DEEP LEARNING JP
[DL Papers]

HyperSeg: Towards Universal Visual Segmentation with Large Language Model

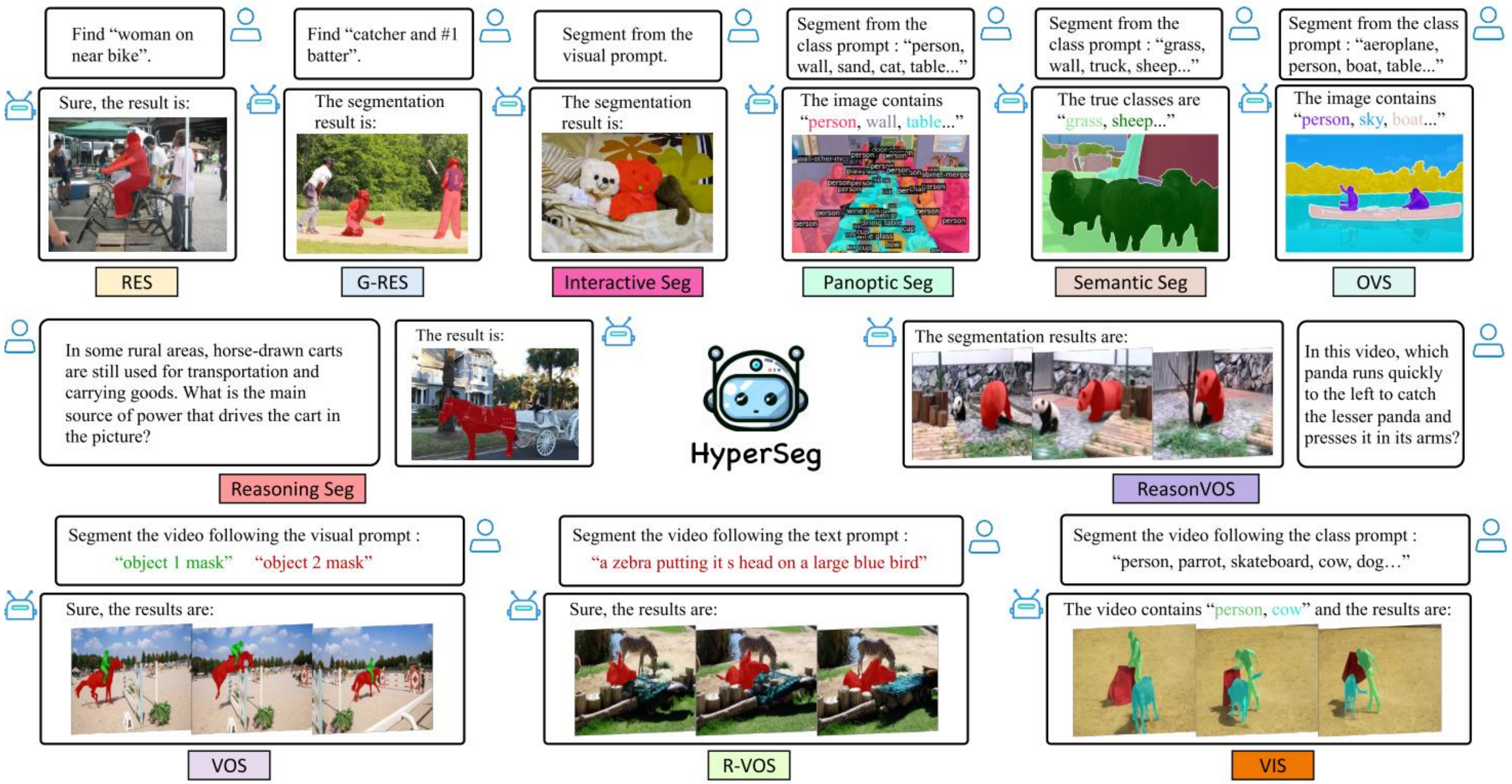
国際航業（株）　林

書誌情報

- タイトル：HyperSeg: Towards Universal Visual Segmentation with Large Language Model
- 投稿先：arxiv(2024年11月末)
- Code：https://github.com/congvvc/HyperSeg
  - 学習コードはまだ未公開（公開予定あり）
- 選定理由：
  - マルチモーダル（特に言語と画像のフュージョン）に興味
  - 様々がタスクに通用するネットワークで対応
  - 様々なbenchmarkで高精度を示した

