FIX-CLIP: Dual-Branch Hierarchical Contrastive Learning via Synthetic Captions for Better Understanding of Long Text

Supplementary Material

6. Prompting Templates for Long-text Caption Synthesis

To ensure the diversity of the synthesis long-text captions, we have set up multiple prompts to instruct Llama3-LLaVA-NeXT-8b [35] to generate long-text captions with detailed descriptions. During the re-caption process, samples are randomly taken from the following 20 prompts.

```
1.  Provide a comprehensive description of
this image, including all visual elements,
their spatial relationships, and the
overall atmosphere.
2.  Generate a detailed caption explaining
what's happening in this image, covering
actions, subjects, environment, and
temporal context.
3.  Analyze this image in detail,
describing the main subjects, background,
lighting, colors, and composition.
4.  Write an extensive caption that
captures both the explicit visual content
and implicit context or story behind this
image.
5.  Describe this image as if explaining it
to someone who cannot see it, including all
relevant details and visual nuances.
6.  Break down the scene components in this
image, detailing the foreground, middle
ground, and background elements.
7.  Describe the environmental context,
lighting conditions, time of day, and
weather elements visible in this image.
8.  Analyze the spatial arrangement and
relationships between all objects and
subjects in this image.
9.  Detail the setting of this scene,
including architectural elements, natural
features, and atmospheric conditions.
10.  Explain the visual dynamics of this
scene, including movement, direction, and
flow of elements.
11.  Elaborate on the image's details such
as the objects' textures, the direction
of shadows, and how they contribute to the
overall look.
12.  Describe the image from top to bottom
and left to right, highlighting every
element and its significance within the
frame.
13.  Generate a caption that delves into
the emotional undertones suggested by the
image's colors, expressions of the subjects,
and the setting.
14.  Analyze the image to explain how the
placement of elements affects the flow and
balance within the visual space.
15.  Write a detailed description of the
image that includes the sizes of the
objects relative to each other and their
proximity.
16.  Describe the image in terms of the
contrast between light and dark areas and
how it shapes the perception of the scene.
17.  Generate a caption that interprets
the possible narrative connections between
different elements in the image.
18.  Analyze the image to explain how the
colors interact with each other and what
mood they create together.
19.  Write a detailed description of the
image that covers the small details often
overlooked, like tiny patterns on objects.
20.  Describe the image by focusing on
the perspective used and how it makes the
viewer experience the scene.
```

7. Abnormal Synthesized Captions

While synthesized captions provide detailed descriptions, MLLMs usually bring hallucination elements. We apply a simple filtering method on captions to reduce repeated words, meaningless sentences, and short results. Fig. 6 shows some abnormal synthesized captions that have been cleaned out from our training datasets.

8. Details of the Setup

8.1. Details of the training datasets

Our model's training corpus comprises six distinct datasets, as enumerated in Tab. 8. The ShareGPT4V [7] dataset, previously employed in Long-CLIP [63] implementation, exhibits exceptional annotation quality. The remaining five established datasets, including CC3M [47], VisualGenome [24], SBU [40], CC12M [47], and YFCC15M [51], underwent

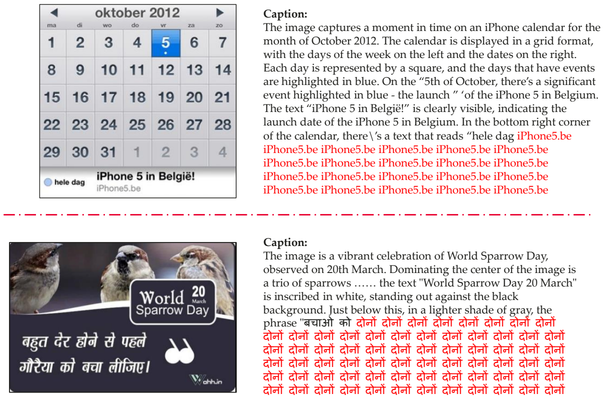


Figure 6. Some incorrect examples from our re-captioned dataset. Both images are wrong captioned with repeating words.

Dataset	Image-text pairs	Sentences per Text	Tokens per Text
CC3M [47]	2760314	6.31	116.96
VisualGenome [24]	107653	6.52	117.68
ShareGPT4V [7]	1246901	9.22	172.94
SBU [40]	835333	6.01	110.33
CC12M [47]	8523767	6.84	131.13
YFCC15M [51]	14994664	6.14	115.38

Table 8. Details of training datasets. We cleaned the data, so the number of image-text pairs is slightly less than that of the original datasets.

