一、本项目学术成果

1）本项目发表的期刊论文（按引用次数排序）

[1] Bang Yang, Meng Cao, Yuexian Zou*. Concept-Aware Video Captioning: Describing Videos with Effective Prior Information. IEEE Transactions on Image Processing (TIP), 2023. [IF:10.6] [JCR:Q1] [CCF-A]（期刊论文，引用 58）

[2] Meng Cao, Can Zhang, Long Chen, Mike Zheng Shou, Yuexian Zou*. Deep Motion Prior for Weakly-Supervised Temporal Action Localization. IEEE Transactions on Image Processing（TIP）,2022. [IF:11.041][JRC:Q1][CCF-A]（期刊论文，引用 42）

[3] Fenglin Liu, Xian Wu, Chenyu You, Shen Ge, Yuexian Zou*, Xu Sun. Aligning Source Visual and Target Language Domains for Unpaired Video Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022. [IF:17.861] [JCR:Q1] [CCF-A]（期刊论文，引用 40）

[4] Bang Yang*, Fenglin Liu*, Yuexian Zou, Xian Wu, Yaowei Wang, and David A. Clifton. ZeroNLG: Aligning and Autoencoding Domains for Zero-Shot Multimodal and Multilingual Natural Language Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2024. [IF:17.861] [JCR:Q1] [CCF-A] （期刊论文，引用 24）

[5] Asif Raza, Bang Yang, Yuexian Zou*. Zero-Shot Temporal Action Detection by Learning Multimodal Prompts and Text-Enhanced Actionness. IEEE Transactions on Circuits and Systems for Video Technology（TCSVT）, 2024. [IF:7.1] [JCR:Q1][CCF-B]（期刊论文，引用 13）

2）本项目发表的会议论文（按引用次数排序）

[6] Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Yuexian Zou*. Unsupervised Pre-training for Temporal Action Localization Tasks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, Louisiana, 2022/6/19-2022/6/23. [EI] [CCF-A]（会议论文，引用 87）

[7] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Yuexian Zou*. LocVTP: Video-Text Pre-training for Temporal Localization. European conference on

computer vision (ECCV 2022). Tel-Aviv, Israel. 2022/10/23-2022/10/27. [EI] [CCF-B]（会议论文，引用 85）

[8] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yuexian Zou*. Unify, Align and Refine: Multi-Level Semantic Alignment for Radiology Report Generation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV2023), Paris, France, 2023/10/2-2023/10/6. [EI] [CCF-A] （会议论文，引用 60）

[9] Meng Cao, Can Zhang, Yuexian Zou*. Iterative Proposal Refinement for Weakly-Supervised Video Grounding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, Canada, 2023/6/18-2023/6/22. [EI] [CCF-A] （会议论文，引用 48）

[10] Bang Yang, Tong Zhang, Yuexian Zou*. CLIP Meets Video Captioning: Concept-Aware Representation Learning Does Matter. The 5th Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2022). Shenzhen, China. 2022/10/14-2022/10/17. [EI] （会议论文，引用 31）

[11] Meng Cao, Ji Jiang, Yuexian Zou*. Correspondence Matters for Video Referring Expression Comprehension. 2022 ACM International Conference on Multimedia (ACM MM 2022). Lisbon Portugal. 2022/10/10-2022/10/14. [EI] [CCF-A] （会议论文，引用 30）

[12] Hongxiang Li, Meng Cao, Xuxin Cheng, Zhihong Zhu, Yaowei Li, Yuexian Zou*. Exploiting Auxiliary Caption for Video Grounding. The 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024). Vancouver, Canada. 2024/2/20-2024/2/27. [EI] [CCF-A] （会议论文，引用 28）

[13] Bang Yang#, Fenglin liu#, Xian Wu, Yaowei Wang, Xu Sun, Yuexian Zou*. MultiCapCLIP: Auto-Encoding Prompts for Zero-Shot Multilingual Visual Captioning. The 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, Canada, 2023/7/9-2023/7/14. [EI] [CCF-A]（会议论文，引用 24）

[14] Puzhao Ji, Meng Cao, Yuexian Zou*. Visual Relation-Aware Unsupervised Video Captioning. 31st International Conference on Artificial Neural

Networks(ICANN2022). Bristol, UK. 2022/9/6-2022/9/9. [EI] [CCF-C]（会议论文，引用 16）

[15] Bang Yang, Yong Dai, Xuxin Cheng, Yaowei Li, Asif Raza, Yuexian Zou*. Embracing Language Inclusivity and Diversity in CLIP Through Continual Language Learning. The 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024). Vancouver, Canada. 2024/2/20-2024/2/27. [EI] [CCF-A] （会议论文，引用 12）

[16] Xianwei Zhuang, Hongxiang Li, Xuxin Cheng, Zhihong Zhu, Yuxin Xie, Yuexian Zou*. KDProR: A Knowledge-Decoupling Probabilistic Framework for Video-Text Retrieval. The 18th European Conference on Computer Vision (ECCV 2024), MiCo Milano, Italy, 2024/9/29-2024/10/4. [EI] [CCF-B] （会议论文，引用 10）

[17] Xianwei Zhuang, Xuxin Cheng, Zhihong Zhu, Zhanpeng Chen, Hongxiang Li, Yuexian Zou*. Towards Multimodal-augmented Pre-trained Language Models via Self-balanced Expectation-Maximization Iteration. The 32nd ACM International Conference on Multimedia (ACM MM 24), Melbourne, Australia. 2024/10/28-2024/11/1. [EI] [CCF-A] （会议论文，引用 5）

[18] Puzhao Ji, Bang Yang, Tong Zhang, Yuexian Zou*. Consensus-Guided Keyword Targeting for Video Captioning. The 5th Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2022). Shenzhen, China. 2022/10/14-2022/10/17. [EI] （会议论文，引用 2）

3) 本项目发表的专利

[1] 邹月娴、杨邦，一种面向视觉语言大模型幻觉缓解的注意力对比解码方法，2025，中国，发明专利申请号：202510313574.2，实审；

[2] 邹月娴、李耀伟，一种文本引导的视频-语言跨模态对齐方法、装置及设备，2025，中国，发明专利申请号：202511019736.8，实审

[3] 曹蒙; 邹月娴, 一种基于稀疏候选框的自然语言时序定位方法，2023，中国, 发明专利申请号：202311191588.9，实审；

[4] 邹月娴、李鸿翔，一种面向视频文本检索的相关性感知跨模态对齐方法与装置，2025，中国，发明专利申请号：202510989724.1，实审

[5] 邹月娴、姚子裕，一种基于人体姿态的实时视频动作计数方法，2025，中国，发明专利申请号：202510989719.0，实审

4) 本项目发表的软件著作权

[1] 面向交通场景的自动数据标注系统 V1.0，软件著作权，登记号：2025SR1664922，登记日：20250901

[2] 多模态大模型驱动的交通智能播报系统[简称：交通智能播报系统]V1.0，软件著作权，登记号：2025SR2184818，登记日：20251111

二、本项目主要研究工作

课题组的主要研究内容体现在四个方面：1）动态视觉场景的高效时序运动特征建模方法；2）基于多模态自适应对齐的高效动态视觉场景描述方法；3）基于交互式图网络的多模态高阶语义关系推理；4）端到端视频交通场景描述生成模型设计与实现。

1）动态视觉场景的高效时序运动特征建模方法研究

课题组开展了动态视觉场景的高效时序运动特征建模方法的研究，旨在提出一种任务无关，且所学习到的时序运动表征具备时序不变性、时序平移和尺度等变性的高效动态视觉场景时序运动特征建模方法，以解决时序运动特征提取与下游任务特征不匹配的问题。**a)** 在无监督预训练技术框架下，以视频运动定位为具体任务，设计了一种名为“伪动作定位”（Pseudo Action Localization, PAL）的自监督代理任务，并引入时序等变对比学习范式，通过在合成视频中对齐不同位置和尺度的伪动作特征，强制模型学习对背景干扰鲁棒且对时序变换敏感的特征表示，所提 PAL 伪任务算法流程图如图 1 所示。在 ActivityNet v1.3 数据集上进行了详细的性能评测，由表 1 可见，该方法在时序动作检测任务中取得了 33.4% 的平均 mAP，在动作建议生成任务中取得了 66.8% 的 AUC，这些性能不仅显著优于 MoCo-v2 等已有无监督预训练方法，更超越了同等数据量的有监督分类预训练模型，说明了 PAL 算法的先进性和在实际应用中的有效性。

