Visual Exemplar Driven Task-Prompting for Unified Perception in Autonomous Driving

Xiwen Liang, Minzhe Niu, Jianhua Han, Hang Xu,

Chunjing Xu, Xiaodan Liang†

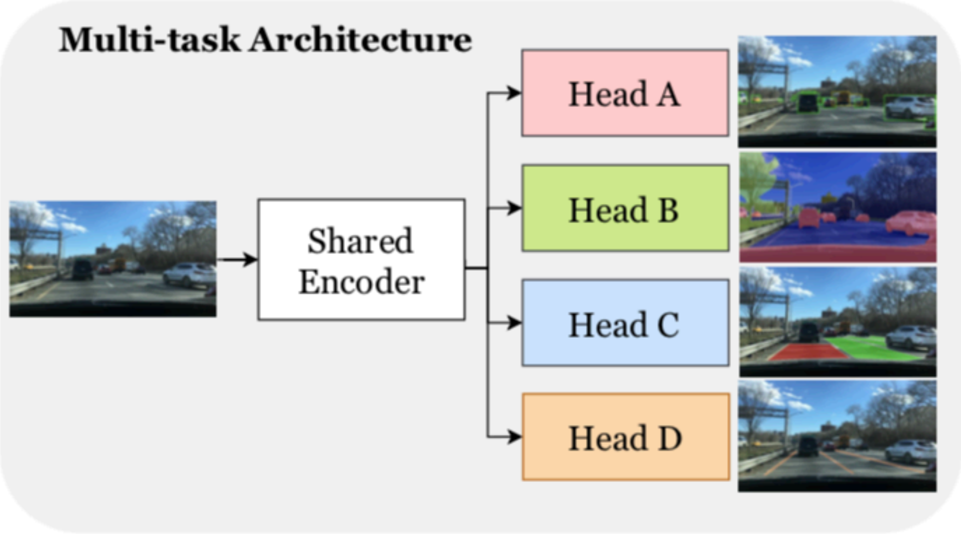
Paper Tag: WED-AM-132

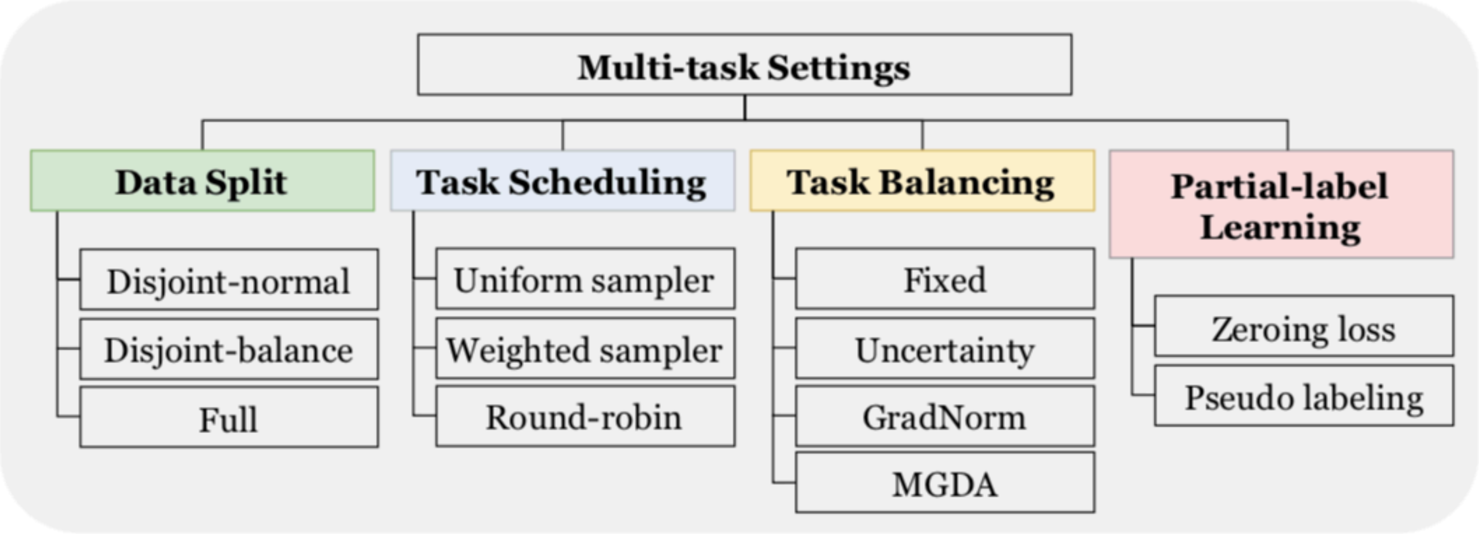
中山大學
SUN YAT-SEN UNIVERSITY

HUAWEI | NOAH'S ARK LAB