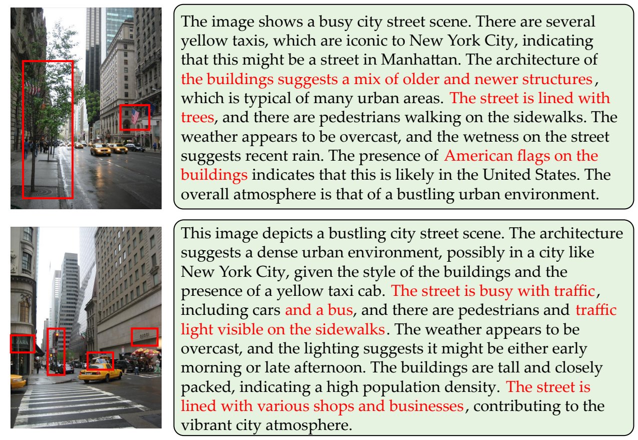


Figure 7. Some examples from our re-captioned dataset. The captions of two similar images are both synthesized by Llama3-LLaVA-NeXT-8b. The key attributes to distinguish these images are marked in red, and highlighted by the red boxes in the images.

	Dataset	Images	texts	Sentences per Text	Tokens per Text
Long-Text	ShareGPT4V [7]	1000	1000	8.15	173.24
	Urban-1k [63]	1000	1000	7.088	129.24
	DCI [52]	7805	7805	10.81	172.73
	IIW [15]	612	612	10.16	39.73
Short-Text	COCO [8]	5000	25000	1.0	11.77
	Flickr30k [43]	1000	5000	1.0	14.03

Table 9. Dataset details of retrieval tasks.

our custom annotation process, utilizing the previously described Llama3-LLaVA-NeXT-8b [35] model for generating extensive long-text caption synthesis. These datasets were systematically organized into three distinct scales: 5M, 15M, and 30M for training purposes.

Tab. 8 provides comprehensive statistics, including the quantity of image-text pairs, sentences per text, and tokens per text. Comparative analysis reveals that our annotated datasets demonstrate marginally lower text lengths relative to ShareGPT4V [7], a characteristic potentially attributed to model-specific limitations, which may impose certain constraints on our model's performance upper bound.

We randomly selected two visually similar images from the VisualGenome [24] dataset, with their corresponding synthesized long-text captions presented in Fig. 7. Despite strong similarities in architectural style, scene elements, weather conditions, and lighting characteristics between these images, our synthesized captions demonstrate precise differentiation of fine-grained details. The text segments highlighted in red accurately delineate the fine-grained visual information contained within the red-bounded regions of the respective images.

8.2. Details of the retrieval tasks

To evaluate our model's cross-modal retrieval capabilities, we conducted experiments on both long-text and short-text retrieval tasks. Traditional retrieval evaluations, primarily conducted onCOCO [8] and Flickr30k [43] with an average text length below 15 tokens, are predominantly focused on short-text image-text retrieval capabilities.

For comprehensive long-text retrieval assessment, we adopted the experimental configurations from established works: Urban-1k [63] and ShareGPT4V [7] settings from Long-CLIP [63], and DCI [52] and IIW [15] configurations from LoTLIP [55], ensuring fair comparative analysis. Tab. 9 presents detailed statistical characteristics of these benchmark datasets.

8.3. Hyperparameters

Training hyperparameters of FIX-CLIP are presented in Tab. 12. For a fair comparison, our training hyperparameters are consistent with Long-CLIP [63].

9. Raw Short Caption versus Synthesis Short Caption

We identified quality limitations in the raw short captions within our training dataset through empirical observation. To address this constraint, we proposed an alternative approach utilizing synthetically generated short captions as model inputs. We conducted comprehensive comparative analyses between models trained on synthetic short captions versus

	Model	DCI		IIW		ShareGPT4V-1k		Urban-1k		Avg.
		I-to-T	T-to-I	I-to-T	T-to-I	I-to-T	T-to-I	I-to-T	T-to-I	
B/16	Raw Short Caption	66.2	67.1	**97.1**	96.7	97.8	97.6	87.7	90.1	87.5
	Synthesis Short Caption	**67.1**	**67.5**	96.9	**96.7**	**98.1**	**97.9**	**88.0**	**90.8**	**87.9**
L/14	Raw Short Caption	66.5	69.1	**97.3**	97.0	97.4	97.6	87.9	92.6	88.1
	Synthesis Short Caption	**68.1**	**69.9**	97.1	**97.2**	**98.5**	**98.0**	**88.1**	**93.0**	**88.7**

Table 10. Train on 5M synthesis long captions as the long-text input, we compare the performance between the raw short captions and the synthesis short captions as the short-text input. The R@1 of long-text-image retrieval on DCI [52], IIW [15], ShareGPT4V-1k [7], and Urban-1k [63] datasets. The best results are in **bold**.

