Progress Report

- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár and Christoph Feichtenhofer. **SAM 2: Segment Anything in Images, Videos**. arXiv:2408.00714, Aug 2024.
- Wei Feng, Xin Wang, Hong Chen, Zeyang Zhang, Wenwu Zhu. **Multi-sentence Video Grounding for Long Video Generation.** arXiv:2407.13219, Jul 2024.
- Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz Khan. **Video-GroundingDINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding.** arXiv:2401.00901, Dec 2023

Introduction

- The paper introduces an **Open-Vocabulary Spatio-Temporal Video Grounding** task, overcoming the limitations of closed-set video grounding.
  - Introduced a novel **spatio-temporal video grounding model** that achieves state-of-the-art results in closed-set benchmarks and outperforms others in open-vocabulary settings.
  - The model leverages **pre-trained spatial grounding mod**els and integrates temporal aggregation modules for spatio-temporal localization.
- **Goal:** Achieve improved open-vocabulary performance while maintaining strong closed-set video-grounding performance.

Related Work

- **Spatial Grounding Models**
  - Foundational models:
    - **GLIP**, **Grounding DINO**, **Kosmos-1**, **Kosmos-2, Ferret, GLaMM**
  - Excel in open-vocabulary tasks
  - Limited to static images.
- **Spatio-Temporal Video Grounding**
  - Existing models:
    - **STGVBert, TubeDETR, Augmented 2D-TAN, OMRN, MMN, STCAT, STVGFormer**
  - Excel in closed-set tasks.
  - Struggled with open-vocabulary generalization.

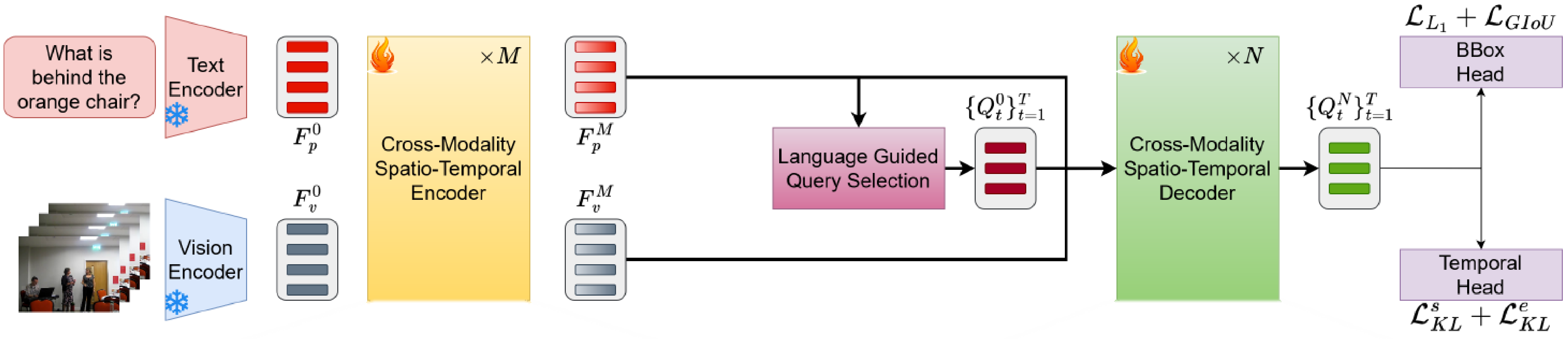
Problems Definition

- **Spatial grounding problem**
  - The localization of one or more objects associated with the text prompt in a frame using a bounding box.
- **Temporal grounding problem**
  - Understanding how objects or actions evolve over time.
- **Spatial-temporal grounding problem**
  - A set of spatio-temporal coordinates associated with the subset of frames where objects exist.

Spatial-temporal Video Grounding

- Problem: limited dataset
- Solution
  - Utilize the generalized representations of these models to enrich the weaker representation of video-grounding approaches.
  - Aim to leverage the strong pretrained representations of spatial grounding methods.