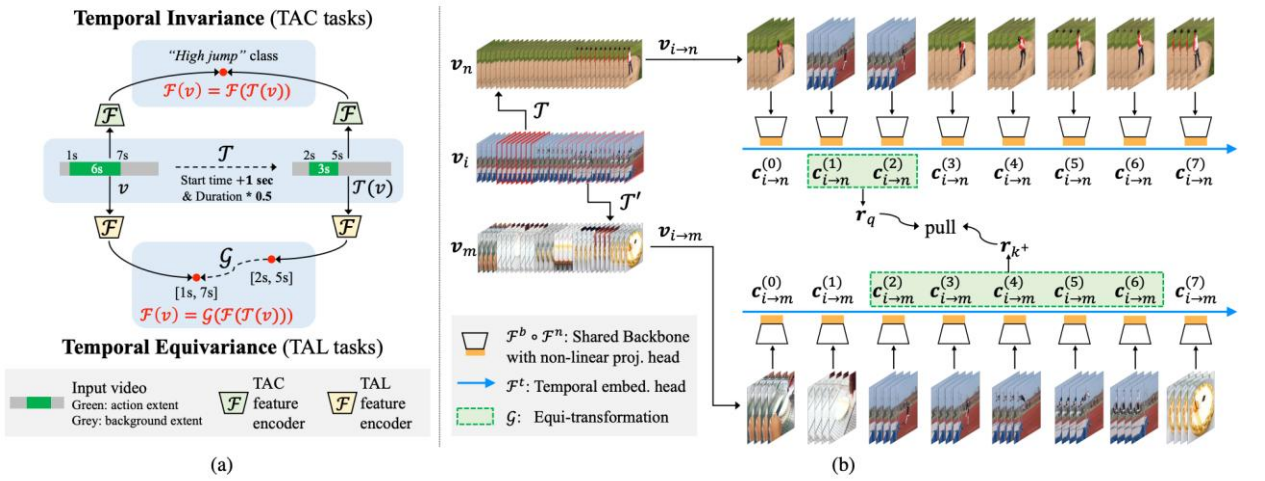

Figure 2. **(a) Schematic depiction of temporal invariance *vs.* temporal equivariance. (b) Overview of our PAL pretext task.** Given

图 1 PAL 伪任务算法流程图

表 1 PAL 在 ActivityNet v1.3 数据集上的实验结果

Method	Modal	Dataset	Backbone	TR×SR^2 (per clip)	FLOPs (per clip)	TAD Task (G-TAD [63])				APG Task (BMN [36])			
						mAP@0.5	@0.75	@0.95	AVG	AR@1	@10	@100	AUC
CoCLR [28]	V+F	K400	S3D	32×128^2	47.2G	47.9	32.2	7.3	31.9	32.7	53.5	73.9	65.0
XDC [2]	V+A	IG65M	R(2+1)D-18	32×224^2	325.2G	48.4	32.6	7.6	32.3	33.2	54.1	74.0	65.4
MoCo-v2 [16] *	V	K400	I3D	8×112^2	3.6G	46.6	30.7	6.3	30.3	30.8	53.5	72.4	64.0
VideoMoCo [44]	V	K400	R(2+1)D-18	32×112^2	81.3G	47.8	32.1	7.0	31.7	31.8	53.9	72.8	65.1
RSPNet [13]	V	K400	R(2+1)D-18	16×112^2	40.6G	47.1	31.2	7.1	30.9	31.5	53.3	72.2	64.1
AoT [57] †	V	K400	TSM-Res50	8×224^2	33G	44.1	28.9	5.9	28.8	-	-	-	-
SpeedNet [5] †	V	K400	TSM-Res50	8×224^2	33G	44.5	29.5	6.1	29.4	-	-	-	-
PAL (Ours)	V	K400	I3D	8×112^2	3.6G	49.3	34.0	7.9	33.4	33.7	55.9	75.0	66.8
PAL (Ours)	V	K700	I3D	8×112^2	**3.6G**	**50.7**	**35.5**	**8.7**	**34.6**	**34.2**	**57.8**	**76.0**	**68.1**
TAC *	V	K400	I3D	8×112^2	3.6G	48.5	32.9	7.2	32.5	32.3	54.6	73.5	65.6
BSP [61]	V	K400	TSM-Res50	8×224^2	33G	50.9	35.6	8.0	34.8	33.7	57.4	75.5	67.6
LoFi-E2E [62]	V	K400+**ANet**	TSM-Res18	8×224^2	14.6G	50.4	35.4	8.9	34.4	-	-	-	-
TSP [1]	V	K400+**ANet**	R(2+1)D-34	16×112^2	76.4G	51.3	37.1	9.3	35.8	35.0	59.0	76.6	69.0

b) 针对现有视频-文本多模态预训练方法缺乏细粒度交互对齐与时序推理能力的问题，开展 VL 预训练方法研究，以提升 VL 预训练模型的时序表征能力。研究表明，已有的 VL 预训练模型在需要跨模态细粒度时序信息对齐方面能力不足，面向诸多下游任务（如细粒度行为定位）时其性能与泛化性不佳。具体地，课题组提出了一种时序行为定位导向的视频-文本预训练框架（LocVTP），见图 2，通过挖掘“片段-词语”级别的细粒度对应关系来补充粗粒度的视频-句子对齐，并利用上下文变形机制构建时序感知对比损失，以增强模型对时序上下文依赖的建模能力。由表 2 可以看出，在 ActivityNet Captions 数据集上，LocVTP 取得了 48.2% 的 R@1 (IoU=0.5) 精度，显著优于 MIL-NCE (41.8%) 等主流 VL