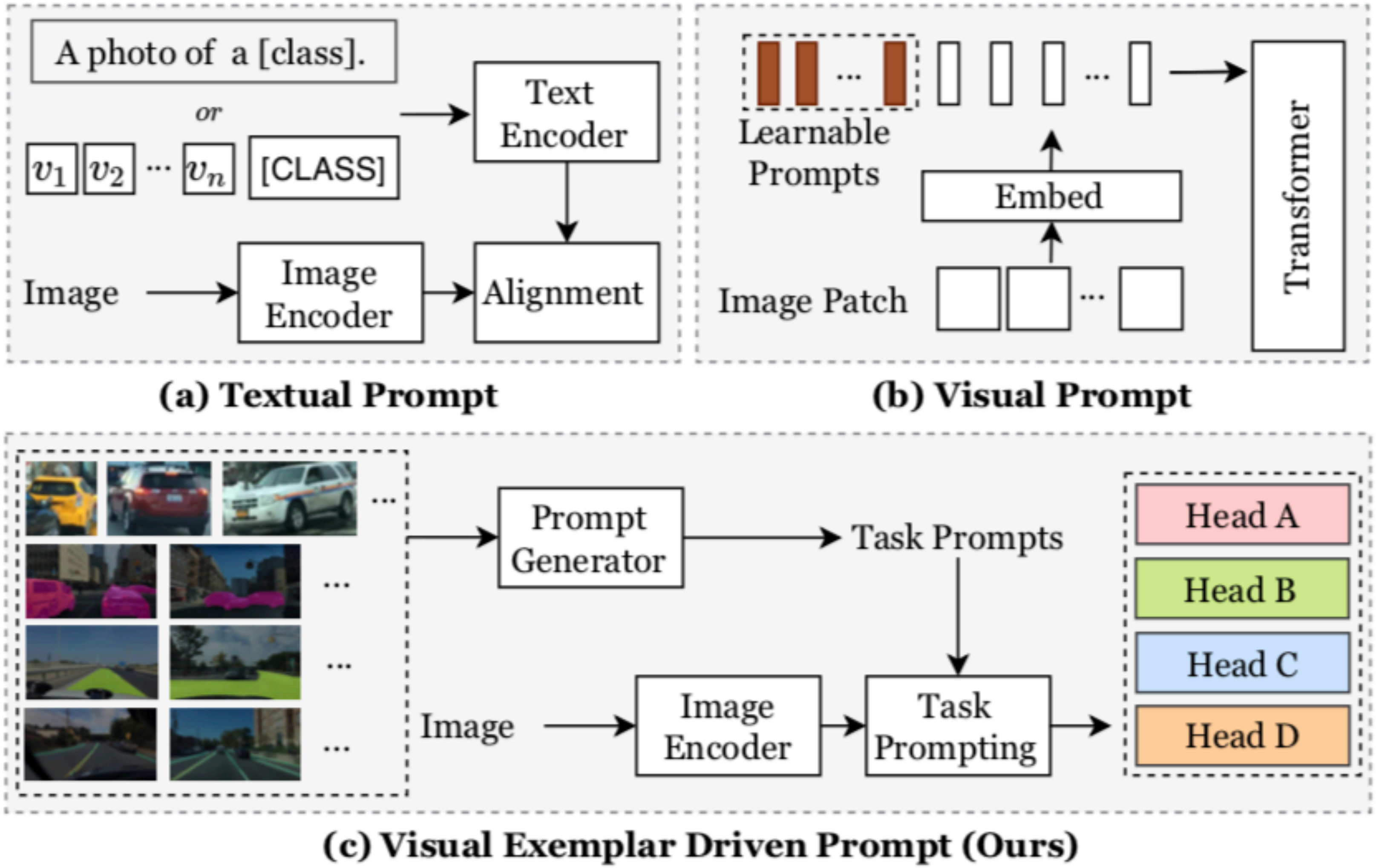
Related Work

- Multi-task Architectures
  - encoder-focused architectures
  - decoder-focused architectures
- Multi-task Settings

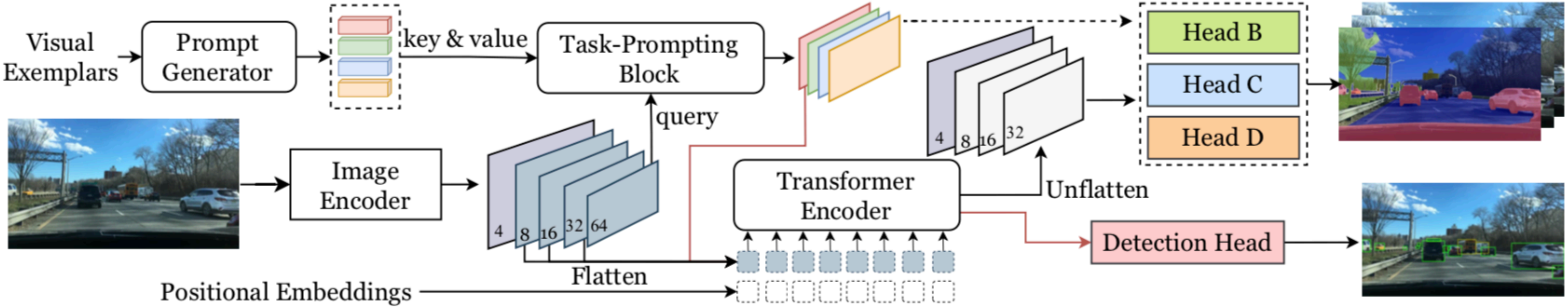




Related Work



VE-Prompt



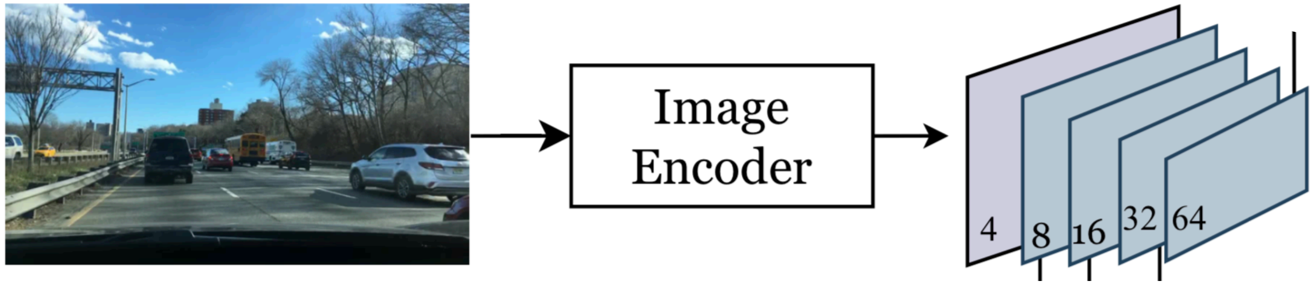
- the image encoder to extract image features
- a shared transformer encoder for feature enhancement
- task-specific prompts generated by the prompt generator with visual exemplars
- a task-prompting block to integrate the visual representation with task-specific prompts
- task-specific heads for different tasks.

