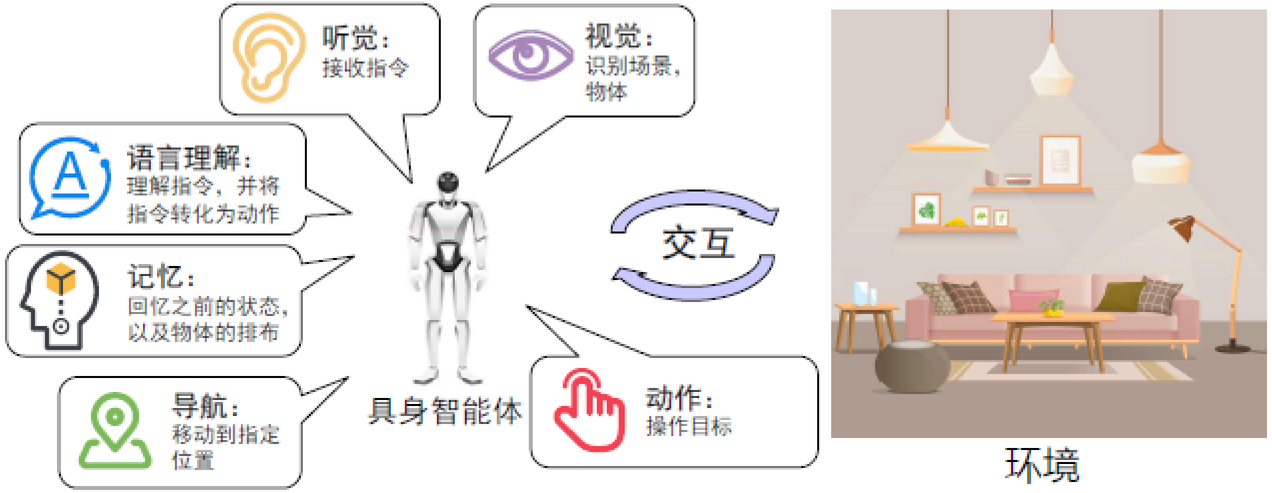
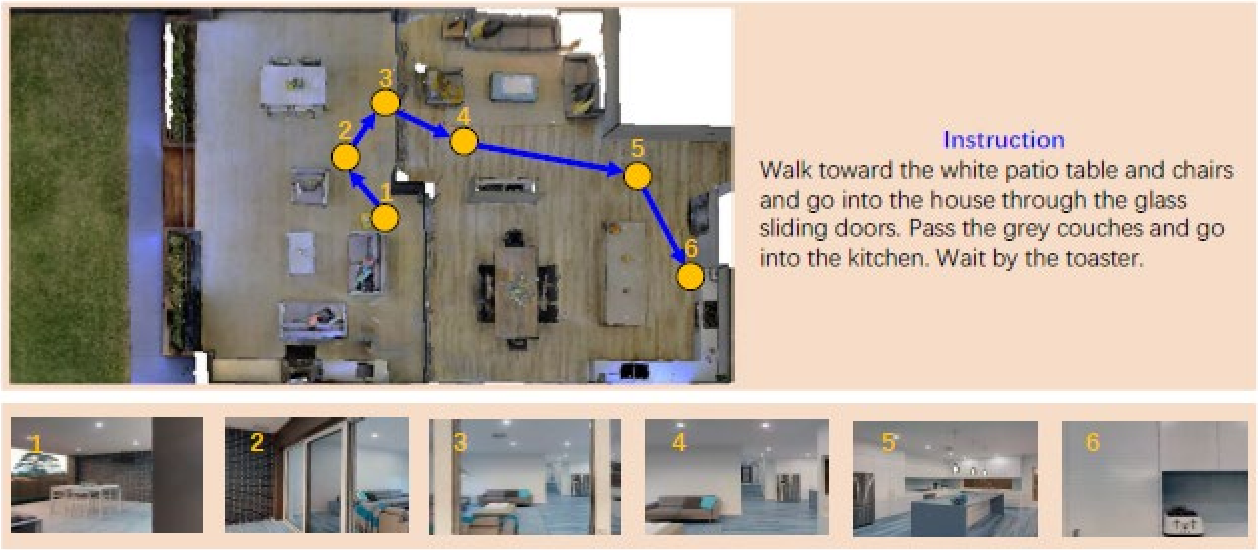
Embodied AI

- 莫拉维克悖论：人工智能系统在执行任务时的反常现象，对于人类而言简单的任务，比如感知和运动控制，对机器却极其困难；而复杂的任务，比如逻辑推理和数学计算，对机器来说却相对容易。
- 研究人员逐步探索人工智能理解物理世界并与之交互的能力，即所谓的具身智能。
- 其中，视觉语言导航(Vision-and-Language Navigation, VLN)融合了自然语言处理、计算机视觉和机器人技术，其目的是通过理解自然语言指令，使智能体能够在虚拟或真实环境中移动到指定位置，为更自然高效的现实世界人机交互铺平道路。