	Model	COCO						Flickr30k						Avg.
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image			
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1
B/16	Raw Short Caption	61.0	84.5	90.8	44.6	70.4	79.5	89.2	98.4	99.7	77.4	94.6	97.2	68.0
	Synthesis Short Caption	**61.3**	**84.9**	**91.2**	**47.0**	**72.4**	**81.4**	**89.9**	**98.8**	**99.7**	**78.4**	**95.2**	**97.7**	**69.2**
L/14	Raw Short Caption	62.5	85.6	91.4	48.5	73.6	82.1	92.3	**99.3**	99.7	81.7	95.9	97.9	71.2
	Synthesis Short Caption	**63.2**	**85.8**	**91.5**	**50.5**	**75.4**	**83.6**	**92.5**	99.1	**99.9**	**82.5**	**96.6**	**98.2**	**72.1**

Table 11. Train on 5M synthesis long captions as the long-text input, we compare the performance between the raw short captions and the synthesis short captions as the short-text input. Results of short-caption text-image retrieval on the 5k COCO2017 [8] validation set and the 1k Flickr30K [43] test set. The best results are in **bold**.

Configuration	FIX-CLIP Training
Batch size	2048
Training Epoch	6
Learning Rate	1e-6
Warm-up Steps	200
LR Scheduler	cosine
Optimizer	AdamW [36]
Optimizer hyper-parameters	$\beta_1, \beta_2, \epsilon$ = 0.9, 0.999, 1e-8
Weight decay	1e-2

Table 12. Summary of FIX-CLIP training hyperparameters.

	COCO		Urban1k		DCI	
	I2T	T2I	I2T	T2I	I2T	T2I
Default	**62.0**	**46.7**	**87.0**	**86.8**	**65.1**	**66.7**
Shared Prompts	60.7	46.0	85.2	86.1	62.9	65.5
R2P	60.3	46.0	85.8	85.7	63.1	65.3
P2R	61.5	46.3	86.6	86.1	64.3	66.1
Short PE (len=77)	61.2	46.3	77.8	75.4	56.3	59.1
Long PE (len=248)	**62.0**	**46.7**	**87.0**	**86.8**	**65.1**	**66.7**

Table 13. Above: ablations on the efficacy of region prompts and masks. "Shared Prompts" refers to all the layers utilizing the same shared prompts, "R2P" and "P2R" denote regional prompts attending to all patch embeddings and vice versa. Bottom: the performance comparison of different position embeddings.

those trained on raw short captions, with results presented in Tab. 10 and Tab. 11. The synthetic short captions were generated by Shikra [6]. Quantitative evaluations demonstrate that incorporating synthetic short captions into the training dataset yields substantial performance gains, suggesting the effectiveness of our proposed approach.