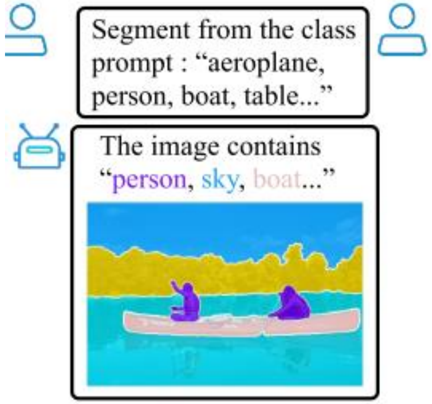
概要

- 様々なpromptに従った多種類のsegmentationタスクを1つのモデルで対応
  - Visual Large Language Models (VLLMs)より、言語に関する知識を習得
  - temporal adapterを提案し、時系列情報を理解して動画にも対応

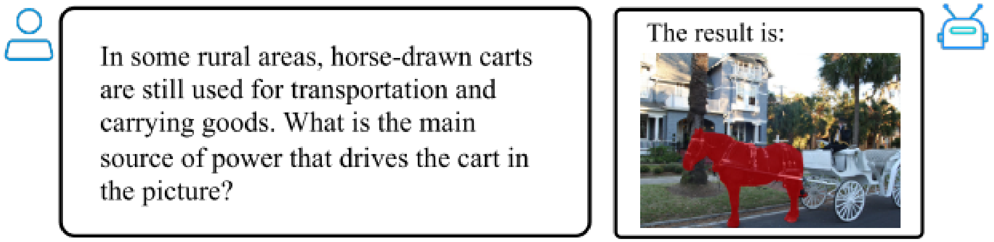


本手法が対応するpromptの種類

- text prompts
  - class names, reasoning questions, referring languages



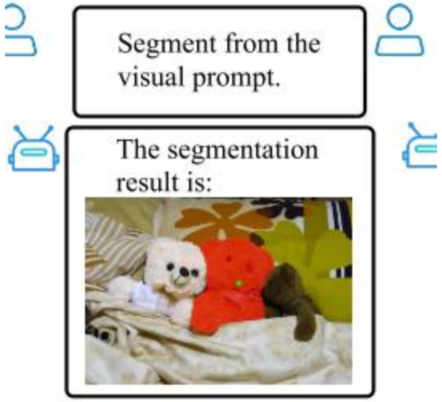
OVS



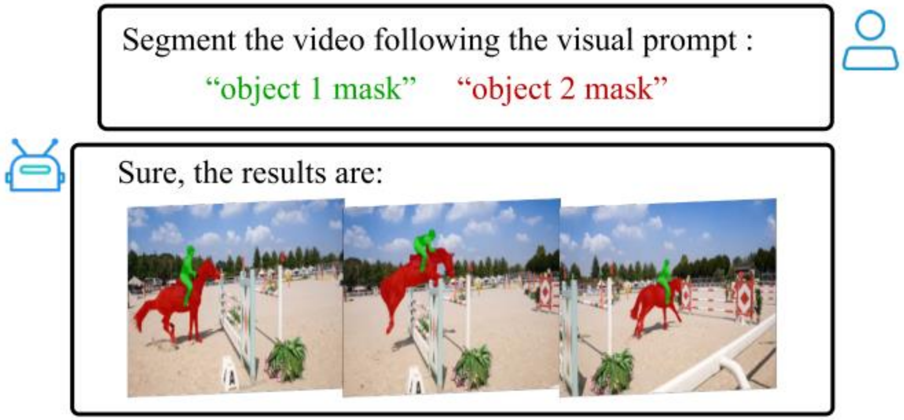
Reasoning Seg

RES

- visual prompts
  - box, mask, pointなど



Interactive Seg



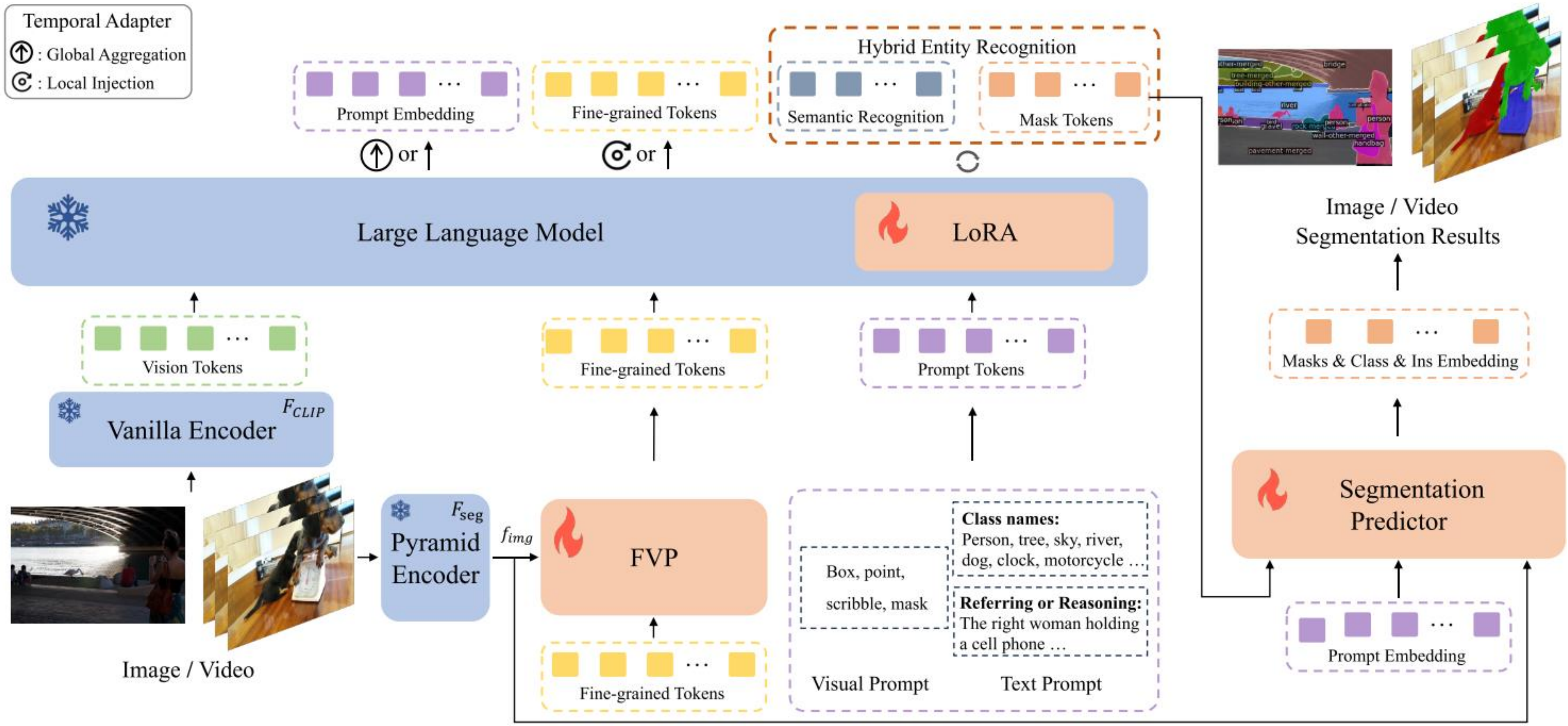
VOS

関連研究

- VLLM
  - 通常は画像のコンテンツを描写する文書を出力。画素レベルの認識に対応不可
  - 代表的な手法：BLIP-2, Flamingo, MiniGPT-4, LLaVA, InstructBLIP , Qwen-VL等
- Perception with VLLM
  - bboxをpromptとして与えて、grounding能力を示した
  - mask decoderをつけることでsegmentationも可能
  - PSALMが初めてVLLMを導入したが、VLLMの性能を十分に引き出せていない
- Unified segmentation model
  - Mask2formerはunifiedネットワークで様々なsegmentationタスクに対応できるが、タスク毎に学習する必要がある
  - OpenSeeDはtext encoderを追加し、Open-Set settingに対応。UNINEXT類似する構造でreferring segmentationに対応。ただし、複雑な文書への対応が困難

提案手法のネットワーク概要

- ネットワーク構成：vision encoder, VLLM, segmentation predictor
- 入力：vision-prompt pairs$\{(\mathcal{V}, \mathcal{P})\}$
- 出力：入力promptに応じたsegmentation masks, class scores, instance embedding（動画の場合）



Prompt design

- モデルの入力：vision-prompt pairs$\{(\mathcal{V}, \mathcal{P})\}$
- prompt $\mathcal{P}$をtext / visual promptに分類
  - $\mathcal{P}_{\mathcal{L}}$: どのようなタスク(instruction)
  - $\mathcal{P}_{\mathcal{C}}$: 具体的なタスク条件
  - visual promptは、その座標でCLIP visual特徴量からsampling