VLN



- 在常见的视觉语言导航(VLN)任务中，智能体接收一条自然语言指令，通过不断观察周围环境并执行动作，移动到指定位置。

- 环境被建模为一个无向拓扑图$\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$，其中$\mathcal{V}$表示可导航的节点，$\mathcal{E}$表示连接边。

- 任务开始时被放置于未见过环境中的一个起始节点。在每一个导航的时间步t，智能体接收到当前所处节点$\mathcal{V}_t$的全局视觉观察，全景观察包含36个独立的视角图像$\mathrm{R}_t=\{\mathrm{r}_{t,\mathrm{i}}\}_{\mathrm{i}=1}^{36}$，而后智能体根据导航指令$\mathcal{W}$和视觉观察$\mathrm{R}_t$从可导航的候选移动方向中选择一个移动到达相邻的其它节点。

PRET: Planning with Directed Fidelity Trajectory for Vision and Language Navigation

Renjie Lu[1], Jingke Meng[1(✉)], and Wei-Shi Zheng[1,2,3]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

lurj3@mail2.sysu.edu.cn,mengjke@gmail.com, wszheng@ieee.org

[2] Peng Cheng Laboratory, Shenzhen, China

[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

ECCV 2024

Motivation

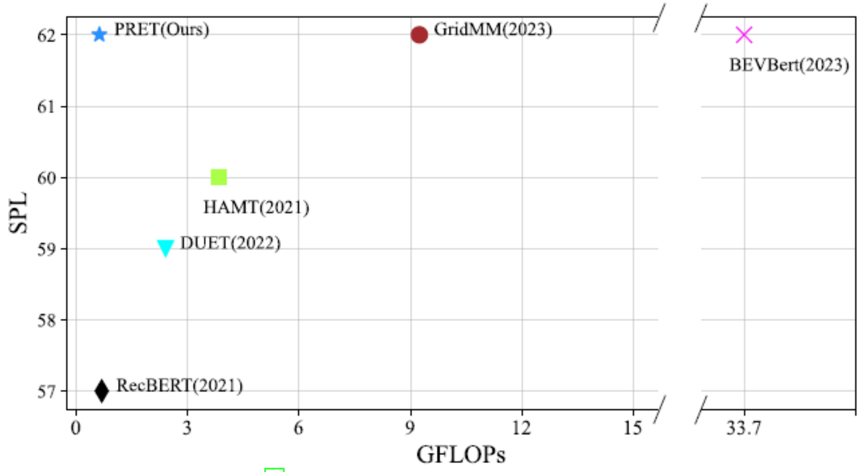
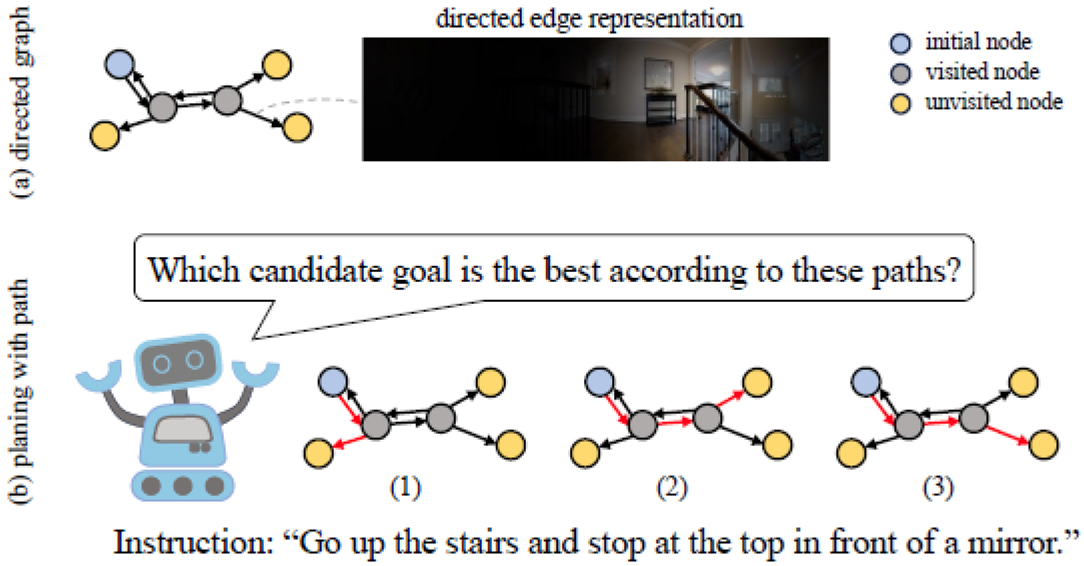
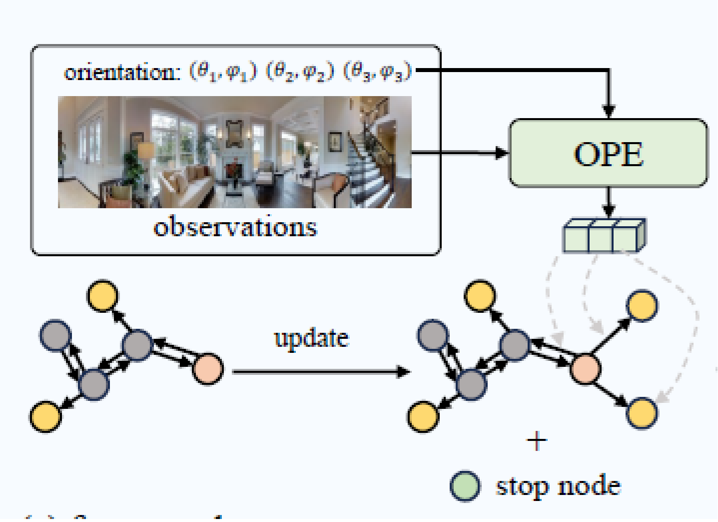