预训练方法及任务型预训练基线 (44.4%)，验证了所提 LocVTP 预训练方法的有效性，特别是在需要精细化时序推理的下游任务中表现优异。

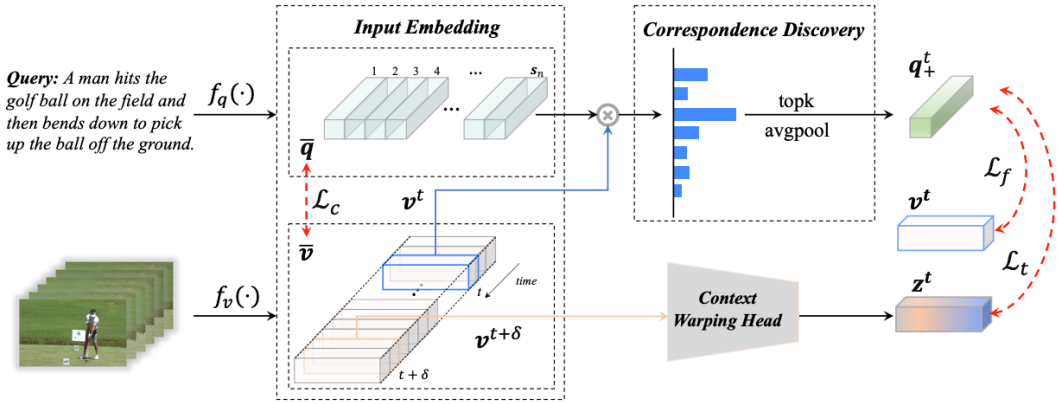

图 2　LocVTP 算法流程

表 2　LocVTP 在 ActivityNet Captions 数据集的时序定位任务实验结果

Models	PT Data	ANet Captions				Charades-STA				TACoS			
		$R_1^{0.5}$	$R_1^{0.7}$	$R_5^{0.5}$	$R_5^{0.7}$	$R_1^{0.5}$	$R_1^{0.7}$	$R_5^{0.5}$	$R_5^{0.7}$	$R_1^{0.3}$	$R_1^{0.5}$	$R_5^{0.3}$	$R_5^{0.5}$
Sep.Pre. [74]	Kinetics	44.4	27.1	77.6	62.1	39.7	23.8	79.6	52.3	37.2	25.6	58.2	45.5
LocVTP (Ours)	**HT‡**	**45.2**	**27.1**	**78.3**	**63.5**	**40.3**	**24.2**	**80.6**	**52.7**	**38.4**	**25.9**	**59.0**	**45.9**
VideoBERT* [49]	HT	37.2	21.0	66.7	53.6	32.7	19.5	68.1	46.2	33.8	22.2	51.6	41.0
MIL-NCE [38]	HT	41.8	24.5	73.5	57.7	37.0	21.2	74.3	50.4	35.1	23.5	53.7	42.5
UniVL [35]	HT	42.2	25.4	75.3	60.5	38.2	22.7	77.2	51.4	35.7	23.7	55.8	43.7
SupportSet* [41]	HT	41.9	25.2	74.7	58.3	37.4	21.6	75.6	50.9	35.5	23.5	54.2	43.2
LocVTP (Ours)	**HT**	**48.2**	**30.5**	**80.1**	**64.7**	**43.6**	**26.3**	**81.9**	**55.3**	**41.6**	**28.9**	**61.4**	**47.6**
Frozen [4]	CC,WV	43.3	25.8	75.8	59.3	38.8	22.9	77.6	50.3	35.7	23.5	54.4	43.7
OA-Trans* [54]	CC, WV	43.6	25.9	76.5	60.2	39.2	22.6	78.5	50.8	35.2	22.5	53.4	42.6
LocVTP (Ours)	**CC,WV**	**46.1**	**27.6**	**78.9**	**63.7**	**41.2**	**24.8**	**81.3**	**53.5**	**39.6**	**27.8**	**60.4**	**47.9**
December [52]	HT	43.0	25.1	76.0	60.2	37.2	21.6	78.3	50.6	34.8	22.9	55.1	43.9
ClipBERT [26]	CO,VG	42.6	24.6	75.3	59.7	37.0	20.8	77.7	50.2	33.7	21.0	54.3	43.3

2）基于多模态自适应对齐的高效动态视觉场景描述方法研究

本课题的核心研究内容是动态视觉场景描述（Visual Captioning）方法的研究。为此，课题组开展了多条技术路线的探索。

a) 针对现有视觉描述方法高度依赖大规模配对标注数据、且难以在多语言动态场景中高效迁移的瓶颈问题，课题组开展了基于域对齐与自编码的零样本多模态及多语言自然语言生成框架（ZeroNLG）的研究，构建了一个统一的零样本学习框架，如图 3 所示，旨在打破模态与语言之间的壁垒，该研究提出了一种创新的跨域对齐策略，将图像、视频等动态视觉数据与不同语言（英、中、德、法）

的文本数据映射到共享的公共潜在空间中，通过对齐各模态在该空间中的坐标来桥接视觉域与语言域。同时，结合无监督多语言自编码器与去噪语言重构机制，使模型在完全无需下游成对标注数据（如视频-文本对），即零样本条件下，实现高质量的跨模态描述生成与跨语言翻译。由表 3 可以看出，在对 ZeroNLG 进行定量分析与消融实验中，完整的模型架构（Full Setting）在视频描述（Video-to-Text）和图像描述（Image-to-Text）任务上，无论是在英语、中文还是德语、法语环境下，均取得了最优的性能指标（如 B-4, CIDEr 等）。对比实验显示，若移除跨域对齐模块（CDA, Setting b/c），模型在零样本设置下的性能会急剧下降甚至无法生成有效内容，证明了在该潜在空间进行自适应对齐的关键作用；而若移除去噪语言重构模块（DLR, Setting a），各项指标亦有明显回落，验证了该机制在提升生成文本鲁棒性方面的贡献。此外，多语言联合训练相比单语言训练带来了普遍的性能提升，说明了 ZeroNLG 算法在实际应用中的有效性，特别是在极低资源或无标注数据的复杂动态场景下，该方法能够通过高效的域对齐机制显著降低对人工标注的依赖，展现出优越的泛化能力和实用价值。

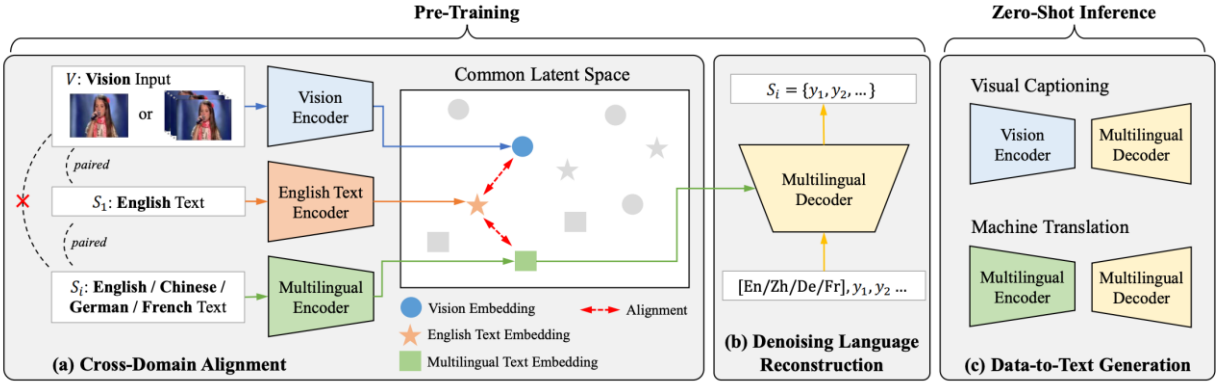

图 3　ZeroNLG 框图

表 3　ZeroNLG 算法在跨域 Video-to-Text 与 Image-to-Text 多语言生成上的实验结果

Setting	CDA	DLR					Video-to-Text								Image-to-Text											
		Data Corruption	Languages				English (En)				Chinese (Zh)				English (En)					Chinese (Zh)			German (De)			French (Fr)		
			En	Zh	De	Fr	B-4	M	R-L	C	B-4	R-L	C		B-4	M	R-L	C	S	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C
Full	√	√	√	√	√	√	**8.7**	**15.0**	**35.4**	**9.9**	**7.1**	**29.6**	**9.8**		**9.6**	**14.4**	**34.9**	**29.9**	**8.7**	**8.4**	**31.8**	**18.0**	**5.7**	**27.2**	**17.1**	**2.8**	**18.6**	**24.8**
(a)	√	-	√	√	√	√	1.5	10.2	24.8	3.7	7.9	18.6	2.9		1.0	7.7	19.0	5.7	3.0	0.7	17.7	3.5	0.0	15.2	3.5	0.5	10.1	8.0
(b)	-	√	√	√	√	√	0.6	7.5	22.7	0.5	0.0	16.8	0.5		0.7	5.3	15.8	1.0	1.0	0.0	14.9	0.7	0.0	11.4	0.7	0.0	7.2	2.8
(c)	-	-	√	√	√	√												**Fail**										
(d)	√	√	√				7.3	14.6	34.0	9.0	-	-	-		9.3	14.2	34.3	27.5	8.3	-	-	-	-	-	-	-	-	-
(e)	√	√		√			-	-	-	-	6.4	28.1	8.6		-	-	-	-	-	8.0	30.4	17.5	-	-	-	-	-	-
(f)	√	√			√		-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	3.6	25.3	13.8	-	-	-
(g)	√	√				√	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	2.3	17.8	21.2

b)开展基于视觉关系感知的多模态自适应对齐动态视觉场景描述方法。该方法以视频中蕴含的主体—关系—客体（subject–relation–object）视觉关系结构为核心语义纽带，实现视觉模态与语言模态在语义层面的精准对齐；同时，如图 4 所示，通过引入视觉关系感知模块（Visual Relation-Aware Module, VRAM），在模型训练过程中对与视觉关系高度一致的词语进行自适应加权，有效抑制跨模态对齐中的语义噪声，从而在无需人工标注的条件下实现高效、稳健的动态视觉场景描述生成。如表 4 所示，该方法在 MSVD、MSR-VTT 等主流视频描述基准数据集上，相较于基于视觉概念对齐的无监督方法，在 CIDEr、METEOR 等语义一致性评价指标上均取得了显著性能提升，验证了所提出多模态自适应对齐机制在复杂动态视觉场景描述任务中的有效性与先进性。

