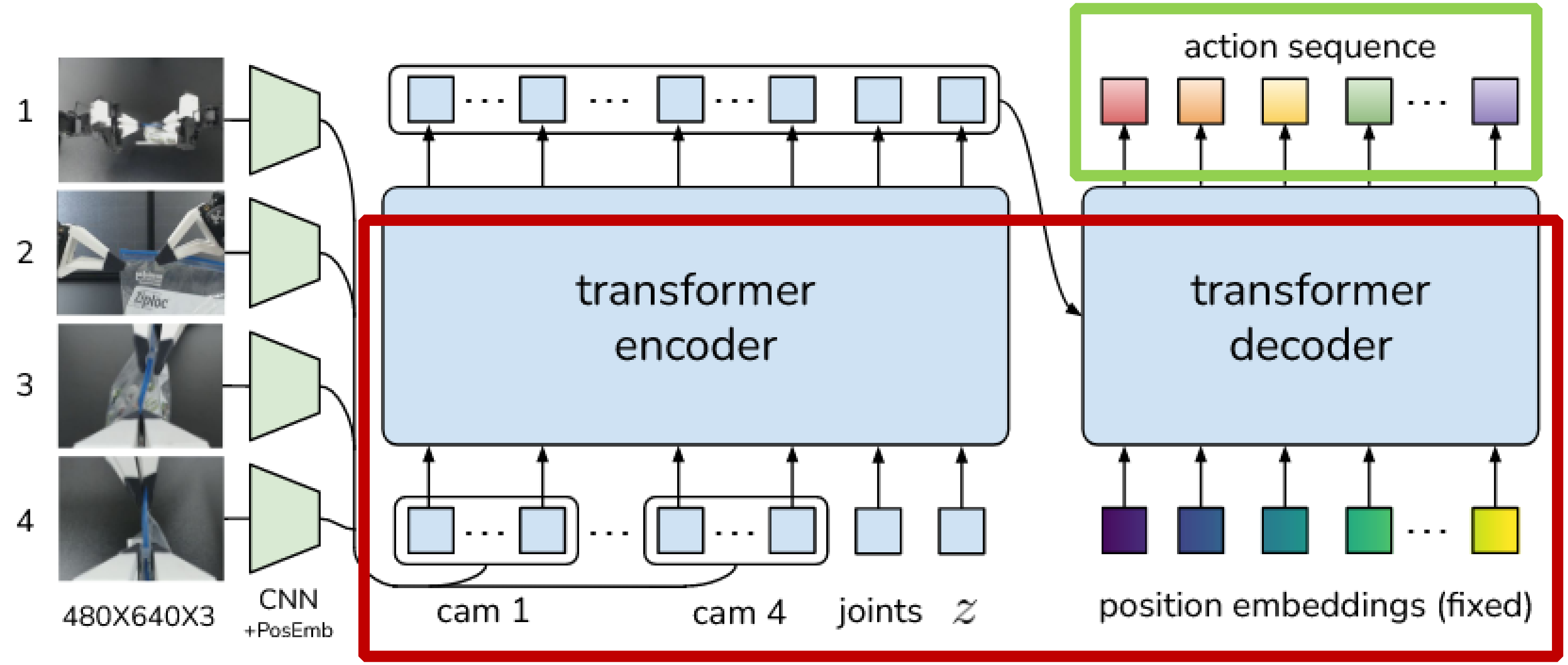
推理时由于没有先验数据丢弃



首先推断CVAE生成的条件变量 z

减少演示数据的非平稳性

生成动作块
减少高频误差



Transformer生成动作序列

Abstract

2022

2022年的SOTA

题目：Think **Global**, Act **Local**: Dual-scale Graph Transformer for **Vision-and-Language Navigation**

VLN：根据语言指令在未知环境中到达精确目标地点

过去的recurrent structure导致指令是一个一个的细节model只允许local action → 需要memory → LSTM

内隐性质导致丢失大量信息 → 直接记忆图片用transformer

针对问题	创新点
原RNN长时序性能有限 信息被压缩，无法精确导航	**引入拓扑图**
BC分布偏移，策略误差放大	**Training中加入pseudo interactive demonstrator**

