Aligning and Prompting Everything All at Once for Universal Visual Perception

Supplementary Material

6. Appendix

6.1. Model Structure Details

In this section, we compare APE with other models from the perspective of model structure. As shown in Tab. 9, our model has a significantly different framework. Compared to GLIPv2 [51] and UNINEXT [48], APE uses a smaller input size for the long side and has only half the number of parameters.

6.2. Training Data Details

We compare the data usage of various APE and other models in Tab. 10. It shows that our method consumes the least images during training while achieving superior performance. The main reason is two-fold: First, we enable to query APE with a large number of prompts, which speeds up model coverage. Second, we design image-centri format to group grounding data, efficiently reducing training iterations and speedup coverage. Based on the three principles in Sec. 3.3, we configure the sampling ratios and loss weights for all datasets as shown in Tab. 11.

6.3. Implementation Details

We build on DETA [32] to implement our model. DETA has a simpler alternative training mechanism to learn an easier decoding function with IoU-based label assignment. We use 900 queries and 6 encoder and decoder layers. For the visual backbone, we adopt pre-trained ViT-L [9] by default and also use ReseNet-50 [13] in our ablation studies. We adopt the pre-trained large model in EVA-CLIP [40] for the language backbone. We use the AdamW [28] optimizer with a weight decay of 0.05 and a learning rate $2e{-}4$, which is decayed at 0.88 fractions of the total number of steps by 10. We also compare our structure to other models for the largest model size in Tab. 9.

For data augmentation, we use the default large-scale jittering [10] augmentation with a random scale sampled from the range 0.1 to 2.0 for all datasets. For COCO [25], instead of panoptic mask annotations, we utilize 80-category instance-level and 53-category semantic-level annotations as the supervision signal. We also apply repeat factor sampling [12] and copy-paste augmentation [10] on LVIS [12]. Detailed descriptions of implementation are available in the supplementary material.

6.4. Additional Result of Visual Grounding

We further conduct experiments on RefCOCO/+/g datasets with other models that only require a single stage of training. As shown in Tab. 12, APE surpasses all other methods with large performance gaps.

6.5. Visualization

In this subsection, we demonstrate the generalization ability to various datasets and flexibility to support task compositions for APE with qualitative visualizations.

In Fig. 3, we first visualize the model outputs for instance and semantic segmentation tasks. Noted that all results for both tasks are the same outputs from APE-D, except for different post-processing. For instance segmentation, we apply non-maximum suppression on predicted regions. For semantic segmentation, we further accumulate the semantic masks for the same concepts as described in subsec. 3.2.

We further present some visualizations in Figs. 4, 5 and 6 on D3 [46], on which APE outperforms all previous methods with a large gap. Our APE presents great generalization on different scenes and text inputs.

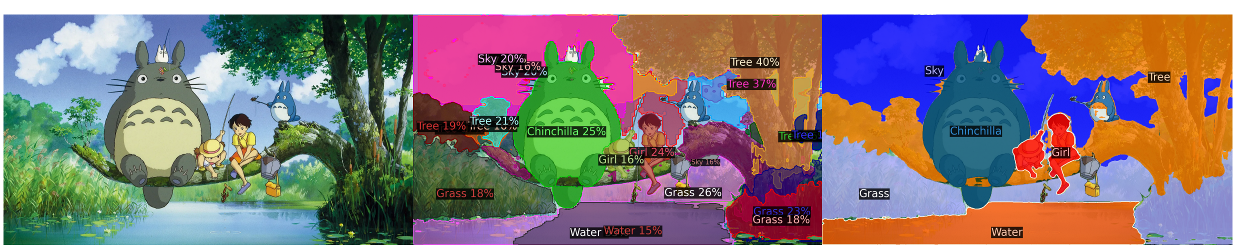
Finally, we visualize some examples on SegInW [56] in Fig. 7.

Table 9. The relevant information of different models including the backbone, base detector, text encoder, and image size.

Method	Backbone	Base Model	Text Encoder	Image Size	
				Short	Long
MDETR [16]	ENB5 (30M)	DETR	RoBERTa	480 ∼ 800	1333
GLIP [21]	Swin-L (197M)	DyHead	BERT	480 ∼ 800	1333
GLIPv2 [51]	CoSwin-H (637M)	DyHead	CLIP	480 ∼ 800	1333
UNINEXT [48]	ViT-H (632M)	DINO	BERT	320 ∼ 800	1333
G-DINO [27]	Swin-L (197M)	DINO	BERT	480 ∼ 800	1333
X-Decoder [56]	DaViT-L (196M)	Mask2Former	UniCL	224, 1024	224, 1024
OpenSeeD [52]	Swin-L (197M)	MaskDINO	UniCL	1024	1024
SEEM [57]	DaViT-L (196M)	X-Decoder	UniCL	800	1333
HIPIE [43]	ViT-H (637M)	UNINEXT	BERT	800 ∼ 1024	1333
ODISE [47]	UNet (860M)	Mask2Former	CLIP	1024	1024
APE-Ti	ViT-Ti (6M)	DETA	CLIP	1024	1024
APE-L (A)	ViT-L (307M)	DETA	CLIP	1024	1024
APE-L (B)	ViT-L (307M)	DETA	CLIP	1024	1024
APE-L (C)	ViT-L (307M)	DETA	CLIP	1024	1024
APE-L (D)	ViT-L (307M)	DETA	CLIP	1024	1024