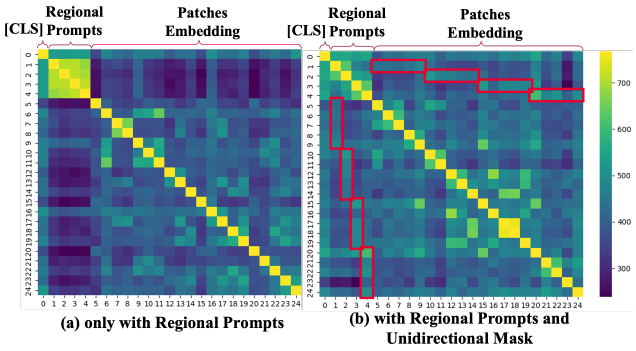


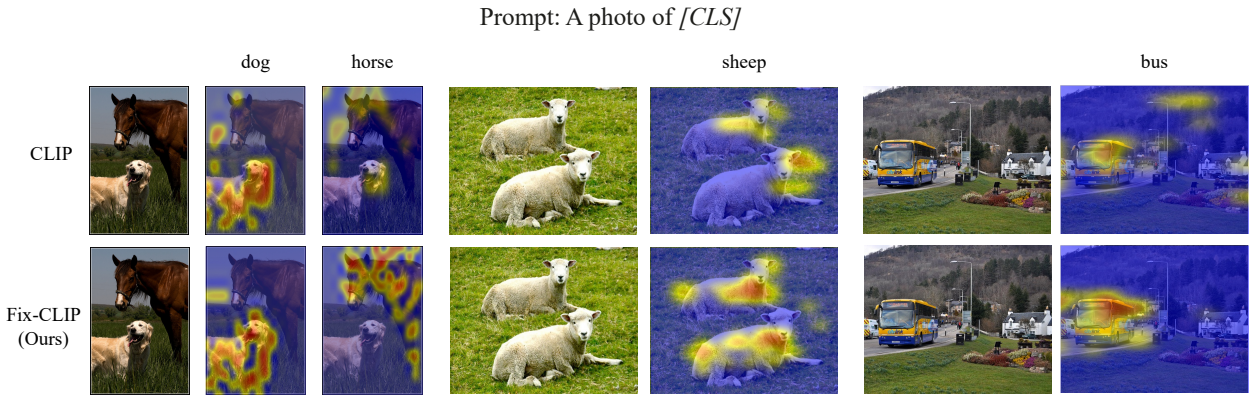
Figure 8. Visualization of the Effects of Unidirectional Masking and Region Prompts.

10. Visualization of the Effects of Unidirectional Masking and Region Prompts

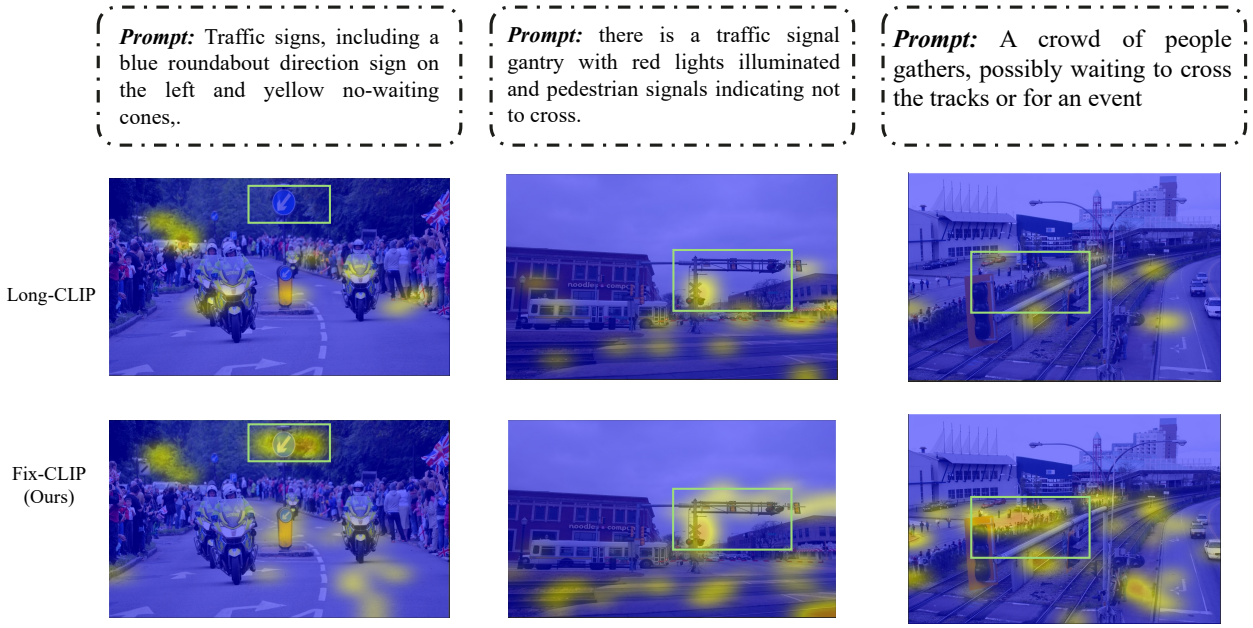
In Fig. 8, the regional prompts obtain stronger responses in the corresponding local patches. The red boxes visualize how regional prompts incorporate local features, highlighting the role of Unidirectional Mask. Moreover, the heatmap (b) exhibits higher global responses compared to heatmap (a).

11. Visualization of the Similarity Heatmap

We visualize the heatmap of similarity between image features and text features, and compare our results with those of CLIP [44] and Long-CLIP [63], as shown in Fig. 9. To evaluate the performance on short texts, the prompt is set as "a photo of *[CLS]*". FIX-CLIP demonstrates superior performance over CLIP [44], accurately identifying instances in the image, as illustrated in Fig. 9a. For long-text under-



(a) Similarity heatmap compared with CLIP [44].



(b) Similarity heatmap compared with Long-CLIP [63].