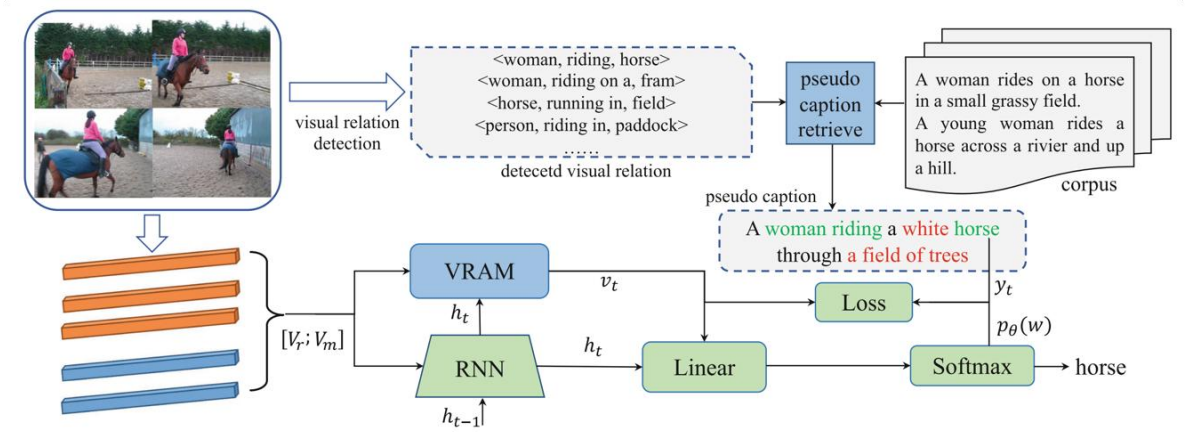

图 4 VRAM 算法流程图

表 4 算法在 CIDEr 数据集上的性能

Corpus	Method	BLEU@4	METEOR	ROUGE_L	CIDEr
TVC	full model	**5.3 ± 0.2**	**16.2 ± 1.1**	**40.6 ± 1.7**	**8.8 ± 0.9**
	w/o VRAM	3.6 ± 1.2	14.3 ± 1.1	39.9 ± 2.0	4.2 ± 0.8
	w/o Relation	4.2 ± 0.7	16.1 ± 0.6	40.0 ± 1.9	3.3 ± 1.8
GCC	full model	4.8 ± 0.8	14.2 ± 0.5	38.4 ± 1.8	**12.9 ± 2.0**
	w/o VRAM	1.4 ± 0.5	11.8 ± 1.1	32.5 ± 1.8	6.5 ± 2.4
	w/o relation	**5.6 ± 1.4**	**15.9 ± 0.7**	**40.7 ± 4.3**	2.8 ± 1.2
VATEX	full model	**9.6 ± 1.3**	**20.1 ± 0.9**	**42.1 ± 3.9**	**13.2 ± 0.8**
	w/o VRAM	6.7 ± 1.1	18.6 ± 0.7	39.2 ± 1.9	8.3 ± 0.9
	w/o relation	5.8 ± 1.7	16.5 ± 1.1	37.1 ± 4.9	1.3 ± 0.3

c) 开展基于概念感知联合表征学习的动态视觉场景描述方法。该方法以细粒度概念语义作为跨模态语义对齐的桥梁，如图 5 所示，针对现有视觉描述模型在视觉表征泛化能力不足、概念检测精度受限以及概念语义利用不充分等问题，通过引入大规模视觉—语言对比预训练模型进行视觉编码，提升复杂动态场景中多概念的可分性；同时，如表 5 所示，结合多模态融合的概念检测机制，利用视觉与语言信息的协同建模提高概念检测的准确性，并通过全局—局部语义引导机制促进概念语义在描述生成过程中的协同利用，从而有效改善视觉与语言之间的语义对齐，提升动态视觉场景描述的生成质量。

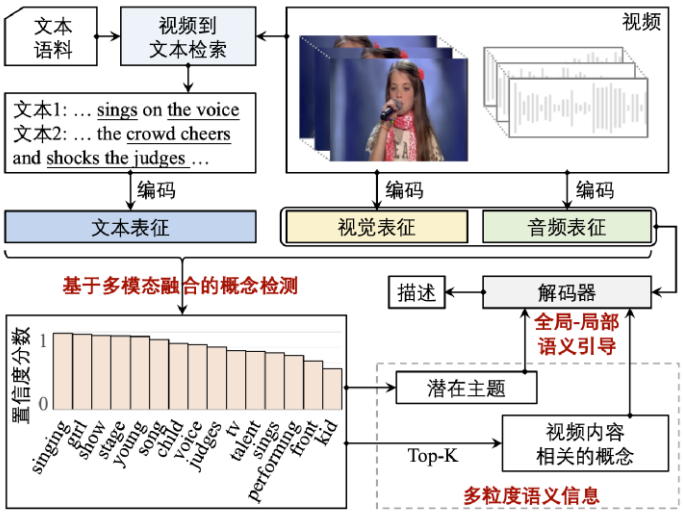

图 5　概念感知视觉描述框图

表 5　VATEX 数据集上与先进视觉描述方法对比的实验结果

方法	模态特征			检测	VATEX性能			
	图像	运动	音频	类型	CIDEr	BLEU4	METEOR	ROUGEL
VATEX[4]	–	I3D	–	–	45.1	28.4	21.7	47.0
OpenBook[32]	IRv2	3D-RX	–	–	54.6	32.5	23.1	49.2
MGCMP[205]	IRv2	3D-RX	–	–	57.6	34.2	23.5	50.3
ORG-TRL[35]	IRv2	3D-RX	–	物体	49.7	32.1	22.2	48.9
HRNAT[206]	IRv2	I3D	–	物体	50.7	32.5	22.3	49.0
CAVC（本文）	IRv2	3D-RX	–	概念	57.9	35.0	23.9	50.7
CAVC（本文）	CLIP B32	3D-RX	✓	概念	**62.6**	**37.6**	**25.0**	**52.4**
CoCa[108]	CoCa L14	–	–	–	57.0	31.7	23.7	49.5
+ CAVC（本文）	CoCa L14	–	–	概念	**58.6**	**32.5**	**23.9**	**49.8**
LLaVA-1.5[171]	CLIPL14	–	–	–	67.1	39.0	25.3	52.7
+ CAVC（本文）	CLIPL14	–	–	概念	**70.3**	**40.5**	25.9	**53.6**

d) 视觉描述涉及视觉模态和语言模态，在跨语言视觉描述场景下，VC 模型的训练面临诸多挑战。因为新的目标语言-视觉配对数据匮乏是一个严重的限制。为此，课题组开展了非配对视频描述生成（Unpaired Video Captioning）方法研究，不同于传统的“视频-枢轴语言-目标语言”的流水线系统（Pipeline System）跨语言视觉描述路线，课题组创新地提出了一种基于视觉注入的非配对视频描述系统（UVC-VI），见图 6 所示。该方法通过设计视觉注入模块（VIM）与多模态协同编码器（MCE），在无需枢轴语言生成的情况下，实现了源端视觉域与目标端语言域的直接动态对齐与信息注入。由表 6 可以看出，UVC-VI 方法在 CIDEr、BLEU 等关键视觉描述指标上显著优于基准流水线系统，甚至超越了部分有监督视频描述模型；同时，引入 MCE 模块后，模型在 MSVD 和 MSR-VTT 数据集上的性能分别获得了显著提升，说明了 UVC-VI 算法在实际应用中的有效性。