抛弃RNN，用transformer

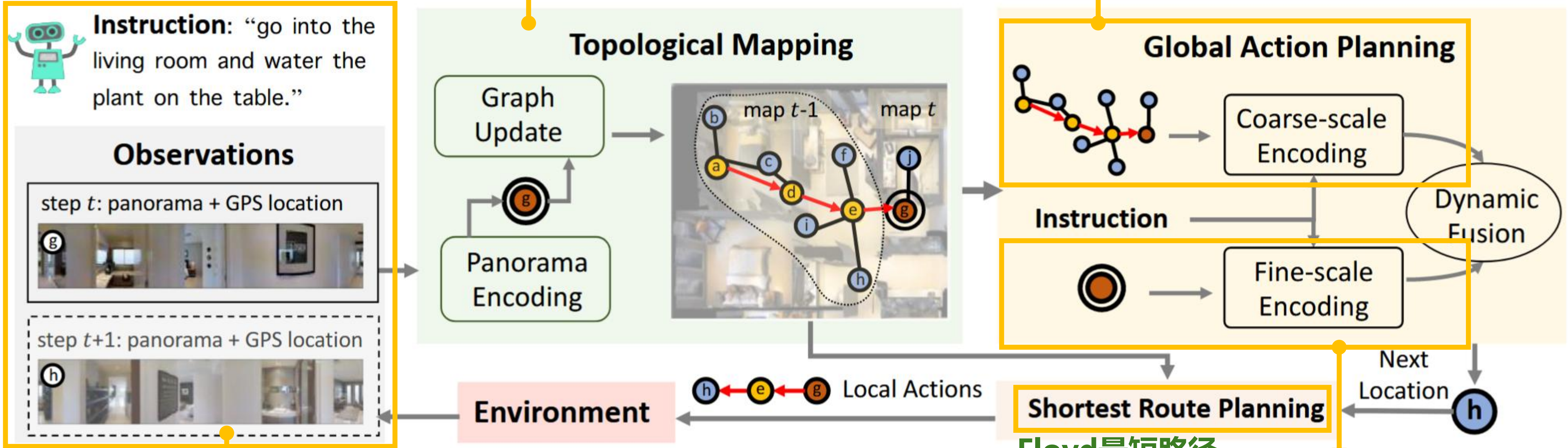
加入粗粒度的节点表示实现dual-scale

trick了一个叫dynamic fusion的方法

Abstract

探索中存储图像构建拓扑图，
是Coarse-scale（粗粒度）的

Coarse-scale
粗粒度的拓扑图，用于global
action（全局规划）



Floyd最短路径

输入信息：Vision+language
全景相机+GPS定位

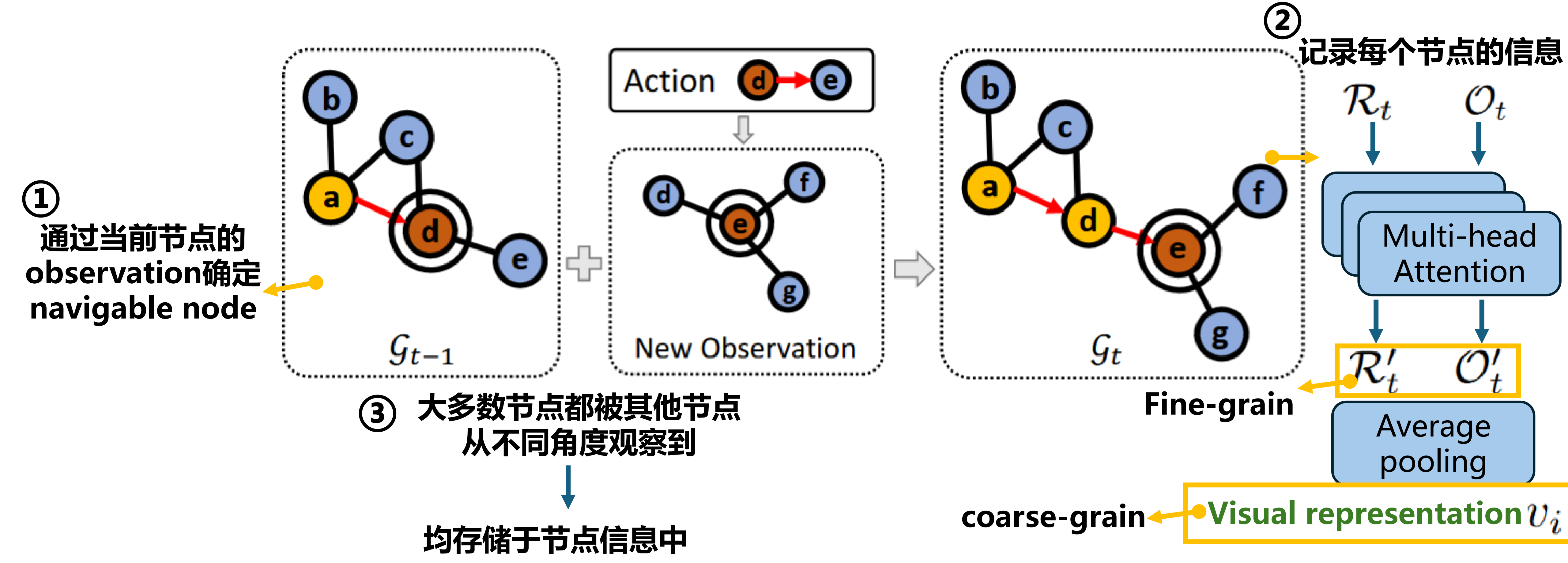
Fine-scale
细粒度的拓扑图，用于local
action（局部规划）

Topological mapping

Goal：理解自然语言指令，穿越整个环境找到指令指定的目标