Table 10. A detailed list of training data for different models. O365: Objects365. OID: OpenImages Detection. VG: Visual Genome. INB: ImageNet Boxes. RefC: RefCOCO/+/g.

Method	Stage	Train Data (Group by annotation types)		Batch Size	Image Consumption
		Instance-level	Image-level		#Epoch × #Image or Batch Size × #Iteration
MDETR [16]	I	COCO, RefC, VG, GQA, Flickr30k	–	64	52M (40 Ep × 1.3M Img)
GLIP [21]	I	O365, OID, VG, INB, COCO, RefC, VG, GQA, Flickr30k	Cap24M	64	64M (64 Bs × 1M Iter)
GLIPv2 [51]	I	O365, OID, VG, INB, COCO, RefC, VG, GQA, Flickr30k	Cap16M	64	64M (64 Bs × 1M Iter)
	II	COCO, LVIS, PhraseCut		64	5.36M (24 Ep × 0.2M Img + 8 Ep ×0.07M Img)
UNINEXT [48]	I	Objects365	–	64	21.8M (64 Bs × 340741 Iter)
	II	COCO, RefC		32	2.9M (32 Bs × 91990 Iter)
	III	COCO, RefC, SOT&VOS, MOT&VIS, R-VOS		32	5.7M (32 Bs × 180000 Iter)
G-DINO [27]	I	COCO, O365, OID, RefC, Flickr30k, VG	Cap4M	64	–
X-Decoder [56]	I	COCO, RefC	Cap4M	32, 1024	200M (50 Ep × 4M Img)
OpenSeeD [52]	I	COCO, O365	–	32, 64	48M (30 Ep × 1.8M Img)
SEEM [57]	I	COCO, LVIS, RefC	–	–	–
APE-Ti	I	COCO, LVIS, O365, OID, VG, RefC, SA-1B, GQA, PhraseCut, Flickr30k	–	64	17.28M (64 Bs × 0.27M Iter)
APE-L (A)	I	COCO, LVIS, O365, OID, VG	–	16	11.52M (16 Bs × 0.72M Iter)
APE-L (B)	I	COCO, LVIS, O365, OID, VG, RefC	–	16	17.28M (16 Bs × 1.08M Iter)
APE-L (C)	I	COCO, LVIS, O365, OID, VG, RefC, SA-1B	–	16	17.28M (16 Bs × 1.08M Iter)
APE-L (D)	I	COCO, LVIS, O365, OID, VG, RefC, SA-1B, GQA, PhraseCut, Flickr30k	–	64	17.28M (64 Bs × 0.27M Iter)



(a) Original Image. (b) Instance Segmentation. (c) Semantic Segmentation.

Figure 3. Visualizations of model outputs for instance and semantic segmentation tasks. All results are inferred in a single forward with prompts of {“Sky”, “Water”, “Tree”, “Chinchilla”, “Grass”, “Girl”}.

Table 11. Training data configures. SR denotes the sampling ratio, and FL denotes federated loss.

Dataset	SR	FL	Loss Weights							
			Encoder			Decoder				
			$\mathcal{L}_{\text{class}}$	$\mathcal{L}_{\text{bbox}}$	$\mathcal{L}_{\text{giou}}$	$\mathcal{L}_{\text{class}}$	$\mathcal{L}_{\text{bbox}}$	$\mathcal{L}_{\text{giou}}$	$\mathcal{L}_{\text{mask}}$	$\mathcal{L}_{\text{dice}}$
LVIS	1.0	✓	1	5	2	1	5	2	5	5
COCO Instance	1.0		1	5	2	1	5	2	5	5
COCO Stuff	1.0		1	5	2	1	5	2	5	5
Objects365	1.0		1	5	2	1	5	2	5	5
OpenImages	1.0	✓	1	5	2	1	5	2	5	5
Visual Genome	1.0		0	0	0	1	0	0	0	0
SA-1B	1.0		1	5	2	0	5	2	5	5
RefCOCO/+/g	0.1		0	5	2	1	5	2	5	5
GQA	0.1		0	0	0	1	0	0	0	0
Flickr30K	0.1		0	0	0	1	0	0	0	0
PhraseCut	0.1		0	0	0	1	0	0	0	0

Table 12. One suit of weights for visual grounding on RefCOCO/+/g. "∅" indicates that the task is beyond the model capability. "–" indicates that the work does not have a reported number.

