CHAIN-OF-TALKERS (COTALK): Fast Human Annotation of Dense Image Captions

Yijun Shen[1,*] **Delong Chen**[2,*] **Fan Liu**[1,*,†] **Xingyu Wang**[1]
Chuanyi Zhang[1] **Liang Yao**[1] **Yuhui Zheng**[1]

[1]Hohai University [2]HKUST

`*` Equal Contribution † Corresponding Author
Email: fanliu@hhu.edu.cn

Abstract

While densely annotated image captions significantly facilitate the learning of robust vision-language alignment, methodologies for systematically optimizing human annotation efforts remain underexplored. We introduce **CHAIN-OF-TALKERS (COTALK)**, an AI-in-the-loop methodology designed to maximize the number of annotated samples and improve their comprehensiveness under fixed budget constraints (*e.g.*, total human annotation time). The framework is built upon two key insights. First, sequential annotation reduces redundant workload compared to conventional parallel annotation, as subsequent annotators only need to annotate the *"residual"*—the missing visual information that previous annotations have not covered. Second, humans process textual input faster by reading while outputting annotations with much higher throughput via talking; thus a multimodal interface enables optimized efficiency. We evaluate our framework from two aspects: intrinsic evaluations that assess the comprehensiveness of *semantic units*, obtained by parsing detailed captions into object-attribute trees and analyzing their effective connections; extrinsic evaluation measures the practical usage of the annotated captions in facilitating vision-language alignment. Experiments with eight participants show our CHAIN-OF-TALKERS (CoTalk) improves annotation speed (0.42 vs. 0.30 units/sec) and retrieval performance (41.13% vs. 40.52%) over the parallel method.

1 Introduction

Language is central to human–AI interaction, while vision enables high-bandwidth perception of the world. Aligning these modalities semantically is essential for AI systems to interpret multimodal information effectively. Typically, this alignment is achieved by learning from images paired with human-generated captions. Early datasets, such as COCO (Chen et al., 2015), provide brief single-sentence captions per image, offering basic alignment but limited comprehensiveness. Recently, research has shifted toward creating denser captions, significantly improving vision-language models through enhanced semantic grounding, interpretability, and downstream task performance (Cho et al., 2025; Shabbir et al., 2025).

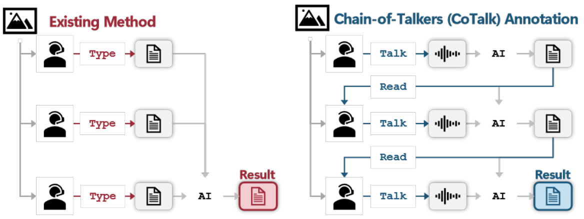


Figure 1: Comparison between Existing Annotation Frameworks and **CHAIN-OF-TALKERS (COTALK)**: Traditional methods require annotators to independently type complete image descriptions. In contrast, COTALK has the first annotator provide a complete spoken annotation, while subsequent annotators focus solely on the *"residual"*—the overlooked details.

Currently, annotation interfaces commonly display images to annotators who examine visual content and generate captions by typing (Garg et al., 2024; Hua et al., 2024). Parallel annotation, wherein multiple annotators independently describe the same image, aims to boost comprehensiveness by aggregating diverse perspectives (Deitke et al., 2024; Athar et al., 2024; Onoe et al., 2024; Hu et al., 2024). Despite their simplicity and ease of use, these methods are largely heuristic, lacking rigorous theoretical justification or systematic validation. In this paper, we identify two fundamental limitations of current annotation practices and propose a novel, systematically validated methodology to overcome these issues.