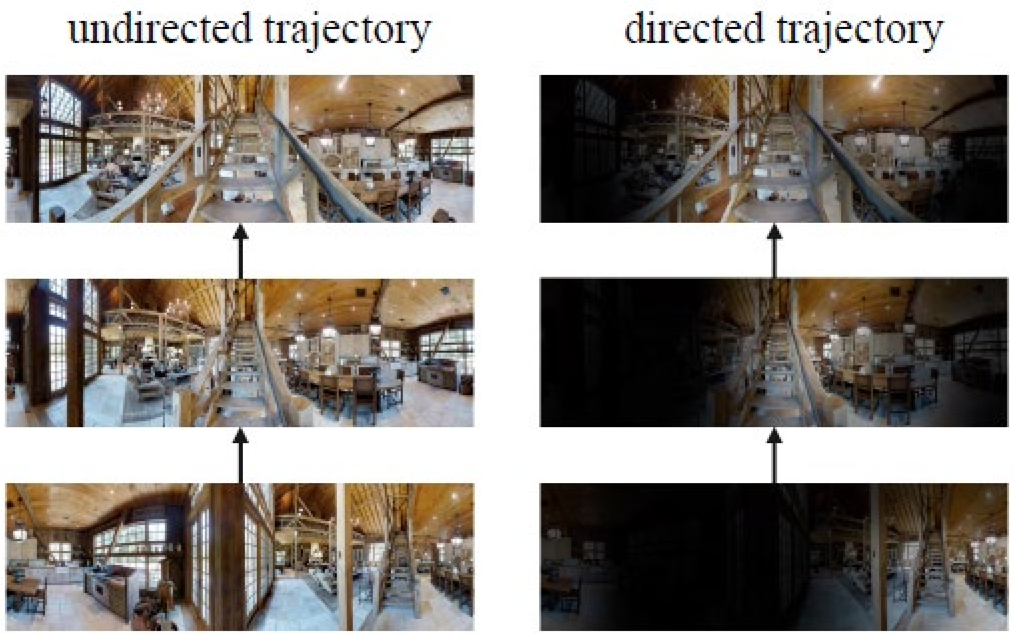
Fig. 1: Comparison of SPL [3] and GFLOPs on R2R test unseen split dataset. Our method is comparable with previous SOTA methods while being more computational efficient. The computational cost of text encoder and visual encoder is omitted for fair comparison.



Method

- Orientation-aware Panorama Encoder





Instruction: Turn right and go up the stairs.