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ 图，包括节点集和边集 $\mathcal{R}_t = \{r_i\}_{i=1}^{n}$ 所有图像信息 $\mathcal{N}(V_t)$ 当前节点的邻居节点

$\mathcal{W} = \{w_i\}_{i=1}^{L}$ 语言指令信息 $\mathcal{A}_t$ 动作空间 $\mathcal{O}_t = \{o_i\}_{i=1}^{m}$ 语义分割的物体信息



Core

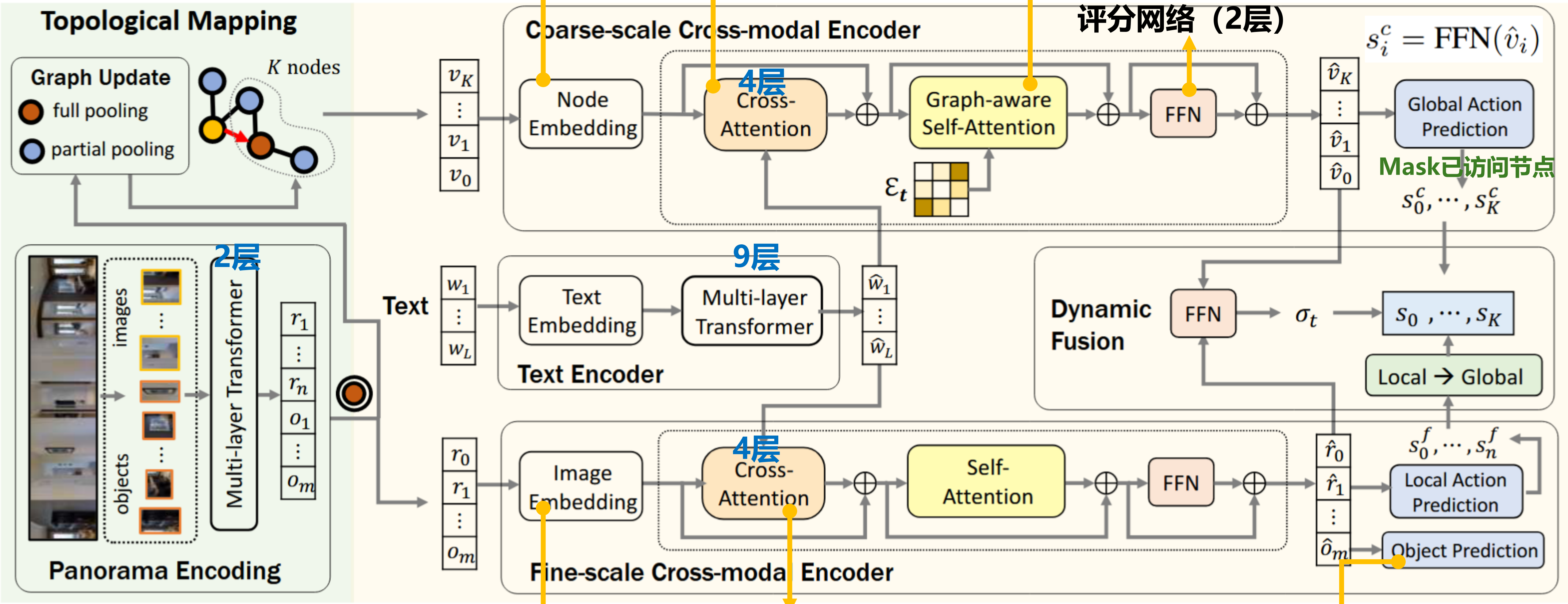
$$\mathrm{GASA}(X) = \mathrm{Softmax}\left(\frac{XW_q(XW_k)^T}{\sqrt{d}} + M\right)XW_v,$$

