A. Architecture Details of the Image Encoder

As shown in Fig. 3, our Uni-Perceiver v2 consists of three main parts: the image encoder, the text encoder, and the unified decoder. In this section, we describe the architecture details of the image encoder.

Backbone Network. Given an input image $x \in \mathbb{R}^{H\times W}$ with height H and width W, a backbone network (*e.g.,* ResNet [11], Swin-Transformer [21]) is firstly employed to extract the multi-scale feature maps $\{\mathcal{F}_l\}_{l=0}^{L-1}$, where $L = 4$ is the number of features scales, and the spatial shapes of the feature maps are $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$, and $\frac{H}{32}\times\frac{W}{32}$. The feature maps are transformed by 1×1 convolutions to match the hidden dimension of the following Transformer-based region proposal network. The transformed feature maps are denoted as $\mathcal{F}'_l$. An additional 3×3 stride 2 convolution layer is applied on $\mathcal{F}_3$ to extract a smaller feature map $\mathcal{F}'_4 \in \mathbb{R}^{\frac{H}{64}\times\frac{W}{64}\times d}$. $d = 256$ is the hidden dimension of the Transformer. In addition, the number of global attention pooling to extract global features is set to $M' = 10$.

Region Proposal Network. The region proposal network is similar to MaskDINO [16], but only considers foreground-background binary classification. A Transformer-based region proposal network is applied on top of the multi-scale feature maps to generate regional representations. Specifically, in the 4-scale setting which is adopted by Uni-Perceiver v2, the input of the Transformer encoder is the backbone feature maps except the first scale $\{\mathcal{F}'_l\}_{l=1}^{L=4}$. A deformable Transformer [60] encoder is employed to extract multi-scale encoded features $\{\mathcal{F}_l^{\text{enc}}\}_{l=1}^{L=4}$ whose spatial shapes and dimensions are the same as the corresponding input features. To generate the region proposals, we apply a deformable Transformer decoder on the multi-scale encoded features. To construct the N input object queries of the Transformer decoder (*e.g.,* $N = 900$), we predict the objectness score and bounding box of each feature pixel in the encoded feature maps $\{\mathcal{F}_l^{\text{enc}}\}_{l=1}^{L=4}$, and select top-$N$ features based on their objectness scores. The selected features are added to N randomly initialized object queries as the input of the Transformer decoder, and their locations serve as the initial guess of the bounding boxes of the region proposals.

The Transformer decoder generates a set of N candidate object proposals $\{q_j^{\text{sem}}, q_j^{\text{box}}, q_j^{\text{mask}}\}_{j=1}^{N}$, where $q_j^{\text{sem}} \in \mathbb{R}^d$, $q_j^{\text{box}} \in \mathbb{R}^4$, and $q_j^{\text{mask}} \in \mathbb{R}^{H\times W}$ are the semantic, bounding box, and segmentation mask representations of the j-th proposal, respectively. Following Mask2Former [52] and MaskDINO [16], the segmentation mask representations are obtained by the dot product of the final-layer hidden state of the j-th proposal q_j and a per-pixel feature map,

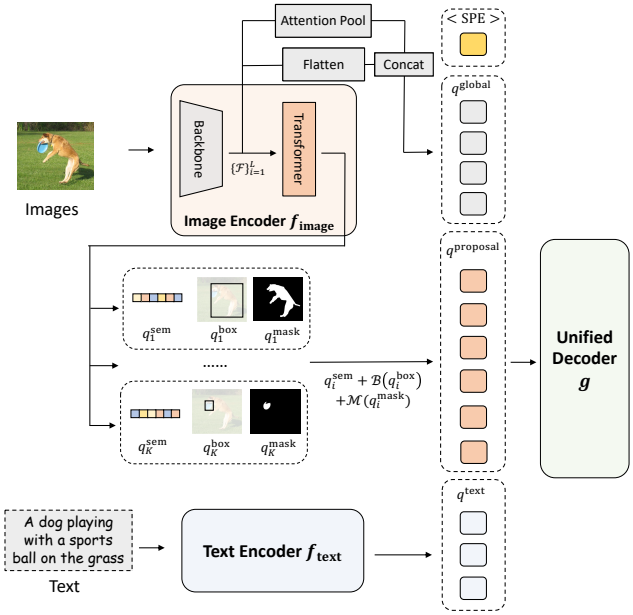

Figure 3. Architecture overview of our Uni-Perceiver v2.