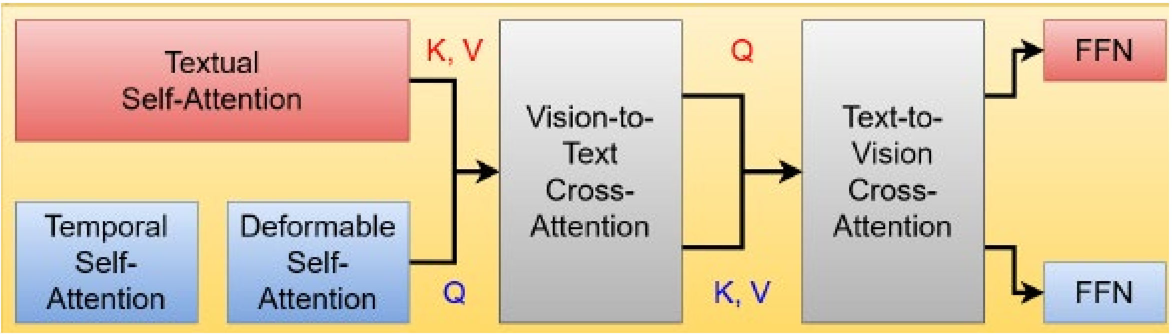
Architecture



Cross-Modality Spatio-Temporal Encoder

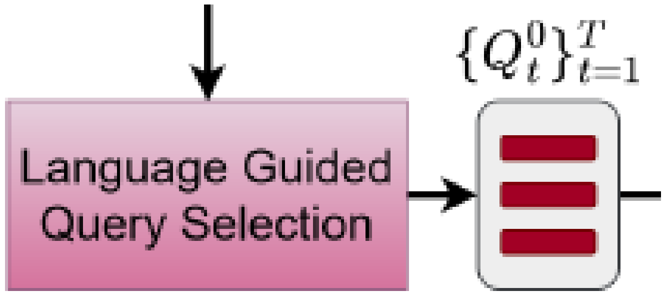
- Multi-Head Self-Attention
  - To visual features along the temporal dimensions
  - Also applied on text features
- Deformable Attention
  - To visual features along the spatio dimensions

$$F_v^{m\prime} = \mathrm{DA}^m_{spatial}(\mathrm{MHSA}^m_{temporal}(F_v^{m-1})),$$
$$F_p^{m\prime} = \mathrm{MHSA}^m_p(F_p^{m-1}),$$
$$\mathrm{Attn}^{\mathrm{m}}_{\mathrm{joint}} = \left(\frac{proj^m_{q,v}(F_v^{m\prime})proj^m_{q,p}(F_p^{m\prime})^T}{\sqrt{d^k}}\right)$$
$$F_v^m = \mathrm{FFN}^m_v(\mathrm{softmax}(\mathrm{Attn}^{\mathrm{m}}_{\mathrm{joint}})proj^m_p(F_p^{m\prime}))),$$
$$F_p^m = \mathrm{FFN}^m_p(\mathrm{softmax}({\mathrm{Attn}^{\mathrm{m}}_{\mathrm{joint}}}^T)proj^m_v(F_v^{m\prime}))),$$



Language-Guided Query Selection

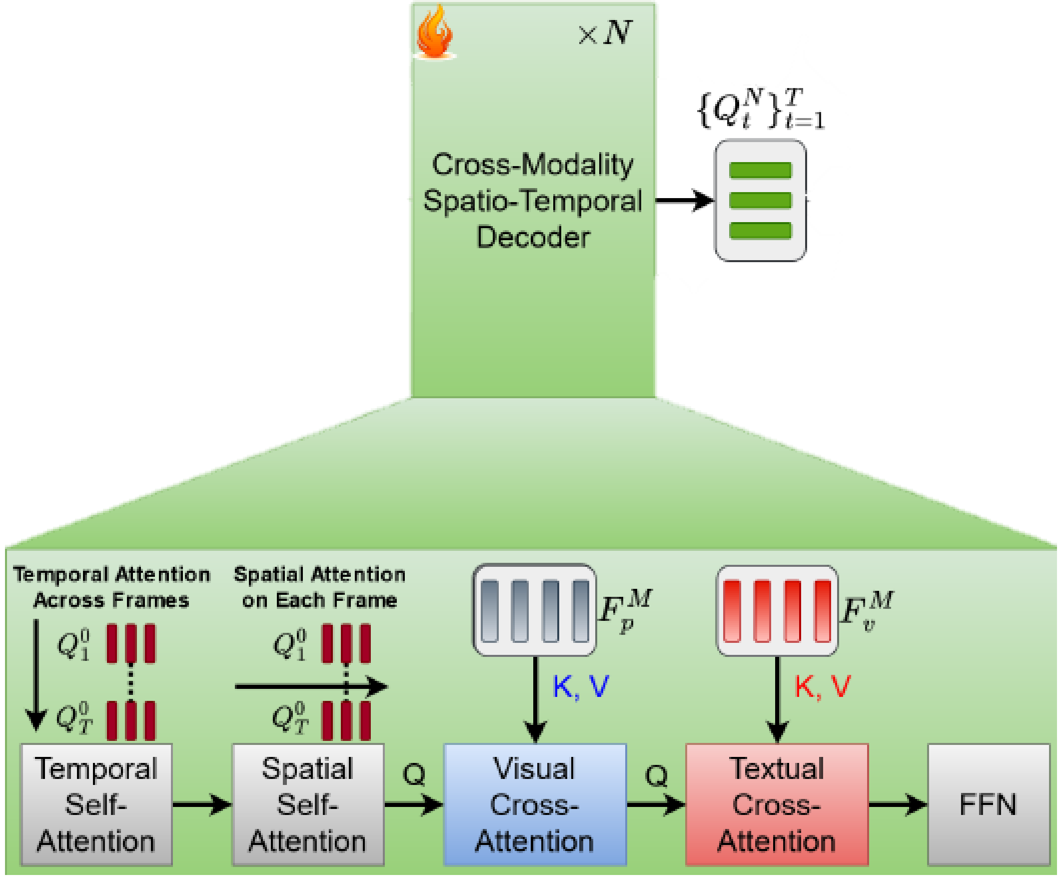
- Input
  - The encoder’s visual and textual features
- Output
  - $\{Q_t^0\}_{t=1}^{T}$:*num_query* indices that correspond to the most relevant features for object detection per frame