$$M = EW_e + b_e,$$

嵌入相对位置和最近
访问时间步

建模Node与
Instruction的关系

不仅仅考虑视觉相关性
加入几何表示



嵌入相对起点位置和
相对邻居位置

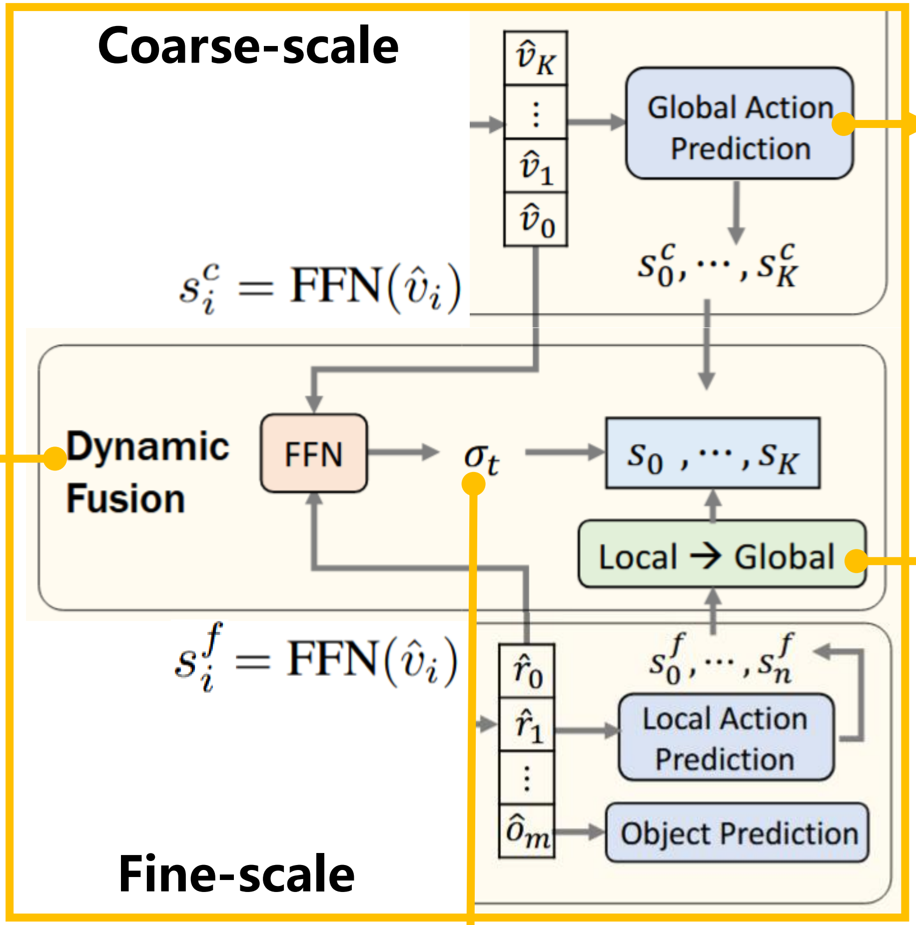
建模图像与
Instruction的关系

若识别物体，则发送stop

$$s_i^f = \mathrm{FFN}(\hat{v}_i)$$

Dynamic Fusion

$$\sigma_t = \mathrm{Sigmoid}(\mathrm{FFN}([\hat{v}_0; \hat{r}_0]))$$



Softmax

回溯机制

$$s_i^{f'} = \begin{cases} s_{\text{back}}, & \text{if } V_i \in \mathcal{V}_t - \mathcal{N}(V_t), \\ s_i^f, & \text{otherwise.} \end{cases}$$

$$s_i = \sigma_t s_i^c + (1 - \sigma_t) s_i^{f'}$$

Training and inference

Training

Pretrain

离线专家演示数据
简单视觉-语言代理任务

- masked language modeling (MLM)
- masked region classification (MRC)
- single-step action prediction (SAP)
- object grounding (OG)

$$L_{\mathrm{SAP}} = \sum_{t=1}^{T} -\log p(a_t^* | \mathcal{W}, \mathcal{P}^*_{<t})$$

$$L_{\mathrm{OG}} = -\log p(o^* | \mathcal{W}, \mathcal{P}_T)$$

BC的Loss

Fine-tune

防止BC的distribution shift

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$

全局图

expert

π^*

专家预测

$$L_{\mathrm{PID}} = \sum_{t=1}^{T} -\log p(a_t^{\pi^*} | \mathcal{W}, \mathcal{P}_{<t})$$

PID（伪监督）的Loss

inference

timestep

- 更新拓扑图
- 预测action
- Floyd最短路径
- Goal/max steps stop

$$L = \lambda L_{\mathrm{SAP}} + L_{\mathrm{PID}} + L_{\mathrm{OG}}$$

总的Loss

Experiment

metrics

Trajectory Length (TL)

Navigation Error (NE)

Success Rate (SR)

Oracle SR (OSR)

SR penalized by Path Length (SPL)

