Supplementary material for Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models

Anonymous Author(s)
Affiliation
Address
email

1 Code

We provide the code in the *DAC_code* sub-folder of this Supplementary zip (includes a README file with usage instructions).

2 Detailed LLM Prompt

The detailed full prompt we used for the LLM-based caption density expansion as part of our DAC-LLM is as detailed below. It provides a list of in-context examples of what we want the LLM to do followed by the:

```
please describe what you might see in a picture of a scene that contains
'{short_description}', write each sentence in a list, and use complete
sentences with all nouns and objects you are referring to
```

sentence where *short_description* is the placeholder for the caption we would like to expand.

```
input_text = f"short: please describe what you might see in a picture of a
scene that contains 'a Christmas tree', write each sentence in a list, and
use complete sentences with all nouns and objects you are referring to\n \
        long:  1  In the center of the room, a majestic evergreen Christmas
tree stands tall, adorned with twinkling lights and colorful ornaments.\n \
   2  Delicate strands of tinsel gracefully drape the tree's branches,
adding a touch of shimmer to the festive display.\n \
   3  An elegant star or angel graces the top of the tree, representing the
Star of Bethlehem or the heavenly messengers present at Jesus' birth.\n \
   4  Wrapped presents in various shapes and sizes are piled beneath the
tree, their festive gift wrap and bows hinting at the surprises inside.\n \
   5  A cozy fireplace crackles nearby, with stockings hung from the mantel,
eagerly awaiting the arrival of Santa Claus.\n \
   6  Lush green garlands and flickering candles decorate the mantel,
enhancing the holiday atmosphere.\n \
   7  Comfortable seating arrangements, such as sofas and armchairs, are
positioned near the tree, complete with plush cushions and warm throw
blankets.\n \
   8  Family members and friends gather around the tree in festive attire,
sharing laughter and conversation.\n \
   9  A beautifully crafted wreath hangs on a nearby wall or window, adding
an additional touch of holiday cheer.\n \
```

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

```
   10 Through the window, a snowy winter landscape can be seen, with
snow-covered trees, rooftops, and gently falling snowflakes, creating the
perfect backdrop for the Christmas scene.\n \
        short: please describe what you might see in a picture of a scene
that contains 'a male hand playing nervously with a pencil on a black
background', write each sentence in a list, and use complete sentences with
all nouns and objects you are referring to \n \
        long:   1  A male hand is positioned prominently in the frame, with
fingers flexing and shifting as they manipulate a pencil.\n \
   2  The pencil, held between the thumb and index finger, twirls and spins
as the hand moves it nervously.\n \
   3  Shadows from the hand and pencil cast dramatic patterns on the stark
black background, emphasizing the sense of tension and unease.\n \
   4  Flecks of graphite from the pencil's tip may be visible, scattered
across the black surface, as a result of the restless movements.\n \
   5  The hand's knuckles and veins are accentuated by the lighting,
highlighting the pressure and force exerted during the fidgeting.\n \
   6  The pencil's eraser end, worn and discolored, suggests frequent use
and a history of anxious behavior.\n \
   7  A hint of perspiration on the hand's skin glistens under the light,
further revealing the nervous energy within the scene.\n \
   8  The positioning of the hand, perhaps slightly off-center or at an
angle, contributes to the visual tension of the composition.\n \
   9  Fingernails on the hand may appear bitten or worn, indicating a habit
of nervousness and stress.\n \
   10 The black background contrasts sharply with the hand and pencil,
isolating them in the scene and focusing the viewer's attention on the
restless, uneasy motion.\n \
        short: please describe what you might see in a picture of a scene
that contains 'a man is programming', write each sentence in a list, and
use complete sentences with all nouns and objects you are referring to \n \
        long:   1  A focused man sits at a desk, his eyes intently scanning
the computer screen in front of him as he works on a programming project.\n
\
   2  The computer display is filled with lines of code, featuring various
colors and syntax highlighting to differentiate between elements of the
programming language.\n \
   3  The man's fingers move swiftly and confidently across the keyboard,
typing commands and adjusting the code as needed.\n \
   4  Beside the keyboard, a mouse and a notepad with handwritten notes or
algorithms offer additional tools for the programmer's work.\n \
   5  A cup of coffee or tea sits nearby, providing the man with a source
of caffeine to maintain his focus and energy.\n \
   6  The room's lighting, either from a desk lamp or overhead lights,
illuminates the workspace, creating a comfortable environment for
concentration.\n \
   7  The man wears casual attire, such as a t-shirt and jeans, reflecting
the informal nature of the programming process.\n \
   8  Reference books or technical manuals may be stacked or spread out on
the desk, offering guidance and information for the programmer.\n \
   9  The man's facial expression, furrowed brows or a slight frown,
conveys his deep concentration and determination to solve the coding
challenge at hand.\n \
   10 Surrounding the man, other electronic devices, like a smartphone or
tablet, may be present, indicating the interconnected nature of his work in
the digital realm.\n \
        short: please describe what you might see in a picture of a scene
that contains '{short_description}', write each sentence in a list, and use
complete sentences with all nouns and objects you are referring to \n \
```

```
        long: "
```