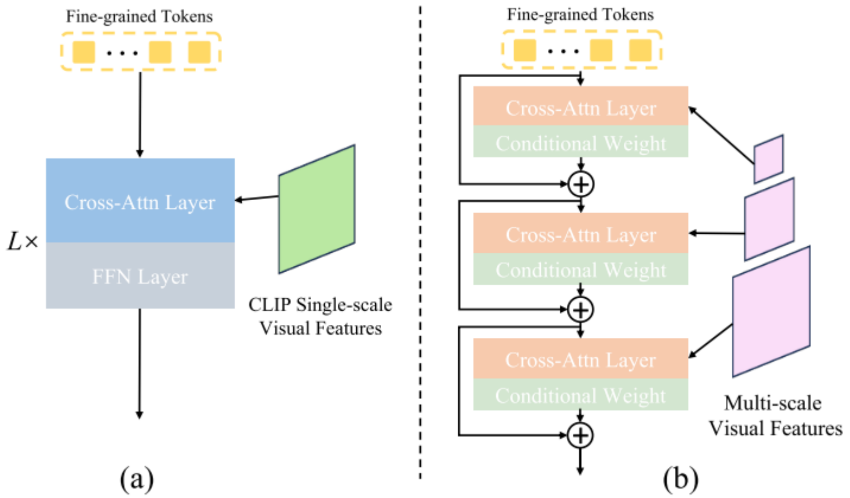
大項目	具体的なタスク	prompt例
class-based segmentation	• panoptic segmentation • open-vocabulary segmentation (OVS) • video instance segmentation (VIS)	$\mathcal{P}_{\mathcal{L}}$: "Please segment all the positive objects according to the following potential categories." $\mathcal{P}_{\mathcal{C}}$: "[category 1, category 2, category 3, ...]"
referring and reasoning segmentation	• referring expression segmentation (RES) • reasoning segmentation • referring video object segmentation (R-VOS) • ReasonVOS	$\mathcal{P}_{\mathcal{L}}$: "Can you perform referring or reasoning segmentation according to the language expression?" $\mathcal{P}_{\mathcal{C}}$: "[referring / reasoning text]"
visual-guided segmentation	• interactive segmentation • video object segmentation (VOS)	$\mathcal{P}_{\mathcal{L}}$: "Please segment according to the given visual region reference" $\mathcal{P}_{\mathcal{C}}$: "[vision 1, vision 2, vision 3, ...]".

Vision Encoder

- 従来手法は、CLIPのvisual encoderの特徴のみ利用することが多い
  - 課題：粒度が高いsegmentationタスクに対して情報が不十分
- Fine-grained Visual Perceiver（FVP）を提案し、粒度の高いvisual情報を抽出（VLLMの入力とする）
  - pyramid vision encoderにより異なるスケールの特徴を抽出
  - 各スケール特徴$f_{img}^{(i)}$とfined-grained token P_jを条件付き重み付きcross-attentionにより、情報を集約

$$\hat{P}_j = \mathrm{MHCA}(P_{j-1}, G_p(f_{img}^{(j)})),$$

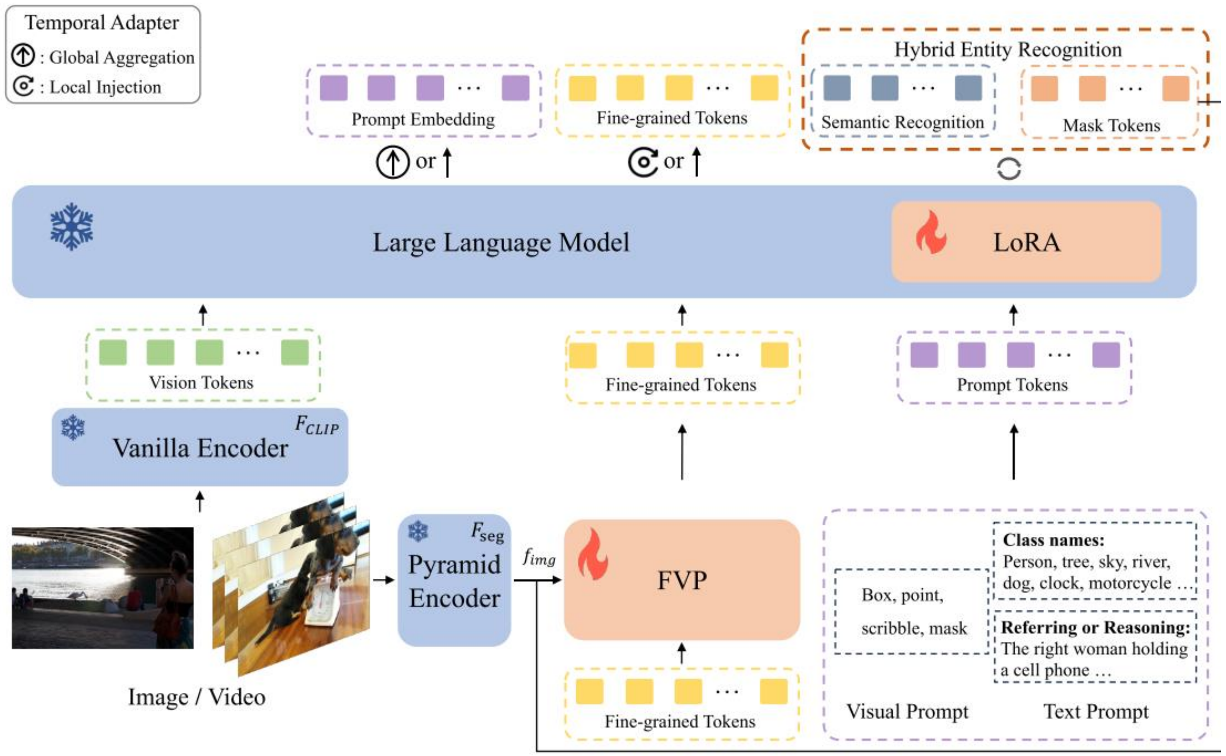
$$P_j = P_{j-1} + \tanh(\mathrm{MLP}(\hat{P}_j)) \cdot \hat{P}_j,$$