VE-Prompt

- Bridging CNN and Transformer
  - Image Encoder
    - Swin-FPN
  - Shared Transformer Encoder

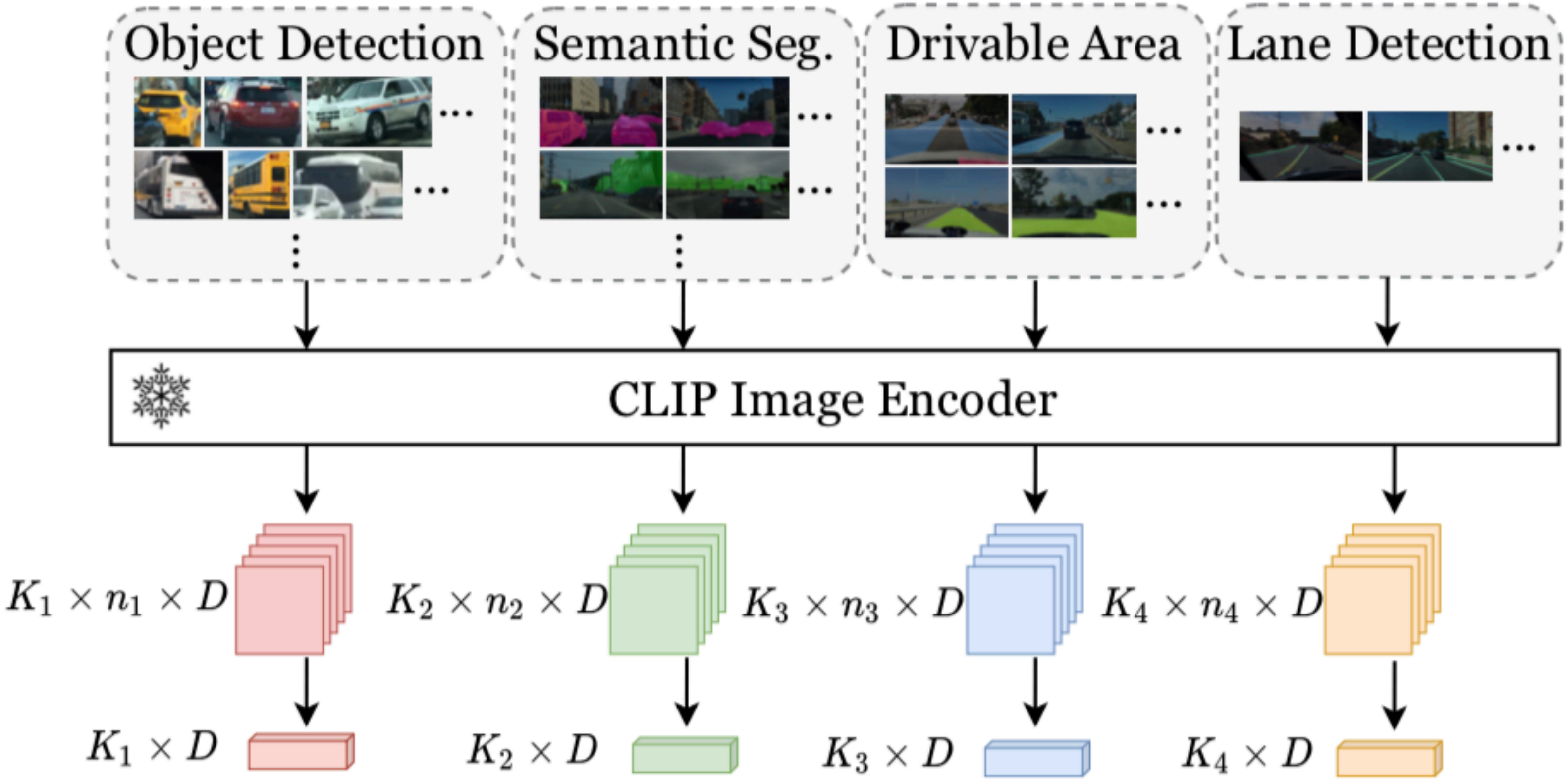
$$O = \text{TransEncoder}(P + p_l).$$

  - Detection Head
    - DINO
  - Segmentation Head
    - Semantic FPN



VE-Prompt

- Prompt Generation with Visual Exemplar



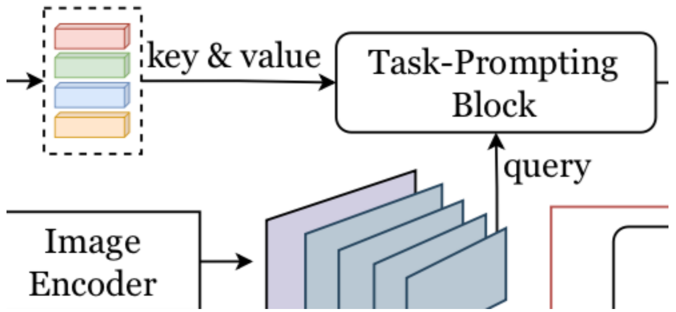
$$\{\hat{p_i^k}\} = \mathrm{L2\_NORM}(\mathrm{IE}(\{r_i^k\})) \in \mathbb{R}^{K \times n \times D},$$

$$p = \frac{1}{n}\sum_{i}^{n} \{\hat{p_i^k}\} \in \mathbb{R}^{K \times D}, i = 1, 2, ..., n,$$

VE-Prompt

- Visual Exemplar Driven Task Prompting
  - Pre-head prompting

$$f_{pre} = \mathrm{TransDecoder}(q = P_6, k = p, v = p),$$



  - Post-head prompting

$$f_{post} = \mathrm{TransDecoder}(q = p, k = P_6, v = P_6).$$

$$v' = \mathrm{MLP}(v \cdot f_{post}).$$

Experiment Setting

- **Disjoint-normal setting**
  - The number of labeled images for each task is as follows: object detection (10k), semantic segmentation (7k), drivable area segmentation (20k), and lane detection (20k).
- **Disjoint-balance setting**
  - There are 28k images in this set and each task has 7k labeled images that are not overlapped with other tasks.
- **Full setting**
  - Full-setting refers to experimenting on all available annotations on ∼74k images in BDD100K and can be used to analyze the upper bound of different methods.

Evaluation Metric

- Whole multi-task performance [1]

$$\Delta_{MTL} = \frac{1}{T}\sum_{i}^{T}(M_{m,i} - M_{b,i})/M_{b,i},$$

where $M_{m,i}$ is the performance of multi-task model on task i, and $M_{b,i}$ indicates the result of single-task baseline.

- Average performance

$$Avg = \frac{1}{4}(mAP + mIoU(SS) + mIoU(DA) + mIoU(LD))$$

[1] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 2021.

Experiment (task scheduling and partial-label learning)