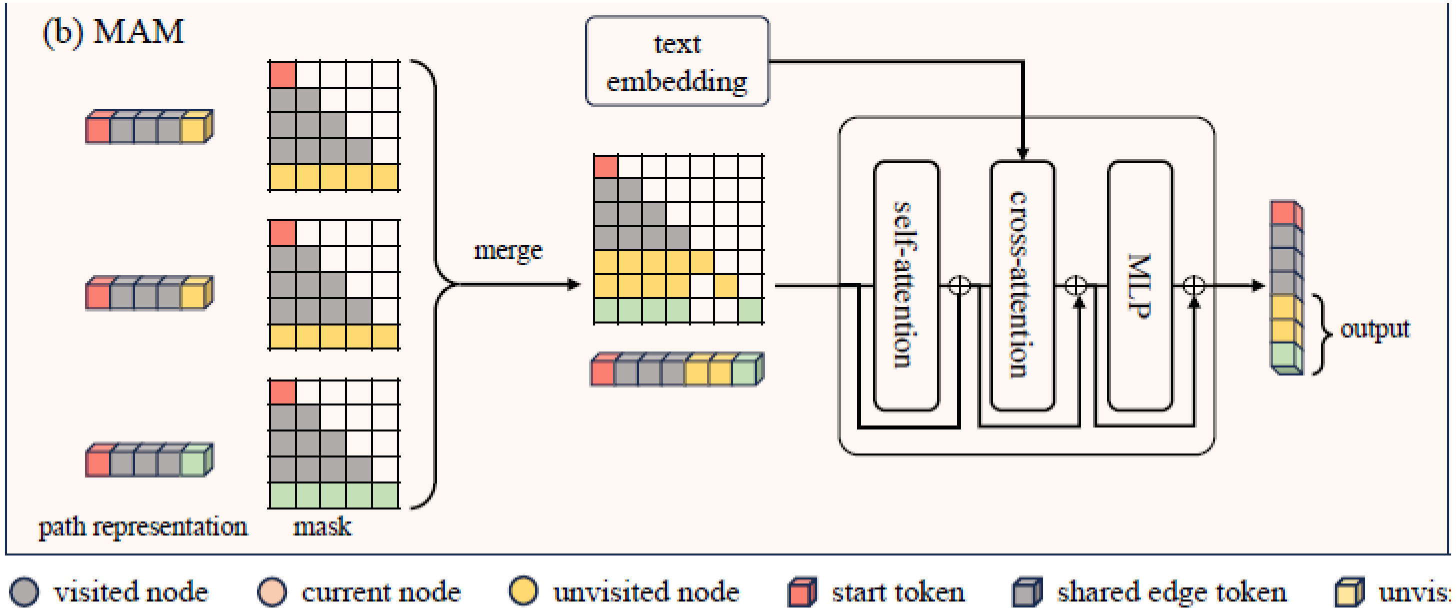
$$x^a = \left[\sin(\phi), \cos(\phi), \sin(\theta), \cos(\theta)\right] W^a, \quad (1)$$

$$x^p_{t,i} = [r_{t,i}; \sin(\phi_{t,i}), \cos(\phi_{t,i}), \sin(\theta_{t,i}), \cos(\theta_{t,i})] W^p. \quad (2)$$