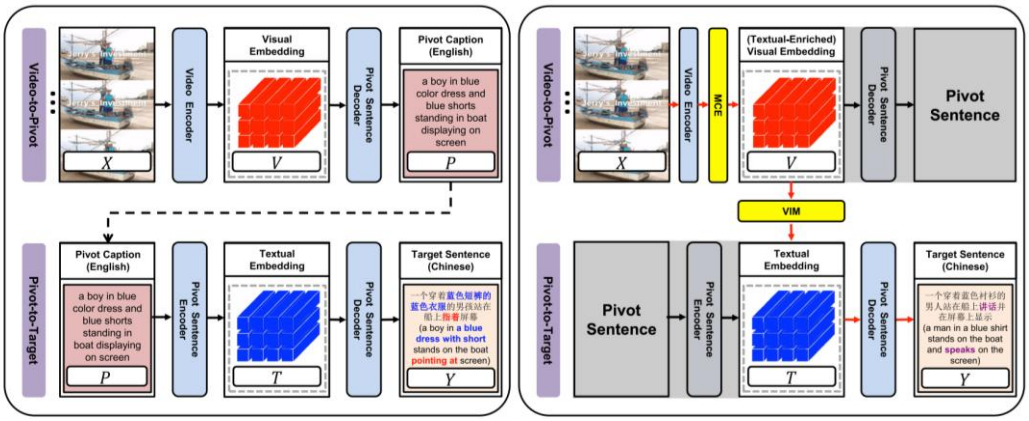

图 6　UVC-VI 框图

表 6 UVC-VI 在 MSVD 等视频描述数据集上的实验结果

Types	Methods	Features	Dataset: MSVD [17]				Dataset: MSR-VTT [8]			
			BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Setting: Conventional Supervised Video Captioning										
Supervised Systems (Section 3.1)	AF(ICCV2017) [62]	I+M+A	-	-	-	-	39.7	25.5	-	40.0
	MA-LSTM(ACMMM2017) [19]	I+M+A	-	-	-	-	36.5	26.5	59.8	41.0
	Two-stream(TPAMI2020) [63]	I+M	**54.3**	33.5	-	72.8	39.7	27.0	-	42.1
	PickNet(ECCV2018) [64]	I	52.3	33.3	69.6	76.5	39.4	27.3	59.7	42.3
	RecNet(TPAMI2020) [65]	I	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
	TDConvED(AAAI2019) [21]	I	53.3	33.8	-	76.4	39.5	27.5	-	42.8
	POS-CG(ICCV2019) [66]	I+M	-	-	-	-	38.3	26.8	60.1	43.4
	STAT(TMM2020) [67]	I+M	52.0	33.3	-	73.8	39.3	27.1	-	43.8
	GRU-EVE(CVPR2019) [22]	I+M	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
	SAAT(CVPR2020) [5]	I+M	46.5	33.5	69.4	81.0	40.5	28.2	**60.9**	49.1
	SGN(AAAI2021) [7]	I+M	52.8	**35.5**	**72.9**	**94.3**	40.8	28.3	60.8	49.5
	MGSA(AAAI2019) [24]	I+M+A	-	-	-	-	**45.4**	**28.6**	-	**50.1**
Setting: Unpaired Video Captioning										
Pipeline Systems (Section 3.2)	MA-LSTM [19] + Google Translator [68]‡	I+M	43.1	29.5	65.8	53.3	31.5	23.9	54.2	30.6
	SAAT [5] + Google Translator [68]‡	I+M	46.4	30.2	66.4	61.1	35.2	25.8	57.2	37.8
	SGN [7] + Google Translator [68]‡	I+M	50.7	32.6	69.2	72.9	38.0	26.7	57.1	39.6
	Base Model	I+M(+A)	47.2	31.8	67.6	68.9	34.7	25.1	55.8	36.3
Proposed (Section 3.3)	UVC-VI	I+M	49.6	34.7	70.3	83.4	37.2	26.8	57.7	43.7
		I+M+A	-	-	-	-	38.9	27.8	59.5	44.5

e) 开展基于共识引导与关键词自适应加权的多模态对齐训练方法（Consensus-Guided Keyword Targeting, CGKT）。该方法针对现有视频描述模型在全监督训练中对大规模高质量标注数据依赖强、人工标注存在长度差异与词分布不均等问题。如图 7 所示，通过引入基于描述共识度的样本重加权机制的损失函数 CGKT，降低低质量标注对模型训练的不利影响，并结合基于词频的关键词加权策略，引导模型关注更具信息量且低频的重要语义词汇，从而促进模型学习更高质量、更具判别性的动态视觉场景语言描述。由表 7 和

表 8 可见，该方法在 MSVD、MSR-VTT 等主流动态视觉场景描述基准数据集上，在 CIDEr、BLEU@4、METEOR、ROUGE-L 等评价指标上均显著优于采用传统交叉熵训练目标的基线模型，并在不改变模型结构的前提下实现了稳定性能提升，验证了所提出多模态自适应对齐策略在提升动态视觉场景描述质量与效率方面的有效性与通用性。

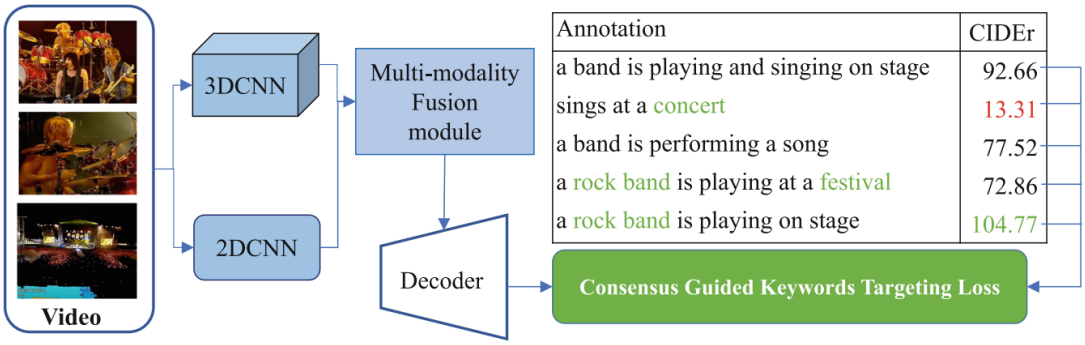

图 7　CGKT 描述模型框图

表 7　在 MSVD 基准数据集上的实验结果

Method	BLEU@4	METEOR	ROUGE_L	CIDEr	ΔC
S2VT *w*/CE	45.8	35.6	70.1	79.2	
S2VT *w*/CGKT	**49.6**	**36.6**	**71.1**	**85.7**	+8.2%
AttLSTM *w*/CE	49.5	36.5	**71.1**	88.2	
AttLSTM *w*/CGKT	**50.6**	**36.7**	**71.1**	**89.4**	+1.3%
Semantic *w*/CE	51.0	37.1	71.2	89.4	
Semantic *w*/CGKT	**51.1**	**37.4**	**71.7**	**93.0**	+4.0%

表 8　在 MSR-VTT 基准数据集上的实验结果

Method	BLEU@4	METEOR	ROUGE_L	CIDEr	ΔC
S2VT *w*/CE	40.8	27.6	59.7	49.8	
S2VT *w*/FL	40.7	27.8	59.9	50.6	+1.6%
S2VT *w*/LS	**41.6**	**27.9**	**60.4**	50.8	+2.0%
S2VT *w*/CGKT	41.4	27.7	60.2	**52.0**	+4.4%

f) 开展基于自适应课程学习的动态视觉场景描述方法。如图 8 所示，视频描述数据中存在样本噪声大、难度分布不均以及随机采样训练易引入数据偏置的问题，该方法通过引入自适应课程学习策略 Adaptive Curriculum Learning (ACL)，对视频—文本样本进行难度排序，并在不同训练阶段自适应地选择与模型当前学习能力相匹配的训练子集，从而引导模型由易到难地逐步学习，如表 9 所示，在无需修改模型结构的前提下方法有效提升动态视觉场景描述模型的训练稳定性与整体性能。