3 Evaluating additional VL models on VL-Checklist

To further illustrate the difficulty of VL models with compositional reasoning, we have extended the evaluation appearing in Table 1 of the main paper with more baseline VL methods. We include also the more recent Meter [1], X-VLM [2], and VLMO [3], as well as BLIP [4] (previous version of BLIP2 [5], exhibiting very similar performance). In Table 1, we report the average of the main VL-Checklist [6] Attribute and Relation metrics for all methods. The results in the table once again demonstrate the difficulty the existing VL methods have with compositional reasoning, and the need for improvement such as offered by our DAC approach.

Method	VL-Checklist [6] Attribute and Relations metrics Average
Meter [1]	55.82%
X-VLM [2]	58.65%
VLMO [3]	54.6%
CLIP [7]	65.47%
NegCLIP [8]	67.87%
BLIP [4]	76.13%
BLIP2 [5]	75.42%
SVLC (CLIP) [9]	70.46%
DAC-LLM(Ours)	81.84%
DAC-SAM(Ours)	**82.79%**

Table 1: Comparison of the results of our method to multiple SoTA VL baselines averaging the VL Checklist “Attribute" and “Relation" performance metrics.

4 Detailed LP evaluation on the ELEVATER [10]

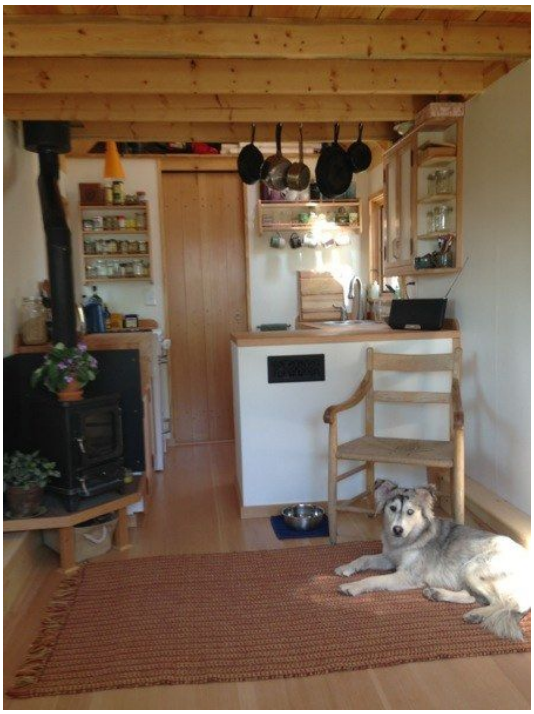
As promised in the main paper, we include the detailed breakdown of LP evaluation results on the ELEVATER [10] benchmark (using the standard [10] evaluation protocol) in Table 2. Each individual dataset results as well as averages over the datasets are reported.