$$q_i^{\text{mask}} = \text{Upsample}\Big(\text{MLP}(q_i) \odot \mathcal{R}\big(\mathcal{G}(\mathcal{F}_0) + \mathcal{H}(\mathcal{F}_1^{\text{enc}})\big)\Big), \quad (8)$$

where $\mathcal{G}$ is a 1×1 convolution layer followed by a Group Normalization (GN) [59], $\mathcal{H}$ is a 1×1 convolution followed by a GN and a bilinear upsampling, and $\mathcal{R}$ is a 3×3 convolution followed by a GN, a ReLU, and a 1×1 convolution.

The regional representations are obtained by fusing the semantic, bounding box, and segmentation mask representations,

$$q_j^{\text{proposal}} = q_j^{\text{sem}} + \mathcal{B}(q_j^{\text{box}}) + \mathcal{M}(q_j^{\text{mask}}), \quad (9)$$

where $\mathcal{B}$ denotes the positional encoding of box coordinates. $\mathcal{M}$ uses adaptive average pooling to scale the mask predictions to the size of 28×28. Both $\mathcal{B}$ and $\mathcal{M}$ are followed by linear projections to match the feature dimension. Note that the bounding box and segmentation mask representations are detached before fusing.

To reduce the computational cost, we predict objectness score for each proposal q_j^{proposal}, and select the top-O proposals as the final regional representations. O is set as 200 by default in Uni-Perceiver v2.

Loss Function. In non-localization tasks such as image classification, the supervision is applied only on the final predictions of the unified decoder as Eq. 7, and there is no special supervision for the proposal generation of the image encoder. In localization tasks such as object detection, additional supervisions are applied for the training of the region proposal network. Specifically, we adopt the contrastive query denoising in MaskDINO [16] for the training of the Transformer decoder. For better convergence of

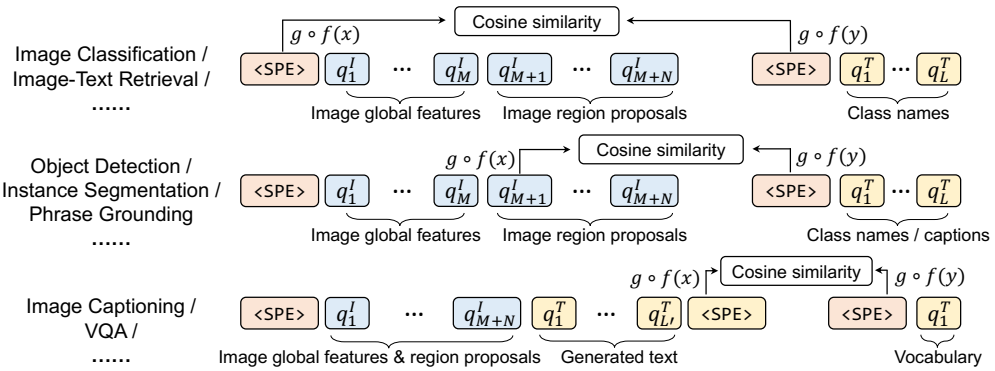

Figure 4. Illustration of the input and prediction formats of our decoder. The input form depends on whether visual-linguistic interaction is required. The token used for prediction depends on whether the required representation is global or local.

the region proposal network, we predict objectness score, bounding box, and segmentation mask for each proposal at the outputs of Transformer encoder and each Transformer decoder layer, and detection losses with binary classification (*i.e.,* predicting the objectness instead of classes) are applied to each output as an intermediate supervision.

B. Implementation Details

Region Proposal Network. The hyper-parameters used in our region proposal network are listed in Tab. 7. These values mainly follow MaskDINO [16], but with small modifications. The number of candidate object proposals ("num_queries" in Tab. 7) used to generate regional representations is 300 and 900 for the ResNet-50 backbone and Swin backbones, respectively. To reduce the computation cost of the unified decoder, the region proposals are filtered depending on their objectness scores and only $O = 200$ region representations are selected as the input for the unified decoder ("topk_queries" in Tab. 7). Moreover, to save computation cost, the point loss used in Mask2Former [52] is adopted to calculate mask loss, where the number of sampled points is 112×112.