Figure 9. Similarity Heatmap between text and image features in different models. (a) presents a comparative analysis between our model and CLIP [44] in short-text scenarios, while (b) illustrates the performance comparison between our model and Long-CLIP [63] in long-text contexts. The text segments highlighted in red represent semantic information successfully comprehended by our model but not accurately captured by Long-CLIP [63].

standing, the prompt consists of a major sentence split from the original long-text captions, enabling a direct comparison with Long-CLIP [63]. The corresponding performance is depicted in Fig. 9b.

12. Analysis of Text-to-Image Generation Examples

In this section, we showcase more text-to-image generation examples in long captions to demonstrate the enhancement in understanding long texts. We replace the original text encoder in the stable-diffusion model with that in Long-CLIP [63] or ours. Then, the reconstructed model would be fed with the long captions in the [15] dataset. Due to the divergence between the original text encoder and our text encoder, the model is restrained to generate coarse images. Therefore, an image-to-image refiner model is utilized subsequently to transfer the coarse images to fine images. The final performance is illustrated in Fig. 10. The result of Long-CLIP [63] has confusion in some details, *i.e.* the background, the direction, and the position relation. Even hallucinations would occur, such as the airplane equipping four jet engines in the 4-th case. For the comparison, our model correctly describes the detailed information and performs better.

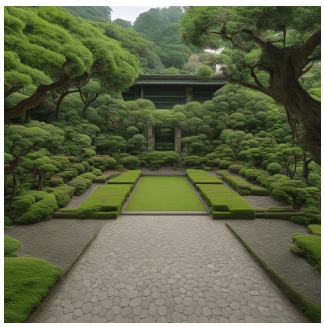
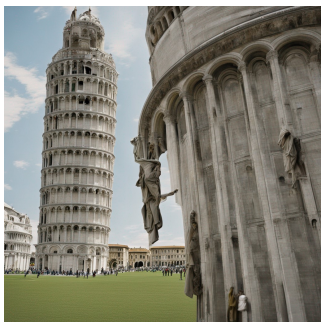
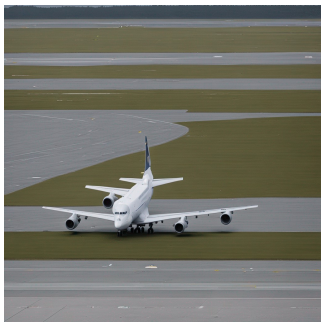
Descriptions	LongCLIP	Fix-CLIP (Ours)
This is a photo of a stone fountain with statues around it with a large building on its left and a large, more ornate building on its right. The stone fountain has a crafted base with a symbol in its middle and has black benches surrounding it. There are people standing and walking around in the background between the buildings. Smaller sections of a building can be seen on either end. The sky has a lot of thin clouds and *there's a mountain or hill range in the background*. There are also little building roof coverings on each side, perhaps for the people to take shade in or sit under. There are multiple street lights behind the stone fountain as well.		
A gardan is shown in the photo. A pinkish pathway extends from the lower left side of the image to just right of the image center. Along this pathway, thin, slab-like carved stones are staggered along both sides. A small, green shrub is planted behind them in the lower right corner of the image. At the end of the pathway, an open-air structure consists of a slightly raised ground slab, four blue cylindrical supports and a roof built in a traditional architectural style. Behind this structure, tall, leafy green trees are partially visible. The sky that peeks through the branches is bright white.		
The Pisa Cathedral with the Leaning Tower of Pisa behind it is the focus of an eye-level, long shot on a clear day with many tourists near the Italian landmark. The front and right side of the cathedral face the viewer as it is positioned angled toward the left. The cathedral is off-white with some weathered yellow along the bottom side. The three tall front doors are open, and people mill about in the distance in front of the entrance. The dome rises behind the cathedral. The Leaning Tower is behind the church to the right. In the front and side of the cathedral is an expansive empty green lawn.		
A Lufthansa airplane taxis to the right on a light-grey airstrip under a light-grey sky in a full outdoor shot. The airplane is white on the top three-quarters and grey on the bottom one-quarter. It has two jet engines, one on either side, plus a blue tail with a yellow circular symbol. The word \"Lufthansa\" is printed on the side of the airplane in dark-blue. The field in front of the airstrip has short green grass and short brown grass. The far side of the airstrip is mostly short brown grass, behind which are low off-white-and-grey buildings.		

Figure 10. More Text-to-Image Generation examples. Images generated by FIX-CLIP are more accurate in detail information such as color, direction, position, quantity, material, light, and shooting angle. The text highlighted in green represents fine-grained details that Long-CLIP [63] fails to capture, whereas our proposed model FIX-CLIP successfully generates these contextual elements with high fidelity.