Model	Avg.	Caltech101	CIFAR10	CIFAR100	Country211	DTD	EuroSat	FER2013	FGVCAircraft	Food101	GTSRB	HatefulMemes	KittiDistance	MNIST	Flowers102	OxfordPets	PatchCamelyon	SST2	RESISC45	StanfordCars	VOC2007
5-Shot Linear Probe																					
CLIP	**66.19**	**90.2**	90.2	66.1	**16.3**	61.4	71.6	44.4	**28.9**	**84.3**	58.6	55.8	**50.5**	**77.9**	93.0	**87.9**	54.5	58.5	79.2	**68.7**	85.0
DAC-LLM	64.92	89.7	**91.4**	68.0	13.8	**62.9**	65.3	**51.9**	28.4	81.8	60.5	**56.1**	36.8	71.1	**93.1**	87.0	50.0	**59.9**	**79.7**	64.9	86.0
DAC-SAM	64.33	89.2	90.1	**69.1**	14.2	56.6	**81.2**	51.4	26.1	77.3	**60.9**	55.5	30.6	57.9	92.9	85.2	**56.7**	62.4	76.8	64.5	**86.8**
10-Shot Linear Probe																					
CLIP	**69.58**	**92.0**	88.3	70.0	**18.7**	68.1	81.8	**53.0**	**35.4**	**84.8**	66.7	**56.7**	**52.2**	89.0	95.2	**89.5**	50.0	58.9	**82.4**	**73.6**	85.2
DAC-LLM	69.2	91.7	**89.1**	**72.2**	16.7	**69.0**	**83.5**	51.4	34.7	82.7	**68.1**	56.4	50.4	86.6	95.1	83.2	52.6	62.4	80.7	71.6	86.2
DAC-SAM	69.41	91.2	87.8	71.8	16.0	68.1	80.0	55.3	34.0	81.1	68.0	55.5	48.9	**89.4**	**95.3**	88.2	**56.7**	**62.4**	81.1	70.5	**87.0**
20-Shot Linear Probe																					
CLIP	71.9	**93.1**	**92.1**	71.7	**19.4**	72.4	83.0	54.4	**40.3**	**85.3**	75.9	55.7	49.8	90.0	95.2	**90.3**	61.0	58.8	**85.4**	**78.2**	85.8
DAC-LLM	**72.98**	92.7	92.1	**74.0**	17.4	**73.2**	87.1	**56.0**	38.1	83.8	76.9	**56.5**	54.0	**94.6**	95.1	89.1	71.3	60.0	84.5	77.0	86.8
DAC-SAM	72.92	92.0	91.4	73.6	17.0	72.3	**87.4**	51.4	39.4	82.4	**77.6**	56.4	**57.4**	93.7	**95.3**	88.5	**73.3**	**60.5**	85.1	76.2	**87.5**
Full-Shot Linear Probe																					
CLIP	78.96	93.3	94.9	80.3	**26.1**	74.2	93.7	**67.4**	45.7	**88.5**	87.8	**64.8**	66.1	98.8	95.2	**91.4**	83.3	**71.0**	87.9	**82.1**	86.3
DAC-LLM	77.44	**93.3**	93.6	**81.9**	24.7	75.0	94.5	66.8	**45.8**	88.3	88.6	63.5	66.1	98.8	**95.6**	89.7	84.2	70.6	**89.1**	81.7	87.2
DAC-SAM	**79.3**	92.6	**95.3**	81.6	24.5	**75.7**	**94.9**	66.9	44.5	87.7	**88.8**	63.7	**72.2**	**98.8**	95.5	91.1	**84.7**	70.5	88.9	81.1	**87.5**

Table 2: Detailed LP evaluation results on the ELEVATER [10] benchmark using the standard [10] LP evaluation protocol.

5 Additional qualitative examples

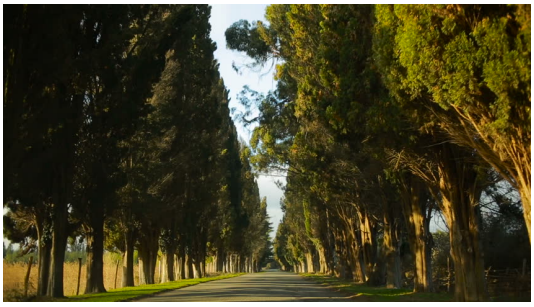
Additional qualitative examples of the automatic caption quality and caption density enhancement pipeline proposed in our DAC (Sec. 3.1 and Sec. 3.2 in the main paper) are included below.