$$E_t = \mathrm{TransformerDecoder}(X^a_t, X^p_t), \quad (3)$$

Method

- Matching Assessment Module



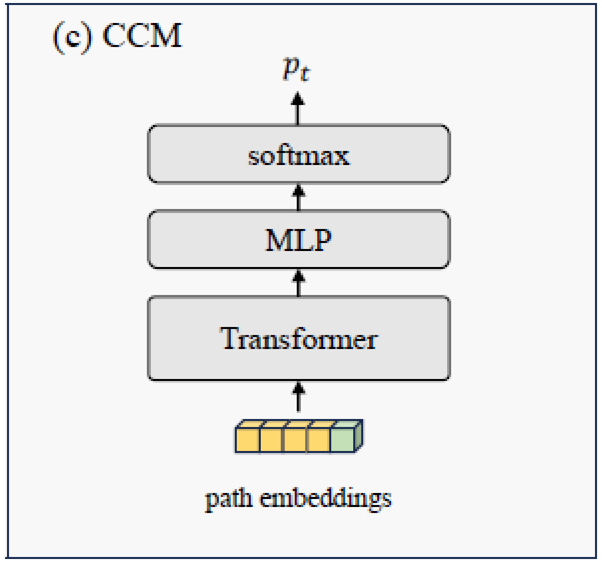
$$X^h = [e_1, e_2, \cdots, e_l]$$

$$X^{h\prime} = X^h + P_l,$$

$$X^o = \text{TransformerDecoder}(X^{h\prime}, X^w), \tag{4}$$

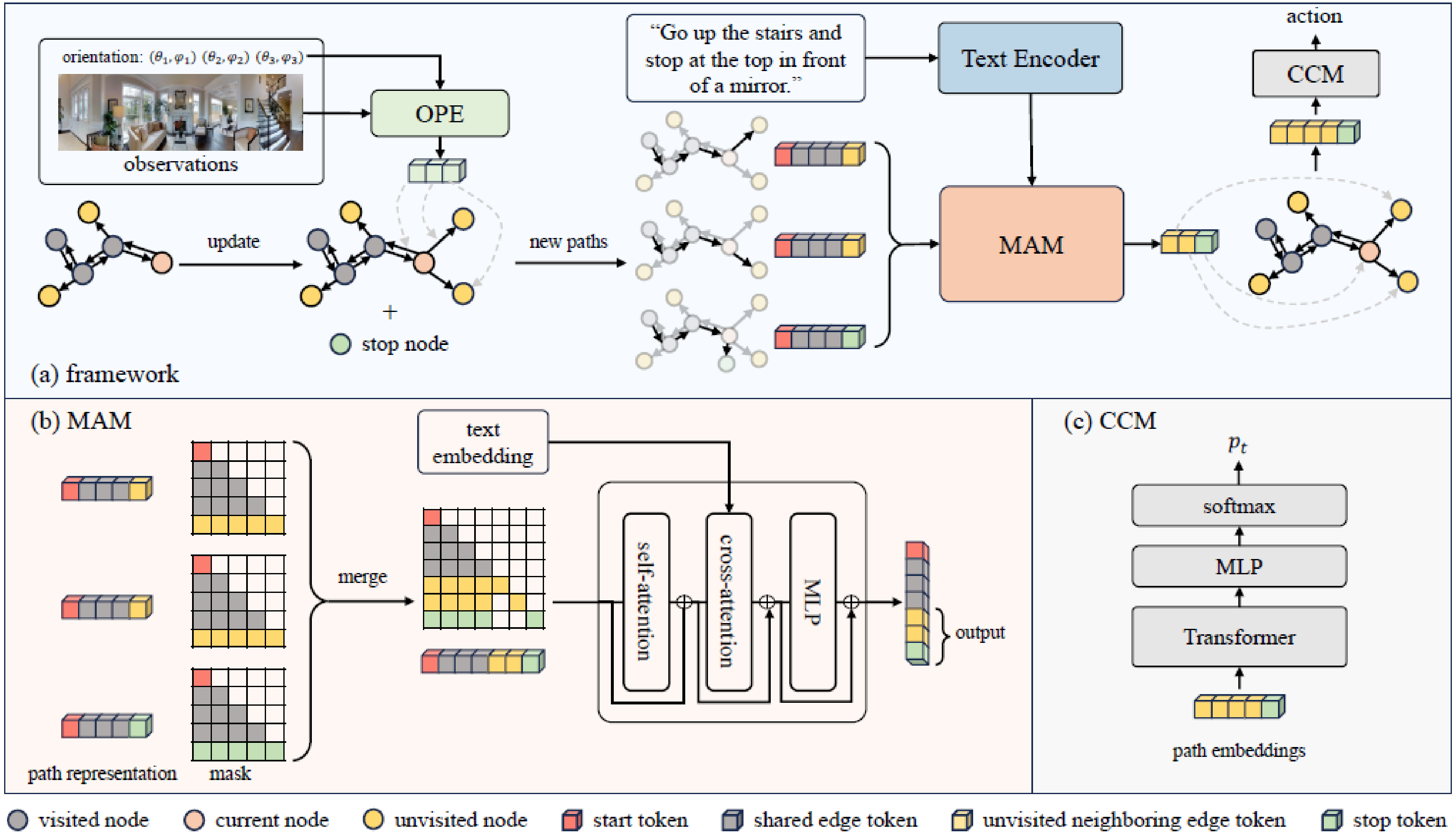
Method

- Candidate Comparison Module



$$
\begin{aligned}
X_t^{e\prime} &= \mathrm{TransformerLayer}(X_t^e), \\
s_t &= \mathrm{MLP}(X_t^{e\prime}), \\
p_t &= \mathrm{softmax}(s_t).
\end{aligned} \tag{5}
$$

Method


orientation: (θ_1, φ_1) (θ_2, φ_2) (θ_3, φ_3)
observations
OPE
update
+
stop node
new paths
"Go up the stairs and stop at the top in front of a mirror."
Text Encoder
MAM
action
CCM
(a) framework
(b) MAM
text embedding
merge
self-attention
cross-attention
MLP
output
path representation
mask
(c) CCM
p_t
softmax
MLP
Transformer
path embeddings
visited node
current node
unvisited node
start token
shared edge token
unvisited neighboring edge token
stop token

Experiments

Table 1: Comparison with other methods on R2R dataset. SPL is considered as the primary evaluation metric.

Methods	Val Seen				Val Unseen				Test Unseen			
	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑
Seq2Seq-SF [4]	11.33	6.01	39	-	8.39	7.81	22	-	8.13	7.85	28	18
Speaker-Follower [14]	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
RCM [46]	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
Regretful [32]	-	3.23	69	63	-	5.32	50	41	-	5.69	56	40
EnvDrop [40]	11.00	3.99	62	59	10.70	5.22	52	48	11.66	5.23	51	47
PREVALENT [17]	10.32	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
NvEM [1]	11.09	3.44	69	65	11.83	4.27	60	55	12.98	4.37	58	54
SSM [44]	14.70	3.10	71	62	20.70	4.32	62	45	20.40	4.57	61	46
RecBert [18]	11.13	2.90	72	68	12.01	3.93	63	57	12.35	4.09	63	57
HAMT [7]	11.15	2.51	76	72	11.46	2.29	66	61	12.27	3.93	65	60
MTVM [28]	-	2.67	74	69	-	3.73	66	59	-	3.85	65	59
DUET [8]	12.32	2.28	79	73	13.94	3.31	72	60	14.73	3.65	69	59
AZHP [16]	-	-	-	-	14.05	3.15	72	61	14.95	3.52	71	60
Meta-Explore [20]	11.95	**2.11**	**81**	**75**	13.09	3.22	72	62	14.25	3.57	71	61
GridMM [47]	-	-	-	-	13.27	2.83	75	64	14.43	3.35	73	62
BEVBert [2]	13.56	2.17	**81**	74	14.55	**2.81**	**75**	64	15.87	3.13	**73**	62
Ours(CLIP)	11.48	2.60	74	69	12.21	3.12	71	63	13.87	3.12	72	62
Ours(DINOv2)	11.25	2.41	78	72	11.87	2.90	74	**65**	12.21	**3.09**	72	**64**