Our first insight addresses the inefficiency caused by redundancy in parallel annotations, where multiple annotators independently describe the same visual content, leading to substantial overlap (*e.g.*, PixmoCap (Deitke et al., 2024)

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4444–4464
November 4-9, 2025 ©2025 Association for Computational Linguistics

cross-annotation overlap is 71.36% as measured by Sentence-BERT (Reimers and Gurevych, 2019)). Our second insight is that speech-based annotation significantly surpasses typing in speed and efficiency. This is supported by prior research demonstrating that the average throughput for spoken words (161.2 words per minute, WPM) substantially exceeds that of typing (53.46 WPM) (Ruan et al., 2016).

Building on these insights, we introduce **CHAIN-OF-TALKERS (COTALK)**, an novel AI-in-the-loop annotation framework (illustrated in Figure 1). Annotators sequentially contribute descriptions: the first provides a comprehensive initial description from scratch, while subsequent annotators incrementally add only the *"residual"*, the missing visual details. Each annotator reads previous annotations and communicates additional details through speech, which are automatically transcribed and synthesized into coherent text by a Large Language Model (LLM). This sequential, multimodal process significantly reduces redundancy and maximizes annotation comprehensiveness.

The design of COTALK is theoretically grounded in an information-theoretic evaluation framework for image captions (Chen et al., 2024), which provides rigorous guidelines for developing efficient annotation methodologies. To validate our method, we conduct comprehensive assessments of human-generated annotations through both intrinsic and extrinsic evaluations.

Intrinsic Evaluation. Traditional metrics such as caption length are biased by linguistic style, and model-based measures (*e.g.,* CLIP-Score (Hessel et al., 2021), CLIP-IMAGE-Score (Ge et al., 2024)) strongly depend on the performance of the underlying model, limiting their interpretability. To overcome this, we introduce an alternative evaluation based on semantic units—object-attribute pairs extracted by LLMs from detailed captions to construct semantic trees. The number of effective object-attribute connections represents annotation comprehensiveness. Using this metric, we demonstrate that COTALK achieves superior annotation comprehensiveness, generating 36.72 semantic units per image compared to 33.61 with parallel annotation. It also reduces annotation time by approximately 48%, increasing the annotation speed from 0.30 to 0.42 semantic units per second. Additionally, we verify that speech-based annotation is not only faster but also captures richer detail compared to typing, and reading previous annotations leads to better comprehension than auditory reviews alone. Since speech-based annotation reaches a speed of 0.40 semantic units per second, significantly faster than typing at 0.17. Meanwhile, reading prior annotations takes 55.20 seconds with 100% accuracy, compared to 70.20 seconds and 94% accuracy for auditory reviews.

Extrinsic Evaluation. To validate the practical utility of COTALK, we employ retrieval-based evaluation, a robust proxy for annotation quality relevant to real-world vision-language tasks. By fine-tuning a CLIP model (Radford et al., 2021) separately using captions generated via COTALK and parallel annotation methods, we systematically compare their downstream retrieval performances across multiple benchmarks. Our results demonstrate that COTALK consistently yields superior retrieval accuracy, achieving an average of 41.13% across three datasets and six retrieval tasks, surpassing parallel annotations (40.52%). This highlights COTALK's ability to produce more semantically rich and practically valuable annotations, underscoring its effectiveness in enhancing vision-language model performance.

2 Methodology

2.1 Preliminary

Chen et al. (2024) propose an information-theoretic framework for image captioning, systematically defining key criteria for high-quality annotations (Figure 2). The framework introduces a semantic space $S \in [0,1]^n$, where each dimension corresponds to a semantic unit ω_i in $\Omega = \{\omega_i\}_{i-1}^n$, representing the probability of that unit appearing in the image. The caption generation process is modeled as a function f^θ, simulating human annotation. For n annotators, individual annotation processes are denoted $f^1, \ldots, f^n$, producing annotations $\widetilde{Y}^k$ for the k-th annotator. These annotations are mapped to the semantic space as vectors Y via a transformation $h(\cdot)$. In this binary semantic space, a value of 1 indicates the presence of a semantic unit, while 0 indicates its absence. The overlap between the source semantics X and received semantics Y defines their semantic consistency. The semantic-level error is given by $Z = Y - X$, capturing the discrepancy between the intended semantics and the annotated interpretation, thereby reflecting annotation quality.