Unified Decoder. The Transformer-based unified decoder is used for general task modeling, whose detailed input and prediction formats are shown in Fig. 4. During training, a uniform drop rate for stochastic depth is used across all layers and the value is set to 0.1. Unlike Uni-Perceiver series [49, 50], the layer-scale technique [57] is not enabled since the instability phenomenon is not observed when the training of the 6-layer unified decoder. In addition, when Conditional MoE is employed in the unified decoder, the number of experts in each layer is set to 8.

Data augmentation. For all tasks except image detection and segmentation, we apply the data augmentation techniques that are similar to Uni-Perceiver [50]. However, image resolution is set to 384×384 and 224×224 for Swin backbones and for ResNet-50 backbone, respectively. And for object detection and instance segmentation tasks, we first randomly resize the input image with its shorter side between 200 and 1800 pixels and its longer side at most 2400. Then we crop the image to a fixed size of 1600×1600 during training. For evaluation, the shorter side is set to 1400, and the maximum longer side is set to 1600.

Others. Tab. 8 lists the batch size, sampling weight s_k, and scaling factor ω_k for each task and dataset in the joint training.

Item	Value
enc_layers	6
dec_layers	6
dim_feedforward	2048
hidden_dim	256
dropout	0.0
nheads	8
num_queries	300/900
topk_queries	200
enc_n_points	4
dec_n_points	4
cls_cost_coef	2.0
bbox_cost_coef	5.0
giou_cost_coef	2.0
mask_cost_coef	5.0
dice_cost_coef	5.0
cls_loss_coef	2.0
bbox_loss_coef	5.0
giou_loss_coef	2.0
mask_loss_coef	5.0
dice_loss_coef	5.0
dn_box_noise_scale	1.0
dn_label_noise_ratio	0.5

Table 7. Hyper-parameters used in our region proposal network.

task	dataset	#data	batch size / GPU	sampling weight s_k	scaling factor ω_k
Image Classification	ImageNet-1k [8]	1.28M	28	0.1	1.0
Object Detection & Instance Segmentation	COCO [19]	118K	1	0.25	1.0
Masked Language Modeling	Books&Wiki [51]	-	256	0.05	0.5
Image Captioning	YFCC [13]	14.8M	24	0.09831	0.16385
	CC12M [4]	11.1M	24	0.08514	0.1419
	CC3M [33]	3M	24	0.04428	0.0738
	Visual Genome [15]	108K	24	0.02973	0.04955
	COCO Caption [7]	113K	24	0.0192	0.032
	SBU [25]	830K	24	0.02328	0.0388
	sum	29.9M	-	0.3	0.5
Image-Text Retrieval	YFCC [13]	14.8M	28	0.09831	0.3277
	CC12M [4]	11.1M	28	0.08514	0.2838
	CC3M [33]	3M	28	0.04428	0.1476
	Visual Genome [15]	108K	28	0.02973	0.0991
	COCO Caption [7]	113K	28	0.0192	0.064
	SBU [25]	830K	28	0.02328	0.0776
	sum	29.9M	-	0.3	1.0

Table 8. Tasks and datasets used for our joint training. "#data" is the amount of visual training samples. For image captioning and image-text retrieval tasks, a combination of image-text-pair datasets is used for training, which has about 29.9M visual samples after filtering the data overlapping with validation sets. To alleviate the data imbalance problem in the combination of image-text-pair datasets during multi-task training, sampling weight s_k and scaling factor ω_k for each dataset is set to be proportional to the square root of the dataset size, which has demonstrated to be effective [49].

