

HIC-12M: A Large Scale Image-Text Dataset with Hierarchical Dense Image Captions

Chenyu Wang*
TranscEngram
Shanghai, China

Boyang Liu*
TranscEngram
Shanghai, China

Gongrui Zhang*
TranscEngram
Shanghai, China

Zibo Zhao
ShanghaiTech University
Shanghai, China

Shiyu Liu
ShanghaiTech University
Shanghai, China

Ziheng Zhang
TranscEngram
Shanghai, China

Jingyuan Zhang
TranscEngram
Shanghai, China

Ruoyu Wang
TranscEngram
Shanghai, China

Shenghua Gao
The University of Hong
Kong
HKSAR, China
TranscEngram
Shanghai, China

Yi Ma
The University of Hong
Kong
HKSAR, China
TranscEngram
Shanghai, China



lake artist cabin

Scene-Level Image Captioning

Short Caption: An artist is painting a serene lakeside scene with a rustic cabin in the background, surrounded by lush greenery.

Long Caption:

The image depicts an outdoor scene where an individual is painting in a lush, forested area near a serene body of water, possibly a lake or pond. The artist, wearing a wide-brimmed hat, a pink shirt, and green shorts, is seen from behind, focused on a canvas mounted on an easel. The canvas features a painting of the surrounding landscape, capturing elements like trees, water, and possibly the cabin in the background. Holding a paintbrush, the artist appears to be adding details. In front of him is a wooden palette with various paint colors, placed on a small table or stand that also holds papers, possibly sketches or notes. In the background, a dense forest with tall trees surrounds a small rustic cabin near the water's edge. Made of wood, the cabin blends harmoniously with the natural surroundings. The water reflects the trees and cabin, adding to the tranquil scene. Overall, the image captures a moment of artistic creation in a peaceful setting, emphasizing the beauty of nature and the artist's dedication to portraying it.



Region-Specific Image Captioning

lake The lake is a calm, still body of water, mirroring the dense evergreen forest on the far bank and patches of blue sky. In the foreground, green foliage and a marshy bank appear on the left. An artist, seen from behind, stands on the near bank, facing a small wooden cabin nestled among the trees across the lake.

artist The artist, seen from behind, wears a wide-brimmed light-tan hat with a dark band, a pinkish-mauve short-sleeved shirt, olive green shorts or trousers, and a dark cloth tucked into the waistband. He is painting at a wooden tripod easel, brush in right hand, left hand (gloved) resting near an open palette box holding colorful oil or acrylic paints and several brushes. Standing on the creek's bank, the artist faces the water, a small wooden cabin and dense forest beyond.

cabin The cabin is a small, rustic, single-story wooden structure with weathered brown shingles and a lighter, possibly greyish, gabled roof. A single dark, multi-paned window is visible on the side facing the viewer. Nestled among tall, dark green evergreens on the far bank of the creek, the cabin's reflection is clearly visible in the still water.

Figure 1: Our proposed HIC-12M dataset is a large scale image-text dataset with hierarchical dense image captions, including both concise and dense scene-level descriptions and detailed region-specific annotations that elaborate on salient objects and regions in the broader scene.

* Authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Dublin

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

Abstract

Recent advances in vision-language models have provided an efficient framework for aligning images and texts at a global, brief level. However, there is a growing interest in extending CLIP-based models to support global dense semantic understanding and capture local, fine-grained details such as object attributes and interactions. Progress in this direction is hindered by the lack of large-scale, fine-grained image-text datasets. To prompt research in fine-grained visual-language understanding, we argue for a new dataset with two

features: (1) images annotated with multi-granularity captions, including both concise and dense descriptions, and (2) region-specific dense captions that emphasize detailed attributes. To this end, we introduce a pipeline for recaptioning images with high-quality concise and dense captions at the image level. Subsequently, we develop a new pipeline for selecting key foreground objects and background regions and generating dense region-specific captions, with a particular focus on object interactions, spatial relationships, and other attribute details. With this pipeline, we curate a large-scale dataset with hierarchical image captions, HIC-12M, comprising over 12 million image-caption pairs. Each image is annotated with concise, dense, and region-specific captions. We further demonstrate the utility of HIC-12M by training CLIP-based models capable of processing long-text inputs and evaluating their performance on retrieval and classification tasks. We believe our work can contribute to the future research in fine-grained visual-language understanding. The dataset and supplementary materials are available at <https://github.com/Chenyu-Wang567/HIC-12M>.