Original caption: "see ... a big dog does just fine in a tiny house"
Quality enhanced caption (Sec. 3.1 in main paper): "a dog laying on a rug in a kitchen"
Caption density expansion using LLM (Sec. 3.2 in main paper): ["Hanging from a rack on the

wall, a variety of pots and pans are ready for use, including a non-stick frying pan and a large cooking pot.", "A dog that could be a terrier or a Pekingese rests in the kitchen corner, near a large bowl of food.", "The scene is illuminated by a fireplace."]
Caption density expansion using Segmentation (Sec. 3.2 in main paper): ["a brown rug", "a dog", "the image shows the wood grain of the wood flooring", "the image shows the floor of a room", "the image shows the front of the door", "the image shows the front of a black plastic handle", "a wooden board", "the image shows a wood stove", "a wooden floor", "the kitchen cabinet in the picture is a corner cabinet", "a wooden table"]

Original caption: "Comic book character is never alone ."
Quality enhanced caption: "Batman sitting on a throne with bats flying around him"
Caption density expansion using LLM: ["Batman is depicted as a large, menacing figure with bat-like bat ears and bat-like wings.", "The picture is filled with bat-like bat wings, resembling wings sprouting from his back and wings sprouting from his sides.", "The bat wings are highlighted by the camera and create an impression of speed and activity, perhaps even flying through the scene."]
Caption density expansion using Segmentation: ["The image shows a computer", "the image shows a black cloak", "the image shows a black hat", "the image shows a tank"]



Original caption: "driving by car along the alley." **Quality enhanced caption:** "a road lined with tall trees in the middle of a sunny day"
Caption density expansion using LLM: ["The tall, stately trees line a broad, straight road, making for a beautiful view as travelers pass through.", "The tree trunks are perfectly aligned to form a

straight line in the sunny autumn day.", "The road is shaded by tall, stately trees, creating a canopy that shades the traveler from the sun's rays.", "Diverse sets of flowers decorate the trees' foliage, adding a splash of color to the scene, while the colors of the pavement and grass blend together to create a soothing blue canvas."]
Caption density expansion using Segmentation: ["a tree", "the shadow of a tree on the ground", "a tree in the sky", "a tree trunk", "a line of trees", "a tree line", "the image shows a statue of a man with a long beard", "a long, straight road", "a green leaf", "the image shows a tree with a leaf on it", "a tree in the sky"]

Original caption: "Running back person breaks off a-yard touchdown run during the first quarter of the game"
Quality enhanced caption: "A football player is running with the ball"
Caption density expansion using LLM: ["A football player is wearing a helmet and cleats, and his feet are placed carefully on the ground.", "He sprints toward the end zone to intercept a pass, sending a football into the air and into the arms of a receiver.", "The receiver, wearing a jersey and a helmet, catches the football for his team and runs with the football for a touchdown.", "A loud roar of cheers and applause erupts from all who heard the play.", "The players celebrate the touchdown with a high-fives and high-fives."]
Caption density expansion using Segmentation:["a football jersey", "the image shows a football player running", "the image shows a person wearing a helmet", "the image shows a football player wearing a jersey with the number 96 on it", "a football player in a blue and orange uniform", "a football jersey", "a blue sky with a few clouds", "the image shows a football helmet", "the image shows a woman's back", "the image shows a helmet with the georgia bulldog logo on it", "a woman in a red dress"]

References

[1] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.

[2] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," *arXiv preprint arXiv:2111.08276*, 2021.

[3] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 897–32 912, 2022.

[4] J. Li, D. Li, C. Xiong, and S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022. eprint: `arXiv:2201.12086`.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[6] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin, "Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations," *arXiv preprint arXiv:2207.00221*, 2022.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[8] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" In *International Conference on Learning Representations*, 2023. [Online]. Available: `https://openreview.net/forum?id=KRLUvxh8uaX`.

[9] S. Doveh, A. Arbelle, S. Harary, R. Panda, R. Herzig, E. Schwartz, D. Kim, R. Giryes, R. Feris, S. Ullman, *et al.*, "Teaching structured vision&language concepts to vision&language models," *arXiv preprint arXiv:2211.11733*, 2022.

[10] C. Li *et al.*, "Elevater: A benchmark and toolkit for evaluating language-augmented visual models," *Neural Information Processing Systems*, 2022.