Visual Large Language Model

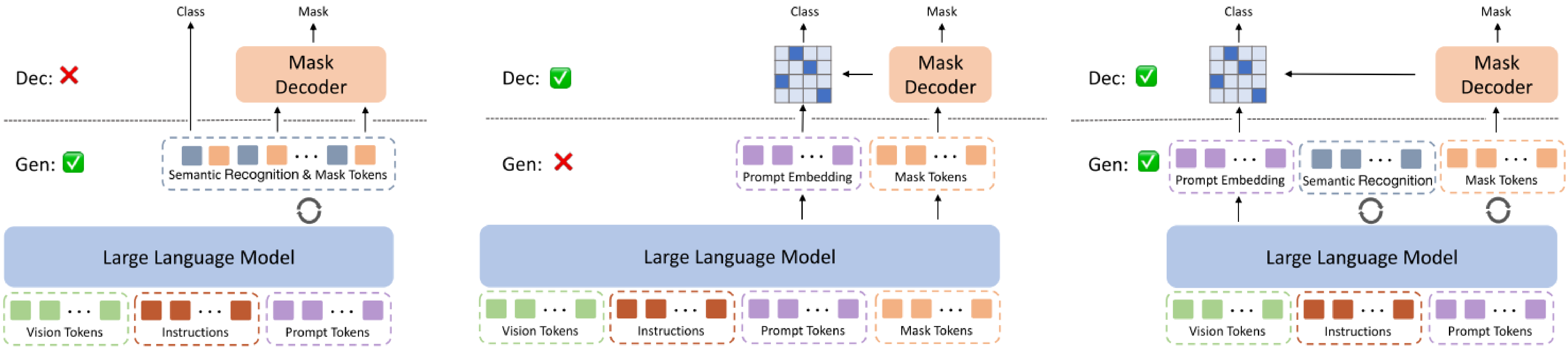
- VLLMは既存モデルを利用
  - visual encoder(CLIP)と軽量化のLLMで構成
- LLMの入力： $f_v = F_{CLIP}(\mathcal{V}), E_O = F_{LLM}(G_c(f_v), P, \mathcal{P}),$
  - vision token f_v: CLIP encoderの出力から取得（画像全体のvisual情報）
  - fined-grained token P ： FVPの出力
  - prompt token $\mathcal{P}$
- LLMの出力
  - prompt embedding
  - semantic recognition
  - mask tokens
  - fine-grained tokens

segmentation predictorに入力



Hybrid Entity Recognition

- LLMを介したsegmentationは3つの流派
  ① クラスとマスクをLLMが生成：漏れや誤検出が多い傾向
  ② クラスとマスクをmask decoderが推定(LLMがprompt tokenをembedする役割):LLMの強力なセマンティックな能力を活かさず
  ③ 本論文はハイブリッドな方式を提案：prompt embeddingをdecodeする。入力画像にあるすべての物体のクラスとそのmask tokenを別々で生成
    - mask tokenと対応するsemantic情報を取得



(a) (b) (c)