Cross-Modality Spatio-Temporal Decoder

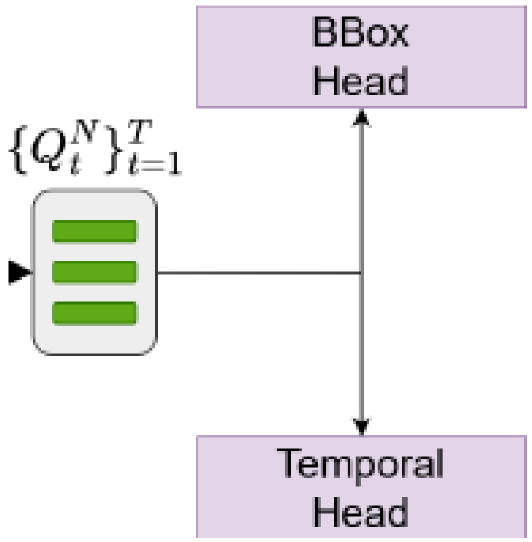
$$Q_t^{n\prime} = \mathrm{MHSA}^n_{spatial}(\mathrm{MHSA}^n_{temporal}(Q_t^{n-1})),$$
$$Q_t^n = \mathrm{FFN}^n(\mathrm{CA}^n_p(\mathrm{CA}^n_v(Q_t^{n\prime}, F_v^M), F_p^M)),$$

$$\mathrm{CA}^n_v(Q_t^{n\prime}, F_v^M) = \left(\frac{proj^n_{q,v}(Q_t^{n\prime})proj^n_{k,v}(F_v^M)^T}{\sqrt{d^k}} proj^n_v(F_v^M)^T\right),$$
$$\mathrm{CA}^n_p(\mathrm{CA}^n_v, F_p^M) = \left(\frac{proj^n_{q,p}(\mathrm{CA}^n_v)proj^n_{k,p}(F_p^M)^T}{\sqrt{d^k}} proj^n_p(F_p^M)^T\right)$$



Prediction Heads

- Output: refined queries
  - Bounding Box Head
  - Temporal Head



Loss Function

$$\mathcal{L}_{spatial} = \lambda_{L_1}\mathcal{L}_{L_1}(\hat{B}, B) + \lambda_{GIoU}\mathcal{L}_{GIoU}(\hat{B}, B)$$

$$\mathcal{L}_{temporal} = \mathcal{L}^{s}_{KL}(\hat{\pi}_s, \pi_s) + \mathcal{L}^{e}_{KL}(\hat{\pi}_e, \pi_e)$$

Evaluation Settings

- Open-Vocabulary Evaluation
  - Training the model on the VidSTG dataset
  - Evaluate it on two different datasets, HC-STVG V1 and YouCook-Interactions to understand how well the model generalizes to new distributions.
    - HC-STVG V1 provides a relatively minor distribution shift given the similar perspective/ objects in the videos.
    - YouCook-Interactions provides a major distribution shift with changes in perspective and annotated objects/ interactions.

Evaluation Settings

- Closed-Set Supervised Evaluation
  - Training on the training set and evaluate each dataset's respective validation/testing set
  - Conducted for three majorly used datasets in spatio-temporal video grounding, namely VidSTG, HC-STVG V1 and HC-STVG V2.