details

image — ViT-B/16 — 朝向特征用sin cos表示

REVERIE — Train — Batch size 32
Iteration 100k
2 Nvidia Tesla P100 GPUs

合成指令扩充数据集 — Fine-tune — Batch size 8
Iteration 20k
1 Nvidia Tesla P100 GPU

ablation

核心创新点

scale	fusion	OSR↑	SR↑	$\frac{\text{SR}}{\text{OSR}}\uparrow$	SPL↑	RGS↑	RGSPL↑
fine	-	30.96	28.86	**93.22**	23.57	20.39	16.64
coarse	-	46.44	36.52	78.64	25.98	-	-
multi	average	**51.86**	45.81	88.33	31.94	**32.49**	22.78
multi	dynamic	51.07	**46.98**	91.40	**33.73**	32.15	**23.03**

GASA重要性

Fusion	GASA	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
average	×	49.22	44.50	30.90	29.88	20.73
average	✓	**51.86**	45.81	31.94	32.49	22.78
dynamic	×	49.25	45.24	32.88	29.91	21.57
dynamic	✓	51.07	**46.98**	**33.73**	**32.15**	**23.03**

Experiment

ablation

Loss

Pretrain SAP	Pretrain OG	Pretrain Aux	Finetune RL	Finetune PID	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
✓	×	×	×	×	38.45	35.30	24.55	-	-
✓	✓	×	×	×	40.24	37.80	26.40	23.89	16.36
✓	✓	✓	×	×	37.63	36.81	27.19	25.05	18.40
✓	✓	✓	✓	×	47.51	42.35	32.97	29.91	**23.53**
✓	✓	✓	×	✓	**51.07**	**46.98**	**33.73**	**32.15**	23.03

Aug生成数据

PID	Aug	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
×	×	37.29	34.56	25.56	23.00	16.64
	✓	37.63	36.81	27.19	25.05	18.40
✓	×	51.07	**46.98**	**33.73**	**32.15**	**23.03**
	✓	**52.09**	46.58	32.72	31.75	22.18

预训练中嘈杂的合成数据训练了辅助代理任务MLM, MRC

微调中的策略学习仍需要纯净数据