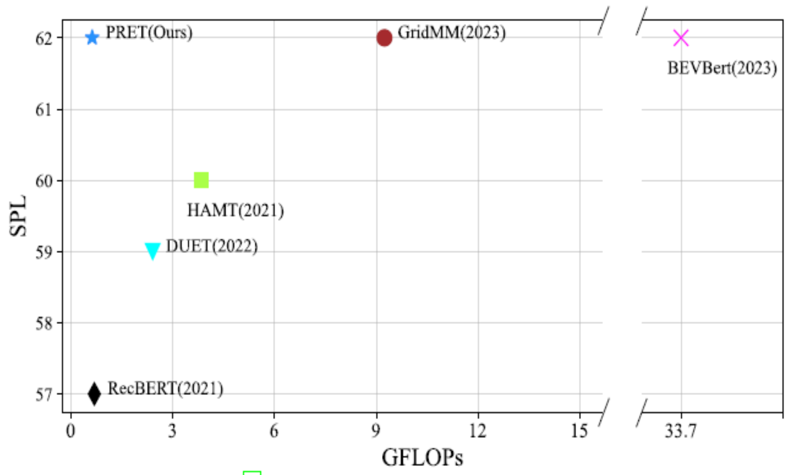


Fig. 1: Comparison of SPL [3] and GFLOPs on R2R test unseen split dataset. Our method is comparable with previous SOTA methods while being more computational efficient. The computational cost of text encoder and visual encoder is omitted for fair comparison.

Experiments

Table 4: Comparison of undirected and directed path representation.

Methods	TL	NE↓	SR↑	SPL↑
undirected	15.62	3.59	68.28	56.77
directed	11.87	2.90	73.78	65.16

Table 5: Ablation study on modules.

	Methods	TL	NE↓	SR↑	SPL↑
1	MAM	12.04	3.99	62.32	54.48
2	MAM+CCM	12.15	3.54	65.94	57.32
3	MAM+OPE	12.18	3.15	71.60	63.07
4	MAM+OPE+CCM	11.87	2.90	73.78	65.16

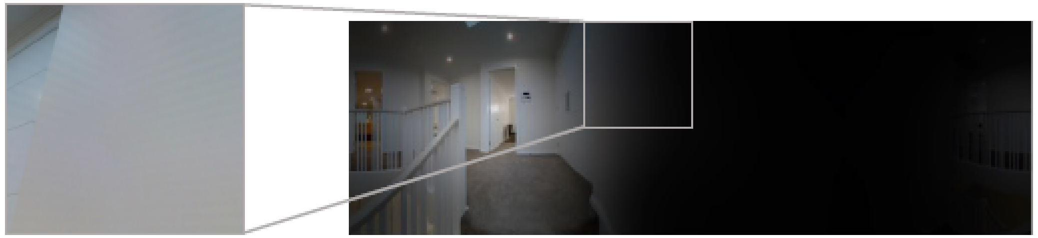


Fig. 4: Comparison of orientation panoramic view and single candidate view.

Exploring Temporal Concurrency for Video-Language Representation Learning

Heng Zhang[1,2†] Daqing Liu[3] Zezhong Lv[1,2] Bing Su[1,2‡] Dacheng Tao[4]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3] JD Explore Academy, JD.com [4] The University of Sydney
zhangheng@ruc.edu.cn, {liudq.ustc,zezhonglv0306,subingats,dacheng.tao}@gmail.com