Chen et al. (2024) then proposes three core objectives to guide high-quality captioning:

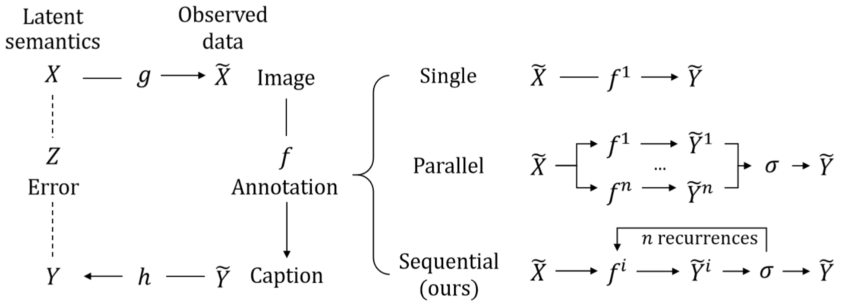


Figure 2: Overview of our formulation. Some latent variable X in a latent semantic space S generates image $\widetilde{X}$ in data space D. The image X is then annotated by f producing a caption $\widetilde{Y}$ which can be mapped back to the original latent space as Y. The semantic-level error measures the annotation quality. The annotation function f can take several forms: in single annotation, a single annotator provides a complete image description; in parallel annotation, multiple annotators independently generate captions that are later merged; and in our sequential annotation, the first annotator provides an initial description, with subsequent annotators incrementally enriching it.

(1) **Information Sufficiency**. $J_{\text{suf}}(\theta) = I(Y; X)$: ensures the caption captures task-relevant semantics by maximizing mutual information;

(2) **Minimal Redundancy**. $J_{\text{min}}(\theta) = -H(\widetilde{Y})$: promotes conciseness by minimizing entropy, reducing repetitive or irrelevant content;

(3) **Human Comprehensibility**. $J_{\text{int}}(\theta) = -D(P_{\widetilde{Y}} \,||\, P_{\text{lang}})$: measures alignment with natural language through distribution distance, enhancing readability. These objectives are integrated into a unified optimization function:

$$J(\theta) = J_{\text{suf}}(\theta) - \beta J_{\text{min}}(\theta) - \gamma J_{\text{int}}(\theta), \quad (1)$$

where β and γ are tunable weights. This formulation offers a flexible and principled approach to both analyzing and optimizing image captioning systems.

2.2 Human Annotation

Traditional human annotation typically involves a single annotator, denoted as f^1, producing the annotation $\widetilde{Y}^1$. Thus, the result of single-round annotation is defined as:

$$\widetilde{Y}_{\text{single}} = \widetilde{Y}^1. \quad (2)$$

In addition to annotation content, time is a crucial factor. The total annotation time in a single round comprises two components: $T_{\text{in}}^{\widetilde{X}}$, the time spent observing the image, and T_{out}, the time spent generating the annotation. The total time can be expressed as: $T_{\text{single}} = T_{\text{in}}^{\widetilde{X}} + T_{\text{out}}^{\widetilde{Y}^1}$.

2.3 Parallel Annotation

Due to the high reliance on a single annotator, single-round annotation often results in inconsistent quality. To address this, *parallel annotation* is introduced, where multiple annotators work independently to produce separate outputs (Deitke et al., 2024; Onoe et al., 2024). These outputs are then aggregated into a unified annotation using a merging function. We define this aggregation process as the **LLM merger** $\sigma(\cdot)$, which consolidates individual annotations into a single, coherent result.

Formally, f^k $(k = 1, \ldots, n)$ generates a description based directly on the image. The annotation is denoted as $\widetilde{Y}^k$.