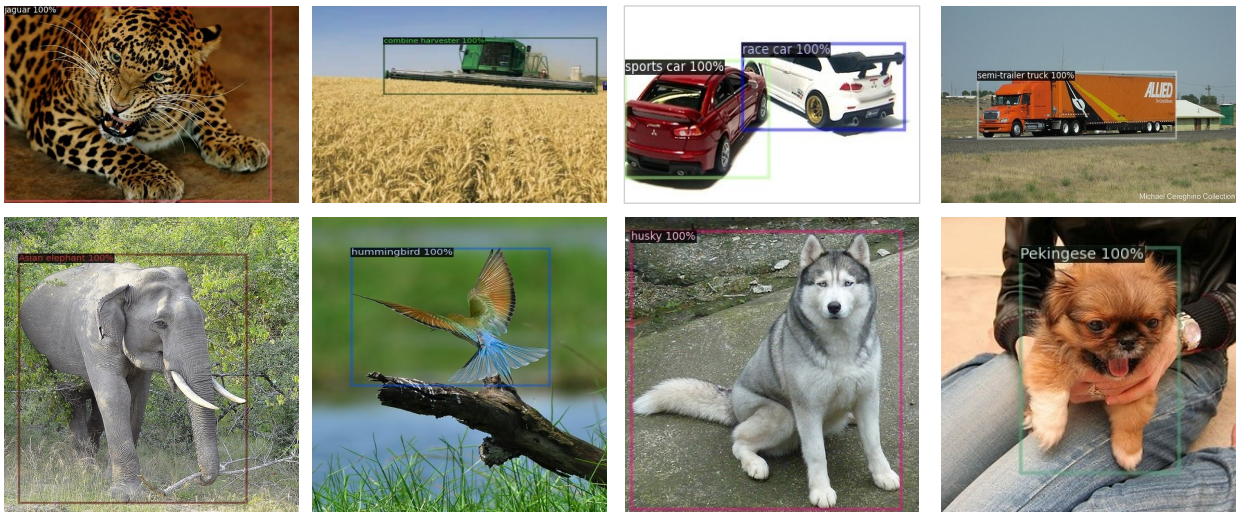

Figure 5. Detection results on novel categories. We show the detection results of images from ImageNet-1k validation set. Note that Uni-Perceiver v2 only uses COCO dataset for the training of image detection task, where most classes in ImageNet-1k are not seen.

C. Detection on Novel Categories

Thanks to the general task modeling of Uni-Perceiver v2, different tasks can borrow knowledge from each other. For example, object detection task can generalize to novel categories in image classification dataset. Fig. 5 shows the detection result of Uni-Perceiver v2 on images from ImageNet-1k validation set whose categories do not exist in COCO dataset. This demonstrates the generalization ability of Uni-Perceiver v2, indicating the benefit of general task modeling.

D. Licenses of Datasets

ImageNet-1k [8] is subject to the ImageNet terms of use [58].

COCO [19] The images are subject to the Flickr terms of use [53].

BookCorpus [51] Replicate Toronto BookCorpus is open-source and licensed under GNU GPL, Version 3.

Wikipedia Most of Wikipedia's text is co-licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) and the GNU Free Documen-

tation License (GFDL) (unversioned, with no invariant sections, front-cover texts, or back-cover texts). Some text has been imported only under CC BY-SA and CC BY-SA-compatible license and cannot be reused under GFDL.
YFCC [13] All the photos and videos provided in YFCC dataset are licensed under one of the Creative Commons copyright licenses.
CC12M [4] is licensed under the Terms of Use of Conceptual 12M [55].
CC3M [33] is licensed under the Conceptual Captions Terms of Use [56].
Visual Genome [15] is licensed under a Creative Commons Attribution 4.0 International License [54].
COCO Captions [7] The images are subject to the Flickr terms of use [53].
SBU Caption [25] The images are subject to the Flickr terms of use [53]

Appendix References

[52] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[53] Inc. Flickr. Flickr terms & conditions of use. https://www.flickr.com/help/terms.

[54] Ranjay Krishna. Visual genome terms & conditions of use. https://visualgenome.org/about.

[55] Google LLC. Conceptual 12m terms & conditions of use. https://github.com/google-research-datasets/conceptual-12m/blob/main/LICENSE.

[56] Google LLC. Conceptual captions terms & conditions of use. https://github.com/google-research-datasets/conceptual-captions/blob/master/LICENSE.

[57] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[58] Princeton University and Stanford University. Imagenet terms & conditions of use. https://image-net.org/download.

[59] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.