Setting	Methods	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)	Avg.	$\Delta_{MTL}(\%)$
Full	Zeroing loss [51]	36.2	61.6	35.9	58.6	89.3	23.8	52.0	-2.68
	Pseudo labeling [15]	36.3	61.6	36.1	60.9	89.3	23.8	52.6	-1.65
	VE-Prompt (Ours)	**39.2**	**64.9**	**39.0**	**63.2**	**89.4**	**24.0**	**54.0**	**+1.52**
Disjoint-normal	Zeroing loss [51]	31.1	54.3	30.2	55.7	88.0	22.2	49.3	-2.64
	Uniform sampler [26]	30.1	52.8	29.0	60.6	88.6	23.4	50.7	-0.10
	Weighted sampler [26]	29.3	51.9	28.7	58.5	**88.9**	**23.8**	50.1	-1.19
	Round-robin [26]	30.2	53.1	29.7	61.0	88.7	23.5	50.9	+2.87
	Pseudo labeling [15]	32.6	54.6	32.3	59.7	88.2	23.0	50.9	+1.19
	VE-Prompt (Ours)	**34.2**	**56.9**	**33.9**	**62.2**	88.3	23.3	**52.0**	**+3.95**
Disjoint-balance	Zeroing loss [51]	29.7	52.3	29.2	57.5	86.7	21.4	48.8	-1.61
	Uniform sampler [26]	28.1	50.2	27.5	60.4	87.1	**22.6**	50.0	-0.44
	Round-robin [26]	28.4	50.8	27.8	60.0	87.1	**22.6**	49.5	-0.34
	Pseudo labeling [15]	31.3	52.8	30.8	60.2	87.0	22.2	50.2	+1.87
	VE-Prompt (Ours)	**33.9**	**56.6**	**33.7**	**61.2**	**87.4**	22.2	**51.2**	**+4.72**

Experiment (task balancing)

Setting	Method	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)	Avg.	$\Delta_{MTL}(\%)$
Full	Fixed [15]	36.3	61.6	36.1	60.9	89.3	23.8	52.6	-1.65
	Uncertainty [21]	36.2	61.6	35.5	61.2	**89.5**	**24.6**	52.9	-0.76
	GradNorm [5]	23.4	40.9	22.8	25.8	51.3	13.0	28.4	-46.24
	VE-Prompt (Ours)	**39.2**	**64.9**	**39.0**	**63.2**	89.4	24.0	**54.0**	**+1.52**
Disjoint-normal	Fixed [15]	32.6	54.6	32.3	59.7	88.2	23.0	50.9	+1.19
	Uncertainty [21]	32.2	54.1	31.5	59.8	**88.6**	23.8	51.1	+1.79
	GradNorm [5]	25.9	43.2	26.1	39.2	39.6	3.7	27.1	-46.18
	MGDA [41]	25.9	44.6	26.0	50.1	85.4	**25.2**	46.7	-7.26
	VE-Prompt (Ours)	**34.2**	**56.9**	**33.9**	**62.2**	88.3	23.3	**52.0**	**+3.95**
Disjoint-balance	Fixed [15]	31.3	52.8	30.8	60.2	87.0	22.2	50.2	+1.87
	Uncertainty [21]	31.2	53.1	30.9	59.9	87.0	22.2	50.1	+1.66
	GradNorm [5]	28.9	49.0	28.7	46.8	57.4	19.6	38.2	-17.26
	GradNorm* [5]	30.7	51.8	30.4	56.6	86.9	21.7	49.0	-0.73
	MGDA [41]	21.0	38.0	20.3	45.5	82.7	**24.3**	43.4	-12.48
	VE-Prompt (Ours)	**33.9**	**56.6**	**33.7**	**61.2**	**87.4**	22.2	**51.2**	**+4.72**