After all n annotators have completed their annotations, the LLM merges all the annotations from the different annotators together.

$$\widetilde{Y}_{\text{Par}} = \widetilde{Y}_{\sigma}^{n} = \sigma(\widetilde{Y}^1, \ldots, \widetilde{Y}^n), \quad (3)$$

then the final semantic unit of the parallel process can be determined as $Y_{\text{Par}} = Y_{\sigma}^{n} = h(\widetilde{Y}_{\sigma}^{n})$

The total time cost for the parallel annotation process is: $T_{\text{Par}} = n \cdot T_{\text{in}}^{\widetilde{X}} + \sum_{i=1}^{n} T_{\text{out}}^{\widetilde{Y}^i}$.

2.4 Chain-of-Talkers Annotation

We introduce the Chain-of-Talkers (CoTalk) method, designed to reduce human annotation time while maintaining high annotation quality. CoTalk is according to two key insights: (1) sequential annotation can be more efficient than parallel annotation, and (2) generating annotations via speech (talk) is faster than typing, while comprehending prior annotations is quicker through text than audio.

At a system level, CoTalk adopts a sequential annotation strategy, where only the first annotator describes the entire image, and subsequent annotators contribute only residual information—i.e., additions or corrections according to what has already been annotated. At the individual level, CoTalk leverages a cross-modal approach: prior annotations are consumed as text, and new annotations are generated through talk. This framework is illustrated in Figure 1.

Formally, f^k ($k = 2, \ldots, n$) generates a talk description based on previous annotation produced by f^{k-1} (when $k = 1$, the description is based directly on the image). After converting the talk to text, the resulting annotation is denoted as $\widetilde{Y}^k$.

$$\widetilde{Y}^k = f^k(\widetilde{Y}_\sigma^{k-1}, \widetilde{X}), \tag{4}$$

After each annotator completes their annotation, the LLM merges the current annotation with the previously aggregated annotations.

$$\widetilde{Y}_{\text{CoTalk}} = \widetilde{Y}_\sigma^k = \sigma(\widetilde{Y}_\sigma^{k-1}, \widetilde{Y}^k), \tag{5}$$

This process continues until an annotator determines that no further information is necessary and the annotation is complete.

To account for time, we introduce $T_{\text{in}}^{\widetilde{Y}_\sigma^{k-1}}$, representing the time of f^k required to read the previous annotation. The total time cost for the CoTalk method is: $T_{\text{CoTalk}} = n \cdot T_{\text{in}}^{\widetilde{X}} + \sum_{i=1}^{n}(T_{\text{in}}^{\widetilde{Y}_\sigma^{i-1}} + T_{\text{out}}^{\widetilde{Y}^i})$.

3 Theoretical Analysis

3.1 CoTalk is Pareto Optimal

In this section, we compare the CoTalk method with parallel annotation and one-round human annotation to highlight its advantages in both quality and efficiency. Annotation quality is evaluated using J_θ(defined in Equation 1), while efficiency is quantified as:

$$E = \frac{J(\theta)}{T}, \tag{6}$$

where T denotes the total annotation time. Then, we first demonstrate that CoTalk achieves higher informational sufficiency with minimal redundancy, indicating superior annotation quality. Subsequently, we show that for annotations of comparable quality, CoTalk requires a lower annotation budget, thereby achieving Pareto optimality in the quality-efficiency trade-off.

Before proceeding with the analysis, we introduce the following assumption:

Assumption 1 (Diminishing semantic contribution in CoTalk annotations) In CoTalk, the amount of new semantic content added by each successive annotator decreases due to the influence of prior annotations, resulting in $Y_{\text{CoTalk}}^k > Y_{\text{CoTalk}}^{k+1}$.

Assumption 2 (Effective and Correlated Merging of Annotations) In both CoTalk and parallel annotation, the merging function $\sigma(\cdot)$ effectively eliminates redundancy and accurately integrates semantic units. In addition, it exhibits a positive input-output correlation: greater input yields more output, as shown in C.