Experimental Results and Analysis

- Open-Vocabulary Evaluation

Method	Pre-training	HC-STVG V1			YouCook-Interactions
		m_vIoU	vIoU@0.3	vIoU@0.5	Accuracy
TubeDETR *(CVPR'22)* [28]	VidSTG	16.84	22.32	9.22	51.63
STCAT *(NeurIPS'22)* [9]	VidSTG	22.58	32.14	20.83	55.90
VideoGrounding-DINO	VidSTG	**27.46**	**40.13**	**29.92**	**57.73**

Experimental Results and Analysis

- Closed-Set Supervised Evaluation

Method	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Factorized:								
GroundeR *(ECCV'16)* [19]+TALL *(ICCV'17)* [7]		9.78	11.04	4.09		9.32	11.39	3.24
STPR *(ICCV'17)* [27]+TALL *(ICCV'17)* [7]	34.63	10.40	12.38	4.27	33.73	9.98	11.74	4.36
WSSTG *(arXiv'19)* [5]+TALL *(ICCV'17)* [7]		11.36	14.63	5.91		10.65	13.90	5.32
GroundeR *(ECCV'16)* [19]+L-Net *(AAAI'19)* [3]		11.89	15.32	5.45		11.05	14.28	5.11
STPR *(ICCV'17)* [27]+L-Net *(AAAI'19)* [3]	40.86	12.93	16.27	5.68	39.79	11.94	14.73	5.27
WSSTG *(arXiv'19)* [5]+L-Net *(AAAI'19)* [3]		14.45	18.00	7.89		13.36	17.39	7.06
Two-Stage:								
STGRN *(CVPR'20)* [32]	48.47	19.75	25.77	14.60	46.98	18.32	21.10	12.83
STGVT *(TCSVT'21)* [24]	-	21.62	29.80	18.94	-	-	-	-
OMRN *(IJCAI'21)* [33]	50.73	23.11	32.61	16.42	49.19	20.63	28.35	14.11
One-Stage:								
STVGBert *(ICCV'21)* [21]	-	23.97	30.91	18.39	-	22.51	25.97	15.95
TubeDETR *(CVPR'22)* [28]	48.10	30.40	42.50	28.20	46.90	25.70	35.70	23.20
STCAT *(NeurIPS'22)* [9]	50.82	33.14	46.20	32.58	49.67	28.22	39.24	26.63
STVGFormer *(CVPR'23)* [11]	-	33.70	47.20	32.80	-	28.50	39.90	26.20
VideoGrounding-DINO	**51.97**	**34.67**	**48.11**	**33.96**	**50.83**	**29.89**	**41.03**	**27.58**

Experimental Results and Analysis

- Closed-Set Supervised Evaluation

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
STGVT *(TCSVT'21)* [24]	18.15	26.81	9.48
STVGBert *(ICCV'21)* [21]	20.42	29.37	11.31
TubeDETR *(CVPR'22)* [28]	32.40	49.80	23.50
STCAT *(NeurIPS'22)* [9]	35.09	57.67	30.09
STVGFormer *(CVPR'23)* [11]	36.90	62.20	34.80
VideoGrounding-DINO	**38.25**	**62.47**	**36.14**

Experimental Results and Analysis

- Closed-Set Supervised Evaluation

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
Yu *et al (arXiv'21)* [30]	30.00	-	-
Aug. 2D-TAN *(arXiv'21)* [22]	30.40	50.40	18.80
TubeDETR *(CVPR'22)* [28]	36.40	58.80	30.60
STVGFormer *(CVPR'23)* [11]	38.70	65.50	33.80
VideoGrounding-DINO	**39.88**	**67.13**	**34.49**

Conclusion

- Performs well in closed-set and open-vocabulary scenarios
  - Surpassing state-of-the-art results in supervised setting on VidSTG and HC-STVG datasets
  - Outperforming recent models in open-vocabulary on HC-STVG V1 and YouCook-Interactions
- Includes learnable adapter blocks for video-specific adaptation, bridging the semantic gap between natural language queries and visual content.

Planned Tasks for This Week

- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár and Christoph Feichtenhofer. **SAM 2: Segment Anything in Images, Videos**. arXiv:2408.00714, Aug 2024.
- Wei Feng, Xin Wang, Hong Chen, Zeyang Zhang, Wenwu Zhu. **Multi-sentence Video Grounding for Long Video Generation.** arXiv:2407.13219, Jul 2024.
- Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz Khan. **Video-GroundingDINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding.** arXiv:2401.00901, Dec 2023
- To be decided.