Segmentation predictor

- 基本構造は、Mask2Formerを採用
  - 3つの入力でmaskと分類scoreを推定
    - task-specific prompt embedding $\{E_{\mathcal{P}}^{k}\}_{k=1}^{K}$, K =カテゴリー数
    - semantically enhanced mask tokens $\{E_{\mathcal{Q}}^{j}\}_{j=1}^{N}$, N =mask推定個数
    - multi-scale visual features f_{img}

$$\{m_j, z_j, e_j\}_{j=1}^{N} = F_p(\{E_{\mathcal{P}}^{k}\}_{k=1}^{K}, \{E_{\mathcal{Q}}^{j}\}_{j=1}^{N}, f_{img})$$

  - 動画を扱う場合、instance embedding eを推定
    - 動画は、フレーム毎にsegmentationを実施

Temporal Adapter

- 動画を処理する場合、フレーム間の整合性をとる必要がある
- 本論文は、global prompt aggregationとlocal space-time information injectionを提案
  - 前の全フレームのprompt embeddingをpooling $E_{\mathcal{P}} = AvgPool([E_{\mathcal{P}}^0, E_{\mathcal{P}}^1, ..., E_{\mathcal{P}}^T]),$
  - 前の1枚フレームのfine-grained tokenから更新 $P_t = G_l[F_{LLM}(P_{t-1})],$

学習目的関数

- 各タスクに共通するloss関数で学習
  - $\mathcal{L}_{text}$: autoregressive cross-entropy loss for text prediction
    - 論文中詳細は言及せず。Supplementaryの情報から、image captionやVQAのようなタスクでvisual-languageの理解を保持（？）
  - $\mathcal{L}_{mask}$: mask推定loss
  - $\mathcal{L}_{cls}$: カテゴリー分類cross entropy loss
  - $\mathcal{L}_{ins}$: contrastive loss for instance association（動画の場合）

$$\mathcal{L} = \mathcal{L}_{text} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{ins}\mathcal{L}_{ins},$$

$$\mathcal{L}_{mask} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice},$$

実験設定

- 合計10個のタスクを同時に学習
  - 各タスクは約16k iterationを学習
  - VLLMはMipha3Bを採用。LLM部分の学習はLoRA採用
    - LLMはPhi-2-2.7Bを採用。Visual encoderはSigLIPを採用
  - Segmentation predictorはMask2Formerを採用
  - 8 NVIDIA A100 GPUsで学習（batch size=32）

実験結果- Referring expression segmentation

- RefCOCO/+/gにおいて、SOTAを達成
- 更に難しいgeneralized referring expression segmentationでも有効
  - Zero-shot形式で評価

Table 1. Comparison with the state-of-the-art models on the closed-set referring segmentation benchmarks (RefCOCO series) and more challenging generalized referring expression segmentation benchmark gRefCOCO. ‡ denotes models using pre-trained SAM [21] for mask generation. * means using gRefCOCO for training while other methods are evaluated in zero-shot manners. Our HyperSeg exhibits excellent performance over other zero-shot models like LaSagnA [45] and PSALM [60].

Type	Method	RefCOCO			RefCOCO+			RefCOCOg		gRefCOCO		
		val	testA	testB	val	testA	testB	val(U)	test(U)	val	testA	testB
Segmentation Specialist	VLT [11]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7	52.5*	62.2*	50.5*
	CRIS [44]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	55.3*	63.8*	51.0*
	LAVT [54]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	57.6*	65.3*	55.0*
	PolyFormer-B [30]	74.8	76.6	71.1	67.6	72.9	59.3	67.8	69.1	-	-	-
MLLM-based Segmentation Network	LISA-7B [22] ‡	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	38.7*	52.6*	44.8*
	PixelLM-7B [41]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	-
	F-LMM-7B [49] ‡	76.1	-	-	66.4	-	-	70.1	-	-	-	-
	GSVA-7B [50] ‡	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0	61.7*	69.2*	60.3*
	GroundHog-7B [33]	78.5	79.9	75.7	70.5	75.0	64.9	74.1	74.6	66.7*	-	-
	SAM4MLLM-7B [6] ‡	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	66.3*	70.1*	63.2*
	LaSagnA-7B [45] ‡	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	38.1	50.4	42.1
	OMG-LLaVA [59]	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9	-	-	-
	GLaMM [40] ‡	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-	-
	PSALM [60]	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	42.0	52.4	50.6
	HyperSeg	**84.8**	**85.7**	**83.4**	**79.0**	**83.5**	**75.2**	**79.4**	**78.9**	**47.5**	**57.3**	**52.5**

実験結果-Reasoning segmentation

- 動画と画像ドメインにおいて、SOTAを達成

Table 2. Comparison with the state-of-the-art models on more complex and challenging reasoning segmentation benchmarks: ReVOS in video domain and ReasonSeg in image domain. ‡ denotes the same meaning as Tab. 1. Our HyperSeg outperforms all the previous VLLM-based models in both video and image reasoning segmentation tasks.