Method	Backbone	RefCOCO						RefCOCO+						RefCOCO					
		val		testA		testB		val		testA		testB		umd-val		umd-test		google-val	
		P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU
MDETR [16]	ENB5	73.4	∅	–	∅	–	∅	58.8	∅	–	∅	–	∅	57.1	∅	–	∅	–	∅
GLIP [21]	Swin-T	50.4	∅	54.3	∅	43.8	∅	49.5	∅	52.7	∅	44.5	∅	66.0	∅	66.8	∅	–	∅
G-DINO [27]	Swin-T	73.9	∅	74.8	∅	59.2	∅	66.8	∅	69.9	∅	56.0	∅	71.0	∅	72.0	∅	–	∅
KOSMOS-2 [33]	ViT-L	52.3	∅	57.4	∅	47.2	∅	45.4	∅	50.7	∅	42.2	∅	60.5	∅	61.6	∅	–	∅
APE-Ti	ViT-Ti	72.7	57.1	79.7	64.1	65.2	50.8	66.7	51.2	71.9	57.0	54.8	41.7	67.3	52.3	65.8	50.0	66.1	50.7
APE-L (A)	ViT-L	34.2	25.1	34.8	28.0	36.1	25.7	33.5	26.3	32.3	26.6	36.0	26.0	38.9	28.1	40.5	28.3	39.4	28.4
APE-L (B)	ViT-L	83.3	70.2	88.4	76.0	77.7	63.9	74.0	59.4	82.0	67.6	62.9	47.8	79.9	62.8	79.9	62.8	**80.5**	**64.3**
APE-L (C)	ViT-L	79.8	66.3	86.8	74.0	76.2	61.8	72.2	56.6	78.4	64.1	60.9	45.6	79.8	63.2	79.5	61.2	79.5	62.6
APE-L (D)	ViT-L	**84.6**	**72.3**	**89.2**	**77.7**	**80.9**	**68.4**	**76.4**	**61.9**	**82.4**	**68.0**	**66.5**	**51.2**	**80.0**	**64.2**	**80.1**	**63.2**	79.9	63.3

(a) “a sofa with no pillow on it in the room”

(b) “aircraft in the air”

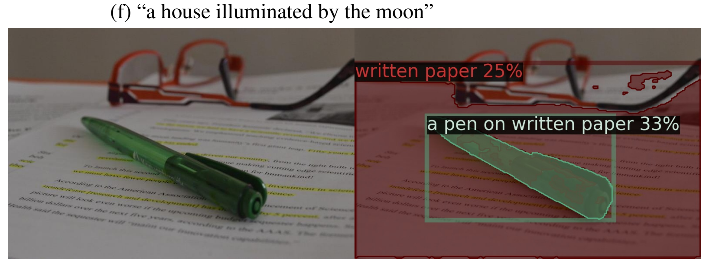
(c) “christmas tree full of decorations”, “a person in santa claus clothes without bags”

(d) “aircraft not on the ground”

(e) “a house illuminated by the moon”

(f) “a house illuminated by the moon”

(g) “a knife being used to cut vegetables”



(h) “written paper”, “a pen on written paper”

(i) “chess piece of horse head”

(j) “peacock standing on the grass”

(k) “donut with colored granules on the surface”

Figure 4. Visualizations of model outputs on D3 [46]. In each group, the **left** image is the original image and the **right** image shows the predictions, and corresponding prompts of predicted objects are listed in the **subcaption**. All results are inferred in a single forward with all provide prompts.

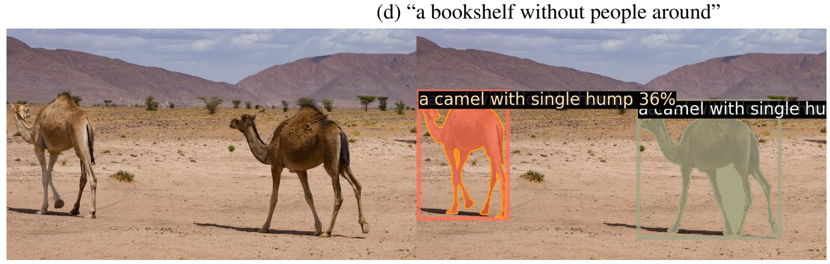
(a) “a plush toy”

(b) “a plane flying to the right”

(c) “a child wearing a mask”

(d) “a bookshelf without people around”

(e) “a bed with patterns in the room”, “the lamp on the table beside the bed”