We now present theorems according to the assumptions:

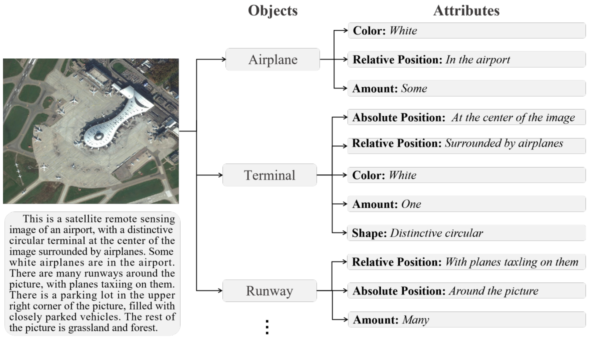
Theorem 1 (CoTalk Enhances Annotation Quality): As defined in Equation 1, high quality annotation is characterized by high information sufficiency, minimal redundancy, and strong human comprehensibility.

Information Sufficiency: Information sufficiency is defined as the completeness of semantic unit coverage, representing how thoroughly annotations capture the semantic content of an image. Under Assumptions 1-2, CoTalk demonstrates superior information sufficiency compared to single-round and parallel annotation. In single-round annotation, a single annotator provides limited semantic coverage. While parallel annotation improves coverage by aggregating multiple independent annotations, it suffers from redundancy, as each annotator samples from the entire semantic space. In contrast, CoTalk allows each annotator to build on the previous ones; the f^k samples from the residual semantic space $Y - Y_\sigma^{k-1}$, focusing only on what has not yet been covered. This targeted supplementation reduces redundancy and increases semantic completeness. Consequently, the incremental information gain per round, defined as $\Delta I(Y_\sigma^k; X) = I(Y_\sigma^k; X) - I(Y_\sigma^{k-1}; X)$, is greater for CoTalk than for parallel annotation when $k >= 2$:

$$\Delta I_{\text{CoTalk}}(Y_\sigma^k; X) > \Delta I_{\text{par}}(Y_\sigma^k; X) \tag{7}$$

This indicates that CoTalk accumulates semantic information more efficiently over successive rounds, thereby achieving higher information sufficiency.

Minimal Redundancy: Single-round annotation inherently contains no redundancy, as it involves only one annotator. In contrast, under Assumption 2, both CoTalk and parallel annotation involve multiple annotators and rely on LLMs to merge their outputs. While parallel annotation gathers full independent annotations from each annotator, CoTalk captures more refined *"residual"* inputs—focused supplements according to prior contributions. Given that LLMs typically produce


Objects
Attributes
Airplane
Color: *White*
Relative Position: *In the airport*
Amount: *Some*
Terminal
Absolute Position: *At the center of the image*
Relative Position: *Surrounded by airplanes*
Color: *White*
Amount: *One*
Shape: *Distinctive circular*
Runway
Relative Position: *With planes taxling on them*
Absolute Position: *Around the picture*
Amount: *Many*
⋮

This is a satellite remote sensing image of an airport, with a distinctive circular terminal at the center of the image surrounded by airplanes. Some white airplanes are in the airport. There are many runways around the picture, with planes taxiing on them. There is a parking lot in the upper right corner of the picture, filled with closely parked vehicles. The rest of the picture is grassland and forest.

Figure 3: Semantic Unit Tree: The first layer is a virtual root node representing the image. The second layer contains all objects in the image, and the third layer captures the attributes of each object.

output proportional to input length, and that word count serves as a proxy for redundancy (being directly related to entropy $H(Y)$, we can compare the overall entropy of the merged outputs. According to Shannon's estimate of 11.82 bits per English word, higher word counts imply higher redundancy (Grignetti, 1964). Since CoTalk reduces redundancy through sequential refinement, it achieves lower overall entropy than parallel annotation:

$$J^{\text{CoTalk}}_{\text{min}}(\theta) \approx J^{\text{Single}}_{\text{min}}(\theta) > J^{\text{Par}}_{\text{min}}(\theta), \tag{8}$$

indicating that CoTalk minimizes redundancy more effectively than parallel annotation, matching the minimal redundancy of single-round annotation while achieving greater information coverage.