Keywords

Fine-grained Vision-Language Understanding, Multi-granularity Caption, Region-specific Caption

ACM Reference Format:

Chenyu Wang, Boyang Liu, Gongrui Zhang, Zibo Zhao, Shiyu Liu, Ziheng Zhang, Jingyuan Zhang, Ruoyu Wang, Shenghua Gao, and Yi Ma. 2025. HIC-12M: A Large Scale Image-Text Dataset with Hierarchical Dense Image Captions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

HIC-12M: A Large Scale Image-Text Dataset with Hierarchical Dense Image Captions

Supplementary Material

In these supplementary materials, we provide the following:

- Dataset access and license (Section 1)
- Experimental Details (Section 2)
- Constructed prompts (Section 3)
- More Dataset Samples Visualization (Section 3)

1 Dataset Access and License

To ensure transparency and facilitate future research, we are publicly releasing the HIC-12M dataset on the Hugging Face platform. In compliance with the licensing requirements of the original LAION-400M dataset, we do not redistribute the raw images directly. Instead, we provide the metadata, including original URLs, text captions and image size information, enabling researchers to reconstruct the dataset by downloading images from their original sources. This approach respects the copyright and licensing conditions of the source data, while still enabling full reproducibility of our work. The released data is intended solely for academic research, with commercial use explicitly excluded. All materials are shared under the terms of the CC BY 4.0 license to ensure ethical reuse and proper attribution.

2 Experiment Details

Evaluation Dataset. Following [1, 7], we evaluate our method on four long-form text-image retrieval datasets: Dense Captioning Images (DCI) [6], Urban-1k [8], ShareGPT4V-1k, and ShareGPT4V-10k [2]. DCI contains 7,805 image-text pairs, with each caption averaging 172.73 tokens, densely describing multiple aspects of the image. Urban-1k includes 1,000 urban-view images from Visual Genome, annotated with detailed captions generated by GPT-4V. Each caption offers rich descriptions of scene elements such as object types, colors, and spatial relationships. ShareGPT4V-1k and ShareGPT4V-10k are constructed from the SA-1B dataset [5], where each image is paired with a long-form caption produced by GPT-4V. ShareGPT4V-1k is a subset of ShareGPT4V-10k, comprising 1,000 image-caption pairs with an average of 173.24 tokens per caption. ShareGPT4V-10k consists of 10,000 such pairs, with captions averaging 173.66 tokens.

Training Details. To ensure a fair comparison, we strictly follow the original model configurations for all baseline methods. Specifically, COSMOS [4], FLAME [1], FLAIR [7], and Long-CLIP [8] are implemented based on the ViT-B/16 backbone, while LLM2CLIP [3] uses the ViT-L/14 architecture. We keep the model architecture fixed and vary only the training datasets on downstream performance. All reported results for these methods are directly taken from their respective original papers to ensure consistency. Furthermore, we adopt the same number of training epochs as used in the original implementations for each method.

3 Constructed Prompts and More Visualization Samples

We illustrate the design of three key prompts used during the construction of our dataset. In Figure 2, we show the prompt used to verify object labels based on visual content within a targeted bounding box. Given a proposed object label and a corresponding image with a bounding box, the prompt guides the model to either confirm the correctness of the label or suggest a more accurate alternative based on visual evidence. In Figure 3, we present the prompt used to generate region-specific dense captions. This prompt explicitly instructs the model to focus on describing the targeted object itself, including its attributes, appearance, position, and interactions with nearby objects, unless surrounding entities directly interact with the target in a semantically meaningful way. We also impose specific format constraints on the output to ensure consistency and quality across captions in the later step. Finally, in Figure 4, we visualize the prompt used for refining raw image-level captions. This step involves restructuring the initially generated free-form text into a well-defined JSON format.

We further showcase two additional samples in our proposed dataset in Figure 5 and Figure 6.

Label Verification Prompt

You are a Visual-Semantic Validation Assistant. Your task is to evaluate whether a given object label accurately describes the primary object within a specified bounding box in an image. If the label is inaccurate, suggest a corrected label.