Comparison with single-task and multi-task learning baselines

Setting	Methods	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)	Avg.	$\Delta_{MTL}(\%)$
Full	Sparse R-CNN [43]	36.5	61.5	36.1	-	-	-	-	-
	DINO [59]	38.6	64.2	38.2	-	-	-	-	-
	Semantic FPN [22]	-	-	-	59.8	-	-	-	-
	Semantic FPN [22]	-	-	-	-	89.1	-	-	-
	Semantic FPN [22]	-	-	-	-	-	25.9	-	-
	Sparse R-CNN based	36.3	61.6	36.1	60.9	89.3	23.8	52.6	-1.65
	DINO based	**39.4**	64.5	**39.8**	61.5	84.9	22.0	52.0	-2.25
	VE-Prompt (Ours)	39.2	**64.9**	39.0	**63.2**	**89.4**	**24.0**	**54.0**	**+1.52**
Disjoint-normal	Sparse R-CNN [43]	28.8	50.4	28.0	-	-	-	-	-
	DINO [59]	31.2	53.0	30.5	-	-	-	-	-
	Semantic FPN [22]	-	-	-	59.8	-	-	-	-
	Semantic FPN [22]	-	-	-	-	87.8	-	-	-
	Semantic FPN [22]	-	-	-	-	-	25.2	-	-
	Sparse R-CNN based	32.6	54.6	32.3	59.7	88.2	23.0	50.9	+1.19
	DINO based	33.1	55.9	32.2	59.2	87.2	22.7	50.6	+0.83
	VE-Prompt (Ours)	**34.2**	**56.9**	**33.9**	**62.2**	**88.3**	**23.3**	**52.0**	**+3.95**
Disjoint-balance	Sparse R-CNN [43]	28.1	49.2	26.7	-	-	-	-	-
	DINO [59]	29.4	50.8	28.1	-	-	-	-	-
	Semantic FPN [22]	-	-	-	59.8	-	-	-	-
	Semantic FPN [22]	-	-	-	-	85.5	-	-	-
	Semantic FPN [22]	-	-	-	-	-	23.7	-	-
	Sparse R-CNN based	31.3	52.8	30.8	60.2	87.0	**22.2**	50.2	+1.87
	DINO based	33.5	55.6	33.1	58.1	85.2	21.4	50.0	+1.58
	VE-Prompt (Ours)	**33.9**	**56.6**	**33.7**	**61.2**	**87.4**	**22.2**	**51.2**	**+4.72**

Experiment (nuImages)

- We also conduct experiments on nuImages dataset, which covers two tasks, object detection and semantic segmentation.

Model	mAP	AP50	AP75	mIoU	Avg.
Sparse R-CNN based	50.4	76.8	54.5	53.8	52.1
DINO based	55.5	81.6	60.6	56.7	56.1
VE-Prompt (Ours)	**55.8**	**81.9**	**60.7**	**59.1**	**57.5**

Ablation study

- Modules in VE-Prompt

	mAP	mIoU (SS)	mIoU (DA)	IoU (LD)
DINO based	33.5	58.1	85.2	21.4
w/ shared TE	32.2	60.5	86.5	21.4
+ Prompt	**33.9**	**61.2**	**87.4**	**22.2**

- Task-specific prompts

#	Prompt	Post	Pre	mAP	mIoU (SS)	mIoU (DA)
1	✗	✗	✗	32.2	60.5	86.5
2	✓	✓	✗	33.2	58.9	86.4
3	✓	✗	✓	**33.9**	**61.2**	**87.4**

Ablation study

- Initialization for prompt vectors

CLIP Initialization	mAP	mIoU (SS)	mIoU (DA)	IoU (LD)
✗	33.5	61.0	87.2	21.9
✓	**33.9**	**61.2**	**87.4**	**22.2**

Fixed	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)
✓	33.3	55.3	32.5	61.1	87.2	22.1
✗	**33.9**	**56.6**	**33.7**	**61.2**	**87.4**	**22.2**

Conclusion

- We provide an in-depth analysis of popular multi-task learning methods under the realistic scenarios of self-driving, which covers four common perception tasks, i.e., object detection, semantic segmentation, drivable area segmentation, and lane detection.
- We propose visual exemplar driven task-prompting (VE- Prompt), which incorporates visual exemplars of different tasks to provide high-quality task-specific knowledge. Be- sides, the proposed framework bridges transformer and convolutional layers for efficient and accurate unified perception in autonomous driving.
- Experimental results show that VE-Prompt can achieve superior performance on large- scale driving dataset BDD100K.

Thank you