
What can Foundation Models’ Embeddings do?

Anonymous Author(s)

Affiliation

Address

email

1 Implementation Details

Prompt details for Data Engine. In the main paper, Table.1, we miss the part about generated caption constraints because of the page limit. Here we give the constraints:

- rely more to ground truth image captions, rely less on pseudo description..
- the entity in the generated caption should be in entity_proposal.
- the generated entity nouns could rely on object_descriptions in ground truth image captions.
- for [entity_id] and <xxx>, the content in <xxx> is better to include the adj description.
- the generation caption should have less hallucinate description in image description.
- the generated caption should be as short as possible.

Loss Details. The training tasks include panoptic segmentation, interactive segmentation, grounded segmentation, image-text retrieval, interleave retrieval with visual entities from the same image, and interleave grounding. Losses are defined below.

$$\begin{aligned} \mathcal{L} = & \alpha_p \mathcal{L}_{\text{CE_pano}} + \beta_p \mathcal{L}_{\text{BCE_pano}} + \gamma_p \mathcal{L}_{\text{DICE_pano}} + \alpha_g \mathcal{L}_{\text{CE_grd}} + \beta_g \mathcal{L}_{\text{BCE_grd}} + \gamma_g \mathcal{L}_{\text{DICE_grd}} \\ & + \alpha_i \mathcal{L}_{\text{CE_iseg}} + \beta_i \mathcal{L}_{\text{BCE_iseg}} + \gamma_i \mathcal{L}_{\text{DICE_iseg}} + \theta \mathcal{L}_{\text{VLC_imgtexr}} + \phi \mathcal{L}_{\text{IC_intr}} + \alpha_{ig} \mathcal{L}_{\text{CE_intg}} \\ & + \beta_{ig} \mathcal{L}_{\text{DICE_intg}} + \gamma_{ig} \mathcal{L}_{\text{ICE_intg}} \end{aligned} \quad (1)$$

Where $\alpha_p =, \beta_p =, \gamma_p =, \alpha_g =, \beta_g =, \gamma_g =, \alpha_i =, \beta_i =, \gamma_i =, \theta =, \phi =, \alpha_{ig} =, \beta_{ig} =, \gamma_{ig} =$. CE, BCE, DICE, VLC, IC, ICE denotes cross-entropy, binary cross entropy, dice loss, vision-language contrastive loss, interleave contrastive loss, interleave cross entropy loss, respectively.

In the next section, we will focus on experiments to prove the effectiveness of the proposed dataset and model.

Training Details. All the numbers reported in Table.2, Table.3, Table.5 are trained with the X-Decoder vision backbone. Following (3), our tiny, base, and large model are using Focal-T, Davit-d3, and Davit-d5 models respectively. The training tasks include panoptic segmentation, interactive segmentation, grounded segmentation, image-text retrieval, interleave retrieval with visual entities from the same image, and interleave grounding. We fixed the vision and language foundation model, and the interface contains 9 layers. To alleviate the conflict of different tasks, we only compute loss on the last 6 layers of the interface for all tasks. During training, the standard pipeline uses a batch size of 192 with a training resolution of 640×640 . However, in order to compare with higher resolution segmentation results or image-text retrieval results training with larger batch size, we use the resolution of 1024×1024 with batch size 192 in comparison with segmentation baselines, and the resolution of 384×384 with batch size 384 to compare with retrieval baselines in Table.2.

Computation Resource. Our largest model is trained on 16 V100s in 24 hours.

30 2 Baselines

31 **Interleave Retrieval** In order to evaluate the interleave image text retrieval results of ImageBind,
32 BLIP2, and X-Decoder, we tried multiple implementation approaches and eventually found the best
33 implementation as follows. Given a search query $Q = \{q_0, q_1, q_2, \dots\}$ as a sequence, where q_i is one
34 of image entity, text entity. In addition, we denote Q' as the query that we have replaced the image
35 entity to [INTERACTIVE] for a placeholder. And we denote the foundation model as B , and the
36 image as I . The final score is calculated as the following:

$$FI_i = B(I_i) \quad (2)$$

$$FQ_i = B(q_0) + B(q_1) + \dots + B(q_n) \quad (3)$$

$$FQ'_i = B(Q'_i) \quad (4)$$

$$S = FQ \times FI^T + FQ' \times FI^T \quad (5)$$

37 where FI denotes image features, FQ denotes query features, and FQ' denotes sentence features. We
38 forward image features, entity features, and sentence features separately and compute the similarity
39 within the dataset for interleave retrieval.

40 **Interleave Segmentation** We use SEEM as our interleave segmentation baseline as it is able to do
41 both interactive segmentation and grounded segmentation. Given a search query $Q = \{q_0, q_1, q_2, \dots\}$,
42 for each q_i that is either an image or text entity, we compute the interactive segmentation and grounded
43 segmentation result separately. As shown in the main paper Table 4, the proposed FIND method
44 achieved around 8 points better than SEEM on cIoU, which shows the privilege of the unified pipeline.

45 3 Benchmark

46 There are recent works that use LLMs or foundation models as augmentation engines to generate
47 high-quality datasets with new task-specific annotations. For example, LLaVA (1) generates a dataset
48 with question-answering entries for visual instruction tuning, MMC4 (2) creates a dataset with
49 interleaved image-text paragraphs. The recent work ContextDET leverages Flickr30K to generate a
50 question-answering dataset, and GLaMM also generates a grounded-QA dataset with a very heavy
51 pipeline.

52 **Dataset Statistics** As shown in the table below, we compare our proposed benchmark with the
53 recent benchmark that uses pseudo-data to create a new set of ground truths.

	Task	Engine	Source	#data	Box	Mask	GT
LLaVA	Visual Question Answering	GPT4	COCO	80K	-	-	-
MMC4	Interleave Paragraph	CLIP	Web	571M	-	-	-
ContextDET	Grounded Question Answering	-	Flicker30k	30K	✓	✗	✓
GLaMM	Grounded Caption	GPT4+Detectors	SAM	11M	✓	✓	✗
FIND	Grounded Caption	GPT4	COCO	110K	✓	✓	✓

54 As shown in the table above, LLaVA and MMC4 only contain image text data either in question
55 answering or interleave paragraph format. ContextDET takes advantage of the Flickr30k dataset
56 to obtain grounded captions and question-answering pairs. However, constraints by the ground
57 truth annotation type of Flickr 30k, ContextDET only contains bounding box annotations, and the
58 annotated instances are mostly constrained to several categories. **GLaMM is a concurrent work**
59 with us, in which they annotated the ground truth bounding box and mask using SoTA detectors.
60 In addition, their pipeline is very complicated, where we simply use GPT4 as an augmentation for
61 ground truth and generated contents.

62 References

63 [1] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485
64 (2023)

- 65 [2] Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang,
66 W.Y., Choi, Y.: Multimodal c4: An open, billion-scale corpus of images interleaved with text.
67 arXiv preprint arXiv:2304.06939 (2023)
- 68 [3] Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L.,
69 et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF
70 Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)