ICCV 2023

Motivation

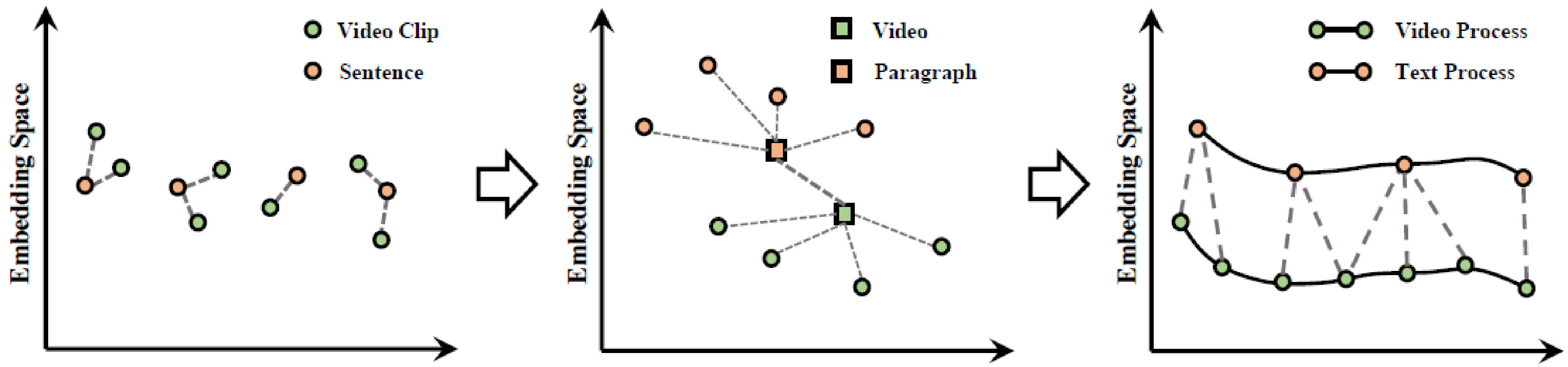
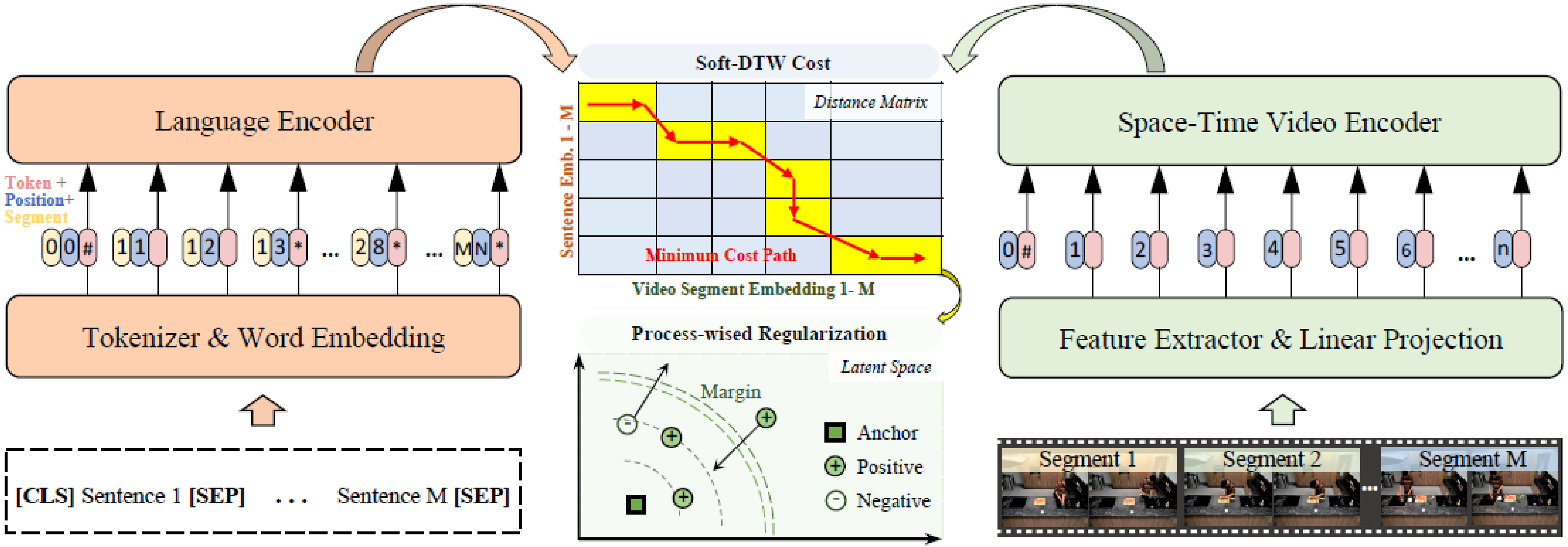


Figure 1. Compare to mainstream video-language representation learning methods. **(a)** *Semantic Alignment* (*e.g.*, HERO [28], Frozen [1]) enforces video-clip sentence pairs to be close in the embedding space, disrupting the inherent temporal dynamics of each modality. **(b)** *Context Dependency* (*e.g.* HD-VILA [50], MERLOT [54]) endows short-range temporal context dependency within each modality, limited on capturing long-range dependencies. **(c)** The proposed *Temporal Concurrency* models video-language pairs as temporal concurrency processes, therefore capturing temporal alignments while maintaining the coherence of each modality.