Method	Backbone	ReVOS-Reasoning			ReVOS-Referring			ReVOS-Overall			ReasonSeg	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	gIoU	cIoU
LMPM [12]	Swin-T	13.3	24.3	18.8	29.0	39.1	34.1	21.2	31.7	26.4	-	-
ReferFormer [47]	Video-Swin-B	21.3	25.6	23.4	31.2	34.3	32.7	26.2	29.9	28.1	-	-
LISA-7B [22] ‡	ViT-H	33.8	38.4	36.1	44.3	47.1	45.7	39.1	42.7	40.9	52.9	54.0
LaSagnA-7B [45] ‡	ViT-H	-	-	-	-	-	-	-	-	-	48.8	47.2
SAM4MLLM-7B [6] ‡	EfficientViT-SAM-XL1	-	-	-	-	-	-	-	-	-	46.7	48.1
TrackGPT-13B [63] ‡	ViT-H	38.1	42.9	40.5	48.3	50.6	49.5	43.2	46.8	45.0	-	-
VISA-7B [52] ‡	ViT-H	36.7	41.7	39.2	51.1	54.7	52.9	43.9	48.2	46.1	52.7	**57.8**
VISA-13B [52] ‡	ViT-H	38.3	43.5	40.9	52.3	55.8	54.1	45.3	49.7	47.5	-	-
HyperSeg-3B	Swin-B	**50.2**	**55.8**	**53.0**	**56.0**	**60.9**	**58.5**	**53.1**	**58.4**	**55.7**	**59.2**	56.7

実験結果-Generic image segmentation

- closed-set and open-vocabulary segmentation両方に効果を確認

Table 3. Quantitative results on the closed-set COCO-Panoptic segmentation, open-vocabulary segmentation (-OV) benchmarks. Our model HyperSeg achieves remarkable performance compared with the previous state-of-the-art methods.

Type	Method	Backbone	COCO-Panoptic		ADE-OV		Citys-OV	PC59-OV	PAS20-OV
			PQ	mIoU	PQ	mIoU	PQ	mIoU	mIoU
Segmentation Specialist	Mask2former [7]	Swin-B	55.1	65.1	-	-	-	-	-
	OneFormer [19]	Swin-L	57.9	67.4	-	-	-	-	-
	SEEM [65]	DaViT-B	56.1	66.3	-	-	-	-	-
	MaskCLIP [13]	ViT-L	30.9	47.6	15.1	**23.7**	-	45.9	-
	DeOP [16]	ResNet-101c	-	-	-	22.9	-	48.8	91.7
	SimBaseline [51]	ViT-B	-	-	-	20.5	-	47.7	88.4
	DaTaSeg [15]	ViTDet-B	52.8	62.7	12.3	18.3	28.0	51.1	-
MLLM-based Segmentation Network	OMG-LLaVA [59]	ConvNeXt-L	53.8	-	-	-	-	-	-
	PSALM [22]	Swin-B	55.9	66.6	13.7	18.2	28.8	48.5	81.3
	HyperSeg	Swin-B	**61.2**	**77.2**	**16.1**	22.3	**31.1**	**64.6**	**92.1**

実験結果-Common video segmentation

- 具体的には、visual-prompted semi-supervised VOS (DAVIS17), text-prompted referring video object segmentation (Ref-YouTube-VOS, Ref-DAVIS17), video instance segmentation (YouTube-VIS 2019)で評価

Table 4. Results of common video segmentation benchmarks, including DAVIS17, Ref-YouTube-VOS, Ref-DAVIS17, and YouTube-VIS 2019. ‡ denotes the same meaning as Tab. 1.