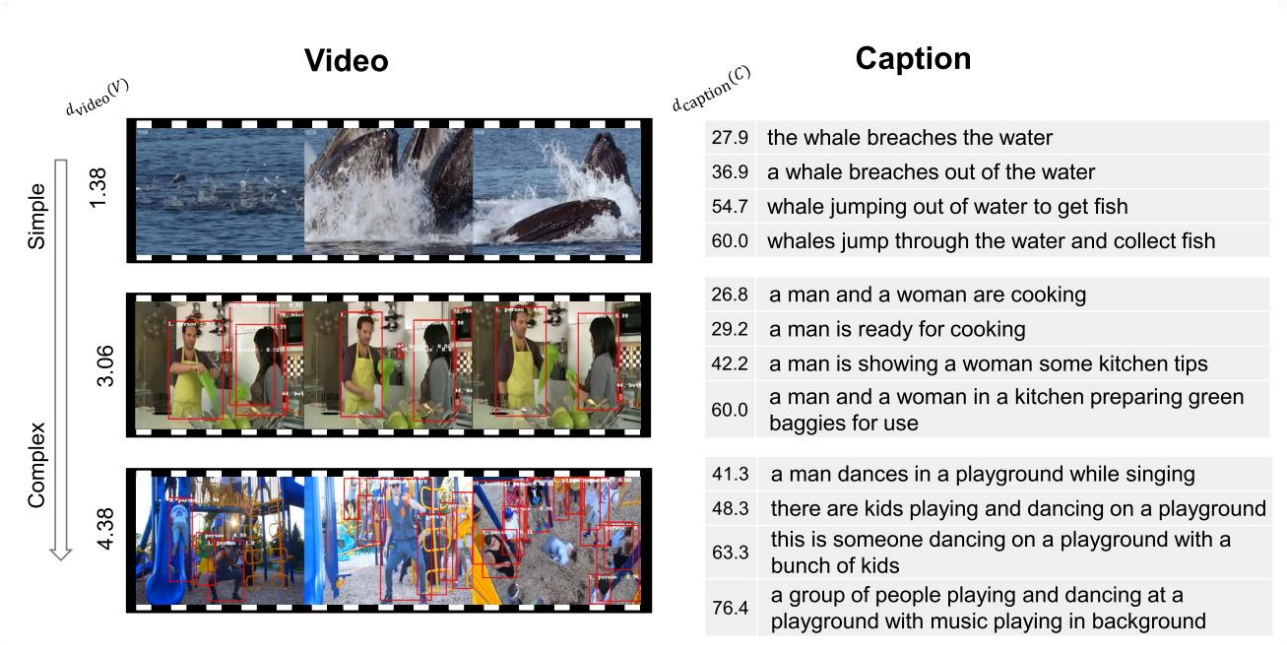

图 8　MSR-VTT 数据集的难度分布图

表 9　在 MSVD 和 MSR-VTT 数据集上对三个基准模型的实验结果

	MSVD				MSR-VTT			
Model	BLEU@4	METEOR	ROUGE-L	CIDEr	BLEU@4	METEOR	ROUGE-L	CIDEr
TopDown [29]	51.8	34.1	71.8	87.6	39.9	27.1	60.5	45.4
w/ PCL	51.9	34.5	71.7	88.1	41.0	**27.3**	60.7	46.0
w/ ACL	**53.6**	**35.2**	**72.9**	**89.6**	**41.5**	27.1	**60.9**	**46.9**
ParAhLSTMat [34]	48.9	34.4	71.3	89.8	40.6	27.7	60.4	47.1
w/ PCL	50.0	34.4	71.9	**90.3**	40.8	27.8	**61.0**	48.1
w/ ACL	**51.0**	**34.7**	**72.2**	89.9	**41.3**	**28.0**	**61.0**	**48.3**
ARB [35]	47.5	34.8	71.1	88.8	41.2	28.6	60.4	48.8
w/ PCL	48.2	34.9	71.3	90.2	41.4	**28.9**	60.9	49.8
w/ ACL	**48.3**	**35.4**	**72.0**	**90.5**	**42.6**	**28.9**	**61.5**	**51.3**

3）基于交互式图网络的多模态高阶语义关系推理方法研究

课题组开展了基于交互式图网络的多模态高阶语义关系推理方法研究，针对跨模态运动表征学习与高阶语义对齐、关系推理开展研究。

a) 开展基于深度运动先验的弱监督时序动作定位方法研究（框图如图 9 **错误!未找到引用源。错误!未找到引用源。**所示）。具体地，在仅有视频级别运动类别信息标注的条件下，设计深度网络，显示学习深度运动先验，即利用图网络对视频上下文依赖实现时序运动显示建模。构建了一种基于图卷积的“运动图”（Motion Graph）以生成高阶“运动度”（Motionness）先验，并设计了与其适配的“运动引导损失函数”（Motion-guided Loss），其中利用光学流特征构建局部运动载体，通过在运动图中引入位置边与语义边，显式推理视频片段间的长时序动态关联。在此基础上，提出利用生成的深度运动先验来动态调节网络训练权重的策略，以解决传统弱监督多示例学习中交叉熵损失函数对动作边界定位不准的问题。由表 10 可以看出，在 THUMOS’14 数据集上进行消融实验对比时，相比于传统的交叉熵损失，采用课题组提出的光流引导的运动损失函数取得了显著的性能提升。具体数据表明，在 IoU 阈值为 0.5 时，模型的平均精度（mAP）由 23.8% 大幅提升至 34.2%，且预测分布与真实标签的 KL 散度显著降低（从 0.087 降至 0.004），说明了 DMP-Net 算法及其核心的运动引导损失机制在抑制背景噪声、精准定位复杂动作边界方面的实际应用有效性。

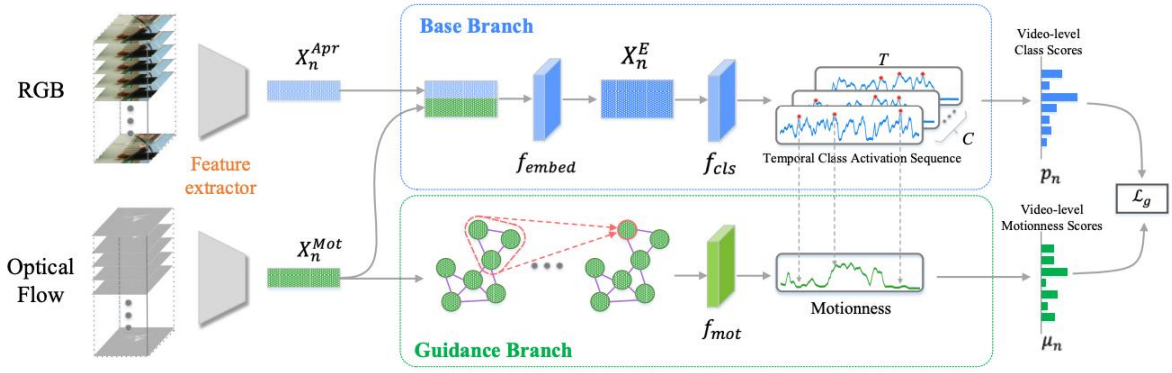

图 9　DMP-Net 框图

表 10　DMP-Net 在 THUMOS'14 数据集不同损失函数的消融实验结果

Loss	mAP@IoU (%)				KL
	0.4	0.5	0.6	0.7	
$\mathcal{L}_{g,Flow}$	**44.0**	**34.2**	**23.5**	**13.1**	**0.004**
$\mathcal{L}_{g,RGB}$	31.9	22.2	13.5	4.6	0.087
$\mathcal{L}_{g,R\&F}$	40.5	28.3	20.8	11.0	0.023
$\mathcal{L}_a$	33.1	23.8	15.2	7.9	–
$\mathcal{L}_{g,Flow'}$	40.7	30.4	19.7	11.4	0.016

b) 开展基于交互式图网络的多模态高阶语义关系推理方法研究。显然，动态时序信息的复杂演化与特征捕获一直是表示学习的一个挑战，课题组提出了基于帧级时序建模（Frame-level Timeline Modeling, FTM）的通用时间图表示学习方法（框图见图 10 所示），该方法有效解决了现有时间图邻域聚合策略难以同时兼顾短时微观特征与长时演化规律的瓶颈。课题组创新性地提出了基于链路的成帧技术（Link-based framing technique）以保留短时特征，并引入时序聚合器（Timeline aggregator）模块来建模图演化的内在动力学，构建了"帧聚合器"与"时序聚合器"两阶段处理架构。由表 11 可以看出，在不同邻域规模（从 S 到 XL）以及极低训练数据比例（如 1%）的严苛测试条件下，集成 FTM 策略的模型在 Reddit 等数据集的未来链路预测任务中，其平均精度（AP）均显著优于原始基干模型（GraphSAGE, GAT, TGAT）。特别是在归纳式（Inductive）设置下，FTM 带来了大幅度的性能增益（例如在 S 邻域规模下平均提升达 14.29%），说