Method

- Cross-modal Sequence Alignment



$$
\begin{aligned}
d(1,1) &= D_{1,1}, \\
d(i,1) &= D_{i,1} + d(i-1,1), \\
d(1,j) &= D_{1,j} + d(1,j-1),
\end{aligned} \tag{3}
$$

where $i \in [2, N]$, $j \in [2, M]$. Then the distance matrix D can be calculated with the dynamic programming:

$$
d(i,j) = D_{i,j} + min\left\{d(i,j-1), d(i-1,j), d(i-1,j-1)\right\} \tag{4}
$$

$$
min^s(d_1, d_2, ..., d_n) = -\lambda \log \sum_{i=1}^{n} e^{\frac{-d_i}{\lambda}}, \tag{5}
$$

$$
\mathcal{L}_{V2P} = \langle S, \hat{D} \rangle \tag{6}
$$

Method

- Intra-modal Sequence Modeling
  - Brownian Bridge Process

$$p(z_t|z_A, z_T) = \mathcal{N}\left((1-\alpha)z_A + \alpha z_T, \alpha(T-t)\right), \quad \text{where } \alpha = \frac{t-A}{T-A}. \tag{7}$$

  - Process-wised Regularization

$$d(z_A, z_t, z_T) = \frac{1}{2\sigma^2}\left\|z_t - (1-\alpha)z_A - \alpha z_T\right\|_2^2, \quad \text{where } \alpha = \frac{t-A}{T-A}. \tag{8}$$

$$\mathcal{L}_{PRT} = [d(z_A, z_t, z_T) - d(z_A, \hat{z}_t, z_T) + \beta]_+ \tag{9}$$

$$\mathcal{L}(V) = \sum_{j=1}^{M}\sum_{t=A+1}^{T-1}[d(v_A, v_t, v_T) - d(v_A, \hat{v}_t, v_T) + \beta]_+^j \tag{10}$$

$$\mathcal{L}(P) = \sum_{j=2}^{M-1}[d(p_1, p_j, p_M) - d(p_1, \hat{p}_j, p_M) + \beta]_+ \tag{11}$$

Published as a conference paper at ICLR 2022

Language modeling via stochastic processes

Rose E. Wang, Esin Durmus, Noah Goodman, Tatsunori B. Hashimoto
Stanford University
{rewang, edurmus, ngoodman,thashim}@stanford.edu

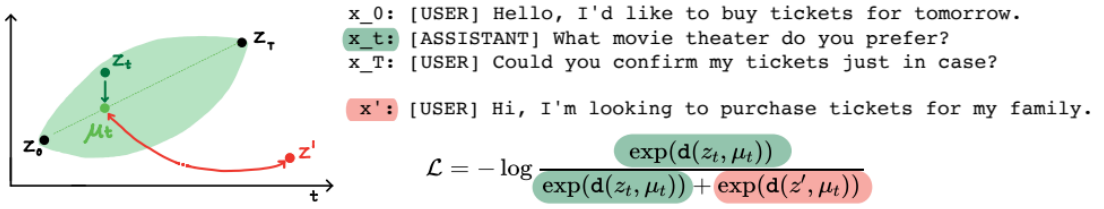
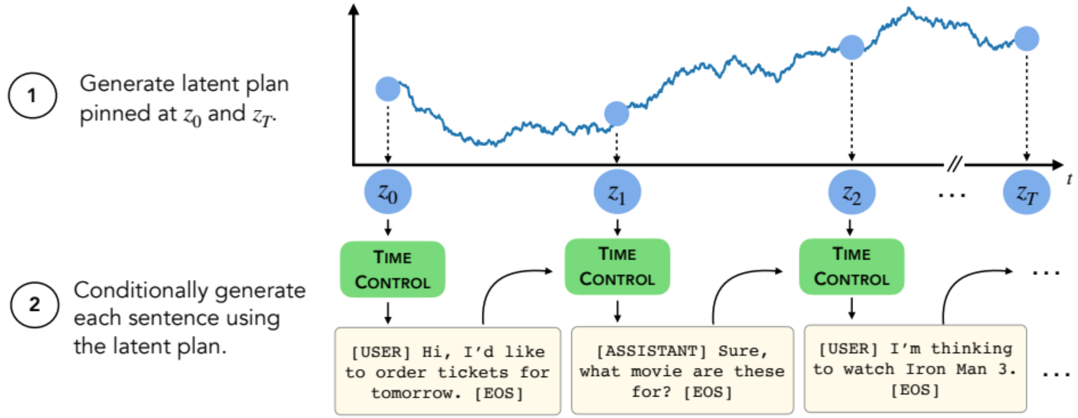


Figure 1: Latent space for a positive triplet of sentences (x_0, x_t, x_T) that are part of the same conversation. Time Control maps positive triplets to a smooth Brownian bridge trajectory. It embeds z_t close to the expected embedding μ_t pinned by z_0, z_T. The green oval area illustrates the uncertainty over z_t as a function of how close t is to 0 and T. In contrast, a negative random sentence x' from a different conversation is not coherent with x_0 and x_T; thus, it is embedded far from μ_t. This is captured by our contrastive loss, $\mathcal{L}$.



Thanks