Input:

- (1) Object label: The proposed label for the object.
- (2) An image with a bounding box: The bounding box containing the object.

Tasks:

- (1) **Label Verification:** Evaluate whether the object label accurately describes the primary object visible within the bounding box.
 - If the label is accurate, return "accurate".
 - If the label is incorrect, or misrepresents the object, return "inaccurate".
- (2) **Label Correction:** If the verification result is "inaccurate", provide a corrected label that concisely and accurately describes the object within the bounding box.

Figure 2: Prompt used for verifying object labels and suggesting corrections based on visual content within a bounding box if necessary.

Region Specific Caption Generation Prompt

You are an Expert Region-Specific Captioner. Your task is to generate a fine-grained, descriptive caption for a specific object within a designated bounding box. The object label is provided.

Input:

- (1) Object label: The label of the primary object to describe.
- (2) An image with a bounding box: The bounding box containing the object which helps you mentally focus on the correct region.

Captioning Guidelines:

- (1) **Primary Focus on the targeted object label:** The caption must be centered around the object label. Describe its salient visual attributes and mention any unique or distinguishing features of the object label clearly visible within its bounding box.
- (2) **Interactions and Positional Context:** Briefly describe the targeted object label’s interaction or relationship with its immediate surroundings or nearby objects. Keep this part concise and directly relevant to the targeted object label.
- (3) **Avoid Detailing Other Objects:** Do NOT provide detailed descriptions of other objects in the scene, even if they are nearby, unless they are directly interacting with the targeted object label in a significant way. The goal is a caption about the targeted object label, not a general scene description.

Output Format: <Object label>. <Description>.

Figure 3: Prompt used for generating region-specific captions.

Refining and Post-processing Prompt

You are an expert in text refinement and structuring. Your task is to process a raw image caption, remove specified redundant phrases, decouple the primary object label, and format the output as a JSON object.


Input: A raw caption string.

Processing Guidelines:

- Remove Redundant Phrases:** Identify and remove semantically redundant phrases listed in `ERROR_PHRASES_LIST`. Ensure the removal preserves the caption's factual content, intent, grammatical correctness, and natural language flow. Rephrase minimally if needed to maintain coherence, without adding new information.
- Decouple Object Label:** Identify the primary object label (typically the noun in the first sentence) from the cleaned caption and separate it. Remove the label's redundant references from the caption while preserving meaning.
- Format Output as JSON:** Structure the output as a JSON object with two fields:
 - `"label"`: A string containing the decoupled object label.
 - `"caption"`: A string containing the cleaned caption after removing redundant phrases and decoupling the label.

Error Phrases (`ERROR_PHRASES_LIST`): "border", "red border", "green border", "bounding box", "red bounding box", "within the red bounding box", "green bounding box", "within the green bounding box"

Figure 4: Prompt used for refining raw image captions and structuring them into JSON format.

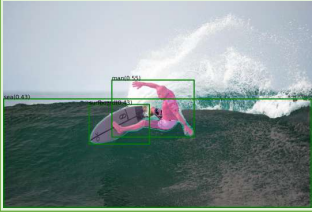


Scene-Level Image Captioning

Short Caption: A surfer in a red shirt and black shorts is performing a maneuver on a white surfboard while riding a wave in the ocean.

Long Caption: The image depicts a surfer riding a wave in the ocean. The surfer is captured in mid-action, performing a maneuver on a white surfboard. The surfer is wearing a red and white wetsuit, with a stylized "OK" logo visible on the board. The wave is breaking to the right of the surfer, creating a dynamic spray of water that contrasts with the clear blue-green ocean water. The sky is mostly clear, with a few scattered clouds, indicating good weather conditions for surfing. The surfer's posture is balanced and focused, with one arm extended for stability and the other bent at the elbow. The overall scene conveys a sense of motion and skill, highlighting the athleticism and precision required in surfing.

sea man surfboard



Region-Specific Image Captioning

sea The sea is a powerful, turquoise-green expanse fading to blue-grey at the hazy horizon beneath an overcast sky. The foreground water is choppy, with foam and spray, while the wave's face and surrounding sea are textured with ripples and whitewater. The surfer, on a white board, cuts sharply across the wave, generating more spray.