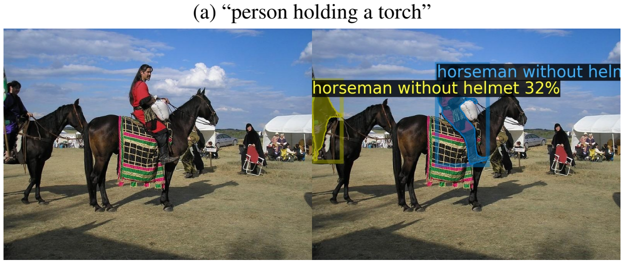


(f) “a camel with single hump”

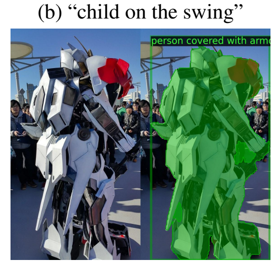
Figure 5. Visualizations of model outputs on D3 [46]. APE is capable to predict multiple instances for one sentence prompts. In each group, the **left** image is the original image and the **right** image shows the predictions, and corresponding prompts of predicted objects are listed in the **subcaption**. All results are inferred in a single forward with all provide prompts.

(a) “person holding a torch”

(b) “child on the swing”



(c) “horseman without helmet”



(d) “person covered with armor”

(e) “a person with golf clubs”

(f) “player with basketball in the hand”, “basketball in hand”

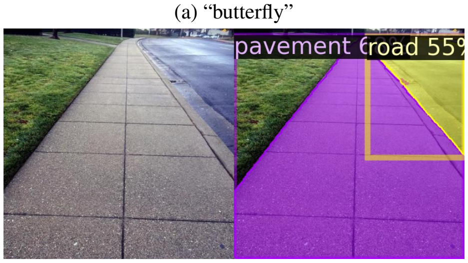
(g) “bartender without suit”

(h) “car contacted by an auto-salon girl”, “an auto-salon girl without bare waist”

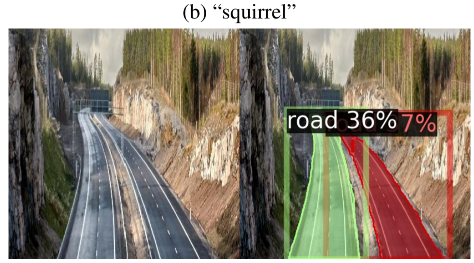
Figure 6. Visualizations of model outputs on D3 [46] for Human-centric grounding. In each group, the **left** image is the original image and the **right** image shows the predictions, and corresponding prompts of predicted objects are listed in the **subcaption**. All results are inferred in a single forward with all provide prompts.

(a) “butterfly”

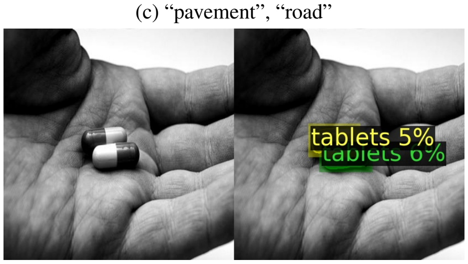
(b) “squirrel”



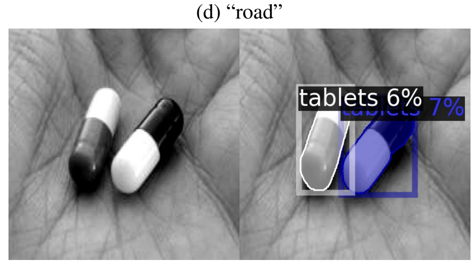
(c) “pavement”, “road”



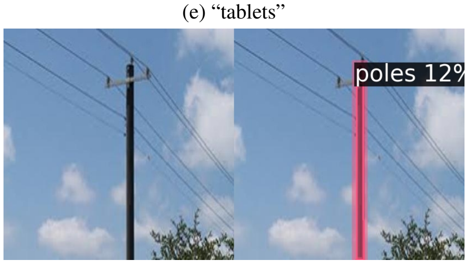
(d) “road”



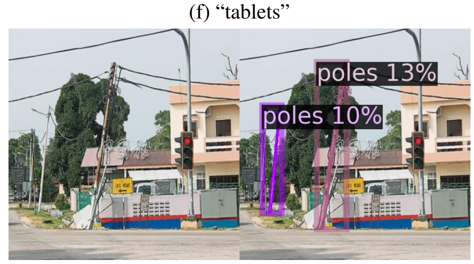
(e) “tablets”



(f) “tablets”



(g) “poles”



(h) “poles”

Figure 7. Visualizations of model outputs on SegInW [56]. In each group, the **left** image is the original image and the **right** image shows the predictions, and corresponding prompts of predicted objects are listed in the **subcaption**. All results are inferred in a single forward with all provide prompts.