明了 FTM 算法在提升时序图神经网络模型的表征能力、鲁棒性、域泛化能力以及在小样本数据场景下实际应用中的有效性。

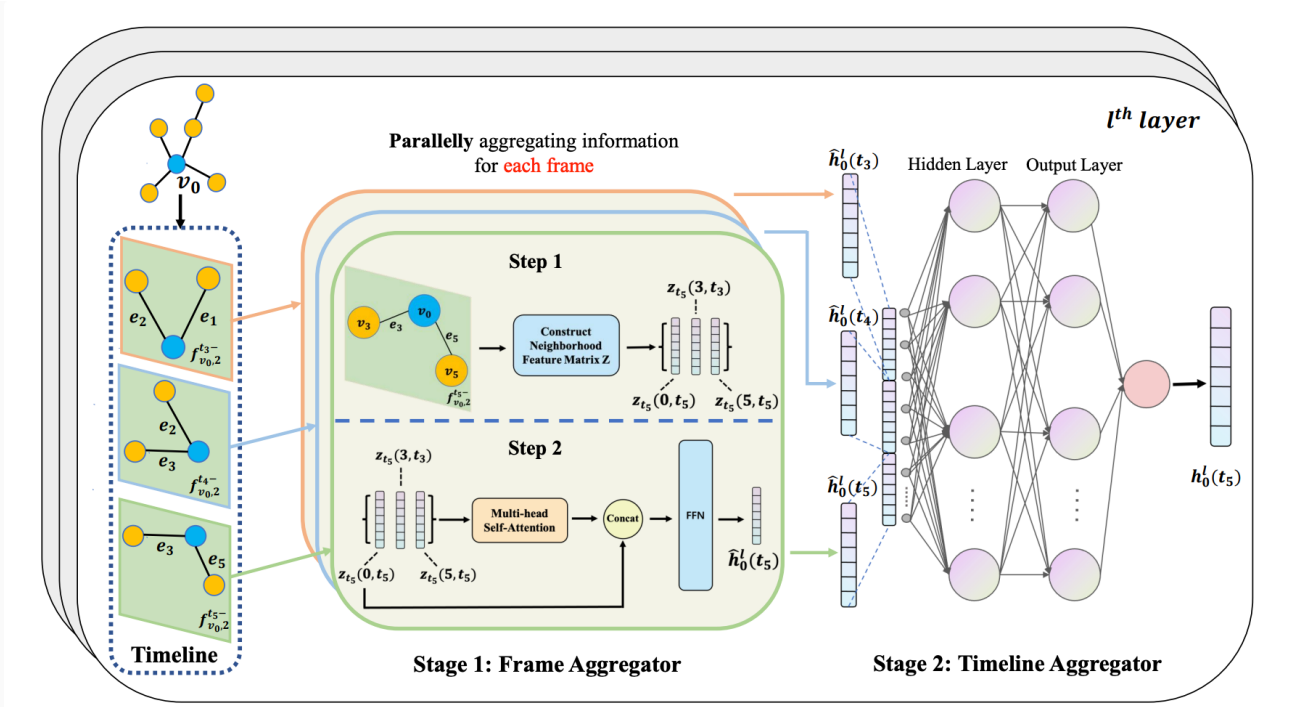

图 10　FTM 框图

表 11　不同邻域规模下 FTM 在 Reddit 等数据集的实验结果

Model	Neighborhood Scale								Percentage of Training Data							
	Inductive				Generalization				Inductive				Generalization			
	S	M	L	XL	S	M	L	XL	1%	5%	10%	50%	1%	5%	10%	50%
GraphSAGE	86.31	88.96	94.19	94.68	70.87	70.83	78.74	83.59	65.31	85.39	91.17	95.64	57.99	62.79	73.04	80.15
w/ FTM	92.24↑	92.31↑	95.53↑	96.28↑	79.37↑	77.26↑	86.30↑	86.53↑	70.40↑	87.58↑	91.95↑	96.65↑	61.98↑	74.92↑	82.71↑	85.34↑
GAT	91.11	93.15	95.56	95.37	69.88	74.96	83.76	85.84	68.99	90.81	93.13	95.10	59.53	76.44	79.70	85.80
w/ FTM	91.85↑	93.40↑	95.84↑	96.75↑	82.38↑	81.75↑	86.20↑	88.97↑	73.13↑	91.02↑	93.70↑	96.68↑	68.45↑	81.99↑	85.91↑	90.30↑
TGAT	91.12	92.63	95.95	96.73	69.22	71.76	85.64	87.34	65.65	88.92	92.67	96.25	74.51	77.16	81.27	86.38
w/ FTM	94.08↑	94.32↑	97.26↑	96.82↑	91.08↑	89.52↑	95.82↑	91.06↑	80.76↑	92.32↑	93.45↑	96.25	81.84↑	87.88↑	87.22↑	88.53↑
Average Gain	3.21	1.76	0.98	1.02	14.29	10.33	6.73	3.26	8.10	1.67	0.64	0.69	5.00	8.25	5.69	2.87

Table 5: Case studies on (1) neighborhood scale, where neighborhood scale expands from S to XL; and (2) the percentage of training data, where models are trained on limited training data of Reddit, *e.g.*, 1% means models are trained/validated on one-percent of the original training/validation data. We do not take TGN into consideration, because the way TGN updates node-wise memory has little to do with the neighborhood scale and the percentage of training data. We report AP(%) of future link prediction on Reddit (inductive; generalize from Wiki).

4）交通态势智能播报生成系统

为了验证本课题所提方法的有效性，面向智慧交通场景的实际应用需求，以视频交通场景交通态势智能播报生成系统（类似路况播报）为应用场景，系统开展了交通场景视频数据资源建设、交通动态场景视觉描述生成模型的设计、训练与系统实现。在数据层面，构建了高质量交通场景中文视频描述数据集 TV-CL，该

数据集覆盖多种典型交通场景、交通参与主体及交通行为类型，能够较为全面地反映真实交通环境中动态视觉场景的时序演化特征与语义复杂性。在模型层面，基于所构建的数据集，完成了视频交通场景描述生成模型的设计与训练，将视频时序运动建模、多模态语义关系推理与语言生成过程进行统一建模与协同优化。基于自建的交通场景中文视频描述数据集 TV-CL 对视频交通场景描述模型进行了优化训练，并结合关键帧提取技术对模型推理流程进行了优化，在保证描述质量的同时有效降低了推理计算开销，提升了模型在实际应用场景中的推理效率。通过在真实交通场景视频上的实验验证，ADSPALB-DRSA 与 ADSPALB-DRSA-Large 两种模型配置在复杂交通场景下均表现出优秀的动态视觉场景解析与时序语义理解能力。如图 11 所示，模型能够对交通场景中的场景解析与时序分析；由图 12 可见，在带有场景解析与时序分析的基础上，模型能稳定生成符合交通语义逻辑的动态场景描述，可见所提出的视频交通场景描述模型在复杂动态交通环境中具备较好的适应性与鲁棒性，验证了本项目所提方法在实际交通场景中的有效性、泛化性与应用潜力。