Method	Backbone	DAVIS17	Ref-YT	Ref-DAVIS	YT-VIS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	mAP
SEEM [65]	DaViT-B	62.8	-	-	-
OMG-Seg [24]	ConvNeXt-L	74.3	-	-	56.4
ReferFormer [47]	Video-Swin-B	-	62.9	61.1	-
OnlineRefer [46]	Swin-L	-	63.5	64.8	-
UNINEXT [25]	ConvNeXt-L	77.2	66.2	66.7	**64.3**
LISA-7B [22] ‡	ViT-H	-	53.9	64.8	-
VISA-13B [52] ‡	ViT-H	-	63.0	70.4	-
VideoLISA-3.8B [3] ‡	ViT-H	-	63.7	68.8	-
HyperSeg-3B	Swin-B	**77.6**	**68.5**	**71.2**	53.8

実験結果-Ablations

- 複数タスクの同時に学習することで、モデル性能を向上
  - 特に動画segmentationタスクにおいて、画像segmentationも学習する効果が大きい

Table 5. The mutual influence between different tasks. Task-specific means training task-specific models only on data from corresponding tasks, Refer+Reason denotes the model is trained on referring and reasoning segmentation data, and Video and Image denote different training visual types: training on video data and image data, respectively.

Task-specific	Refer+Reason	Video	Image	RefCOCO			COCO		ReVOS			YT-VIS
				val	testA	testB	PQ	mIoU	Reasoning	Referring	Overall	mAP
✓				83.8	85.9	82.2	60.8	75.1	51.2	56.6	53.9	50.7
	✓			83.3	84.9	80.9	-	-	53.1	57.3	55.2	-
			✓	85.6	86.1	82.4	60.9	76.5	-	-	-	-
		✓		-	-	-	-	-	51.1	57.0	54.1	50.4
	✓	✓	✓	84.8	85.7	83.4	61.2	77.2	53.0	58.5	55.7	53.8

- 提案手法は別のLLMでも効果を発揮

Table 6. The comparison of different LLMs and backbone usages. w/o CLIP means without using CLIP vision encoder.

Method	LLM	COCO		ReVOS			ADE-OV	PC59-OV	PAS20-OV
		PQ	mIoU	Reasoning	Referring	Overall	mIoU	mIoU	mIoU
LISA [22]	Vicuna-7B	-	-	36.1	45.7	40.9	-	-	-
VISA [52]	Vicuna-13B	-	-	40.9	54.1	47.5	-	-	-
PSALM(w/o CLIP) [22]	Phi-1.5-1.3B	55.9	66.6	-	-	-	18.2	48.5	81.3
HyperSeg (w/o CLIP)	Phi-1.5-1.3B	61.1	76.0	44.0	49.7	46.9	18.9	60.0	90.6
HyperSeg	Phi-1.5-1.3B	60.9	76.7	50.8	57.0	53.9	20.3	61.5	90.8
HyperSeg	Phi-2-2.7B	**61.2**	**77.2**	**53.0**	**58.5**	**55.7**	**22.3**	**64.6**	**92.1**

実験結果-Ablations

- 提案のFVPとHERの有効性を確認
- 動画に対し、global prompt aggregation（全フレーム情報のpooling）とlocal space-time information injection（前1フレームの情報を更新）の効果を確認

Table 7. Ablation on the core components of HyperSeg. FVP and HER denote the proposed Fine-grained Visual Perceiver and Hybrid Entity Recognition modules.

FVP	HER	YT-VIS	COCO		RefCOCO
		mAP	PQ	mIoU	cIoU
		48.4	54.8	66.2	82.8
✓		50.8	55.8	66.6	84.6
	✓	52.0	59.7	74.6	84.3
✓	✓	**53.8**	**61.2**	**77.2**	**84.8**

Table 8. Ablation on the Fine-grained Visual Perceiver design. CW denotes the Conditional Weight illustrated in Sec. 3.3, and Scale denotes the total scale in the proposed FVP module.

CW	Scale	YT-VIS	COCO		RefCOCO
		mAP	PQ	mIoU	cIoU
	single-layer	49.7	55.8	68.0	83.7
	multi-layers	50.4	58.9	73.4	84.5
✓	multi-layers	**53.8**	**61.2**	**77.2**	**84.8**

Table 9. Ablation on the temporal adapter for video tasks, including global prompt aggregation (global) and local space-time information injection (local).

Global	Local	Ref-DAVIS17	ReVOS	YT-VIS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	mAP
		67.3	54.1	47.9
✓		68.8	54.5	48.5
	✓	69.3	54.8	50.2
✓	✓	**71.2**	**55.7**	**53.8**

まとめ

- VLLMを利用し、様々なsegmentationタスクを一つのモデルで対応できる手法を提案
  - 異なるスケールのvisual情報を利用
  - VLLMの出力形式に工夫
  - 画像と動画両方に対応可能。特に画像関連の複数タスクでは、SOTAを達成
- 所感
  - 既存のモデルをうまく組み合わせて、様々なタスクを一つ比較的に小さいモデルで対応
  - Mask2Formerの形に合わせてLLMを組み合わせた気もしなくない