man The man is an athletic, bald surfer with a focused expression, captured mid-maneuver. He wears a red and black short-sleeved wetsuit top with white text (including "USA" and "OLYMPICS") and black and light blue shorts marked with an "OK" logo. A black watch is on his left wrist. Leaning back sharply with arms outstretched—right arm higher—he balances on a white surfboard, both feet planted. He rides a vibrant turquoise wave, cutting through the water and generating a dramatic arc of white spray behind and above the wave's crest.

surfboard The surfboard is a white shortboard with a pointed nose and rounded tail, featuring black graphics—including a stylized "OK" logo near the nose and a black rectangular area near the tail, likely a traction pad. The surfboard is actively being ridden by the man, positioned at a sharp angle against the steep face of the turquoise wave.

Figure 5: HIC-12M Sample 1.

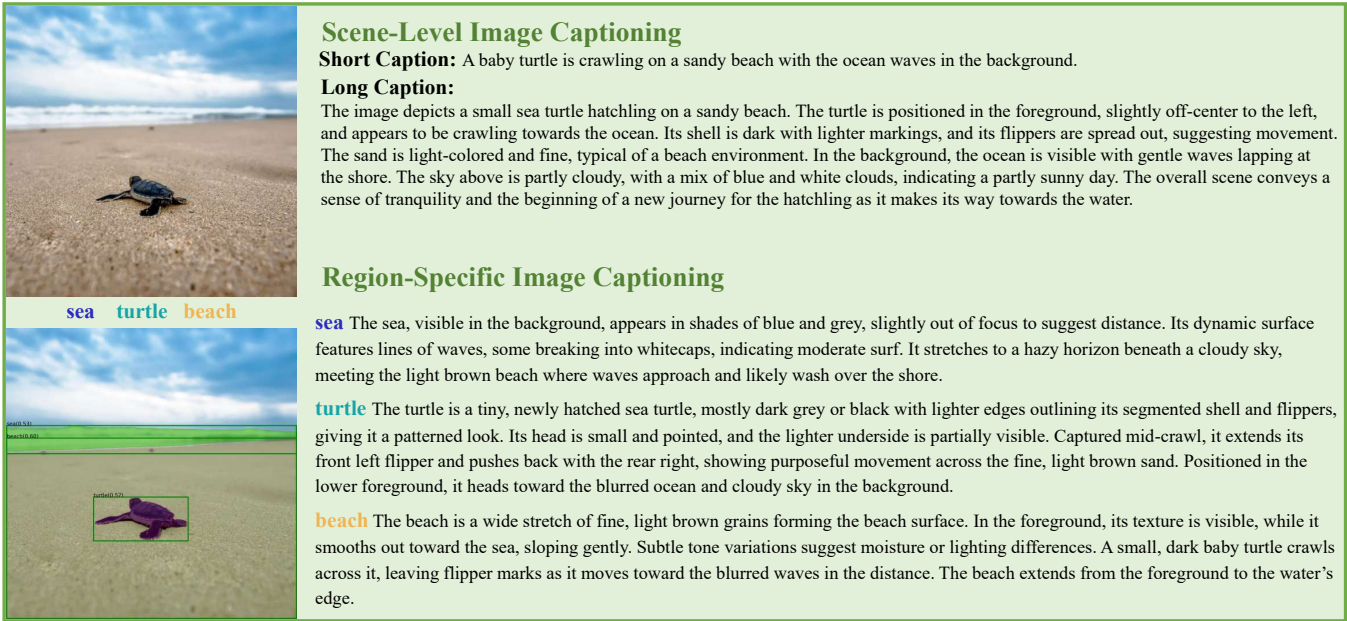


Figure 6: HIC-12M Sample 2.

References

[1] Anjia Cao, Xing Wei, and Zhiheng Ma. 2024. FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training. *arXiv preprint arXiv:2411.11927* (2024).

[2] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793* (2023).

[3] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. 2024. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997* (2024).

[4] Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. 2024. COSMOS: Cross-Modality Self-Distillation for Vision Language Pre-training. *arXiv preprint arXiv:2412.01814* (2024).

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.

[6] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2023. A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. *arXiv preprint arXiv:2312.08578* (2023).

[7] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. 2024. FLAIR: VLM with Fine-grained Language-informed Image Representations. *arXiv preprint arXiv:2412.03561* (2024).

[8] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*. Springer, 310–325.