图 11　ADSPALB-DRSA 交通态势智能播报生成系统界面（场景分析）

图 12　ADSPALB-DRSA 交通态势智能播报生成系统界面（语音播报）

三、其他相关信息

1) 国内外学术合作交流等情况

在 NSFC 项目的支持下，团队积极参与国际学术交流，多位同学参与了、CVPR2022、ICCV2023、ECCV2022 和 ACL2024 等国际学术会议交流，分别进行了 Poster 和 Oral 展示，学术成果获得了国内外同行的正面评价和积极引用，本项成果获得累积引用 600 多次。此外，团队积极开展国际学术合作，

1) 课题组与新加坡南洋理工大学王州霞博士、新加坡国立大学寿政博士、纽约哥伦比亚大学陈隆博士、牛津大学教授 David A. Clifton、腾讯 AI LAB 吴贤博士团队和鹏城实验室副研究员张彤团队进行学术交流与合作，共同发表了 TPAMI、TIP 等领域内顶级期刊和 ECCV、CVPR、ACL 等领域内顶级会议论文。

2) 2022 年 3 月，引进博士后研究人员 RAZA ASIF（国籍巴基斯坦，护照号：HC9891022）与项目组成员杨邦合作开展“视觉语言跨模态学习”方法研究，并联合发表 JCR 一区期刊一篇。

3) 2022 年 10 月，邹月娴教授携项目组成员杨邦、吉普照同学参加在深圳召开的中国模式识别与计算机视觉大会（PRCV 2022）,就模式识别、计算机视觉与机器学习领域前沿理论与方法进行成果展示和学术交流。

4) 2023 月 6 月 2 日-6 月 11 日，邹月娴教授赴希腊参加 2023 届 IEEE 国际声源、语音和信号处理博览会（ICASSP 2023，领域旗舰会议），就机器学习、音视频理解与生成等任务进行成果展示和学术交流。

5) 2024 年 8 月 10 日-8 月 15 日，邹月娴教授携项目组成员庄先炜同学赴泰国曼谷参加第 62 届计算语言学协会年会（ACL 2024，领域旗舰会议），就跨模态预训练模型、多模态场景理解等工作进行成果展示和学术交流。

6) 2025 年 8 月 16 日-8 月 22 日，邹月娴教授赴荷兰鹿特丹参加 2025 届国际语言交流大会（Interspeech 2025，领域旗舰会议），就多模态技术在视觉等领域推动无障碍人机交互的应用、改善视觉问答系统对非标准语音输入的鲁棒性等任务进行成果展示和学术交流。

2) 项目成果转化及应用情况。

1) 本课题部分研究成果获得了上市公司深圳市奥拓电子股份有限公司的关注和认可，为此与课题组所在实验室签订了委托研发项目“人工智能场景认知”（2023.11-2025.01，160 万）。

2) 本课题跨模态理解与生成研究成果获得腾讯团队关注和认可，为此与课题组所在实验室签订了委托研发项目“细粒度跨模态对齐与跨模态内容生成”（2025.11-2026.12，经费 30 万）。

3) 本课题跨模态理解与生成研究成果获得重庆中医院人工智能团队关注和认可，为此与课题组所在实验室签订了委托研发项目“基于中西医多模态信息的皮肤病诊疗专业大模型研究”（2025.11-2028.11，经费 50 万）。

4) 本课题跨模态理解与生成研究成果获得重庆中医院睡眠中心团队关注和认可，为此与课题组所在实验室联合开展“基于中西医知识的睡眠障碍智能诊断大模型研究”（2025.09-2028.09，经费 40 万）。

本课题研究成果正在持续获得产业界关注。

3) 人才培养情况。

在自然科学基金项目的支持下，已经培养了 4 名博士研究生和 9 名硕士研究生，均已获得北京大学理学博士/硕士学位。

序号	姓名	研究生类别	专业/研究方向	毕业论文题目	导师姓名	答辩时间	毕业去向
1	杨邦	博士研究生	计算机应用技术/模式识别技术	基于表征学习的数据高效视觉描述方法研究	邹月娴	2025 年 6 月	深圳，鹏城实验室
2	曹蒙	博士研究生	计算机应用技术/模式识别技术	基于跨模态表征学习的时序行为定位方法研究	邹月娴	2023 年 5 月	深圳，大湾区实验室
3	张粲	博士研究生	计算机应用技术/计算机视觉	面向行为检测的时空可区分性表征学习研究	邹月娴	2022 年 5 月	北京，字节跳动
4	杨东明	博士研究生	计算机应用技术/计算机视觉	人-物交互建模与表征学习方法研究	邹月娴	2022 年 5 月	北京，中国电信
5	刘峰林	硕士研究生	计算机应用技术/多媒体信息处理技术	知识驱动的低资源视觉描述方法研究	邹月娴	2022 年 5 月	英国，牛津大学（读博）
6	李善浩	硕士研究生	计算机应用技术/多媒体信息处理技术	面向视频描述的数据增强和采样策略方法研究	邹月娴	2022 年 5 月	北京，亚马逊
7	吉普照	硕士研究生	计算机应用技术/多媒体信息处理技术	基于数据与运动信息增强的交通视频描述深度模型研究	邹月娴	2023 年 6 月	深圳，快手
8	吴立渝	硕士研究生	计算机应用技术/多媒体信息处理技术	鲁棒行为识别的运动增强和时空表征方法研究	邹月娴	2022 年 5 月	深圳，大疆
9	蒋吉	硕士研究生	计算机应用技术/模式识别与机器学习	基于特征对齐的视频定位深度学习方法研究	邹月娴	2024 年 6 月	北京，腾讯
10	李鸿翔	硕士研究生	计算机应用技术/多媒体信息处理技术	基于语义关系挖掘的视频-语言时序定位研究	邹月娴	2025 年 5 月	香港，香港科技大学（读博）
11	姚子裕	硕士研究生	计算机应用技术/多媒体信息处理技术	基于自回归模型的可控视觉生成方法研究	邹月娴	2025 年 5 月	深圳，自变量机器人
12	庄先炜	硕士研究生	计算机应用技术/多媒体信息处理技术	视觉语言大模型幻觉成因分析与缓解方法研究	邹月娴	2026 年 5 月	/

13	唐乐翔	硕士研究生	计算机应用技术/多媒体信息处理技术	/	邹月娴	2027 年 5 月	/

本项目培养的北京大学博士研究生张粲博士目前在字节跳动担任算法工程师，负责 AI 大模型技术；杨东明博士目前担任中国电信担任天翼云公司产线总经理助理、负责 AI、息壤、智算等工作；杨邦博士加入鹏城国家实验室，担任助理研究员，负责三维沉浸式音视频技术研究工作；曹蒙博士目前在穆罕默德·本·扎耶德人工智能大学担任博士后研究员，开展多模态大语言模型的智能体化与物理空间推理能力研究；本项目培养的硕士研究生李善浩、吴立渝、吉普照、蒋吉、姚子裕，庄先炜分别加入了快手、大疆、腾讯、亚马逊、千问等中国人工智能领军/龙头企业，参与 AI 多模态技术研究；刘峰林和李鸿翔分别在牛津大学、香港科技大学攻读博士学位，开展数字精准预防和长寿医疗健康以及面向生成和理解统一的多模态视觉语言模型的研究。可见，本项目不仅仅在推动学术研究与技术发展方面取得了优秀的预期成果，而且在人才培养上也起到了重要的作用，为相关领域输送了一批具备扎实理论基础和实践能力的专业人才。

4) 其他需要说明的成果。

1) 2024 年 5 月，项目组成员杨邦领衔参加知名国际医疗竞赛“ImageCLEFMedical Caption Prediction 2024”，并荣获冠军（1/11）；
2) 2022 年 8 月，项目组成员曹蒙领衔参加 2022 人工智能大会（WAIC） 赛，黑客松蚂蚁财富赛道:“行情波动下的金融问答挑战”,并荣获冠军（1/4278）。