Human Comprehensibility: Under Assumption 2, we assume that the merging process is lossless. Since all three annotation methods, CoTalk, parallel, and single-round, rely on human annotators to generate content, their outputs are expected to maintain similar levels of interpretability. Therefore, the human comprehensibility of CoTalk is equivalent to that of the other two methods:

$$J^{\text{CoTalk}}_{\text{int}}(\theta) = J^{\text{Single}}_{\text{int}}(\theta) = J^{\text{Par}}_{\text{int}}(\theta) \tag{9}$$

According to the above results, we conclude that CoTalk achieves higher information content, lower redundancy, and comparable human comprehensibility relative to parallel and single-round annotations. Therefore, it offers superior annotation quality, denoted as $J^{\text{CoTalk}}_{\theta}$.

Theorem 2 (CoTalk boosts efficiency): Building on Theorems 1, we establish that CoTalk yields the highest annotation quality. We now compare the efficiency of CoTalk, parallel annotation, and single-round annotation by evaluating the time required to achieve equivalent semantic unit coverage. Since single-round annotation provides significantly lower information sufficiency, we focus on comparing CoTalk and parallel annotation. Assume parallel annotation can match CoTalk's semantic coverage using m rounds, while CoTalk requires only n rounds ($m > n$, as implied by Equation 7). The total annotation time is then:

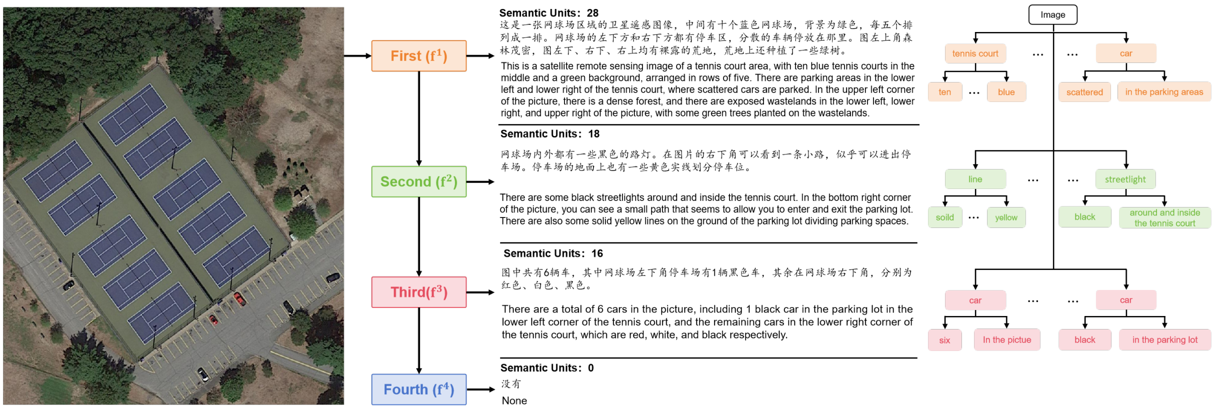
$$\begin{aligned} T^{n}_{\text{CoTalk}} &= n \cdot T^{\widetilde{X}}_{\text{in}} + \sum_{i=1}^{n} (T^{\widetilde{Y}^{i-1}_{\sigma}}_{\text{in}} + T^{\widetilde{Y}^{i}}_{\text{out}}), \\ T^{m}_{\text{Par}} &= m \cdot T^{\widetilde{X}}_{\text{in}} + \sum_{i=1}^{m} T^{\widetilde{Y}^{i}}_{\text{out}} \end{aligned} \tag{10}$$

Here, $T^{\widetilde{Y}^{i}}_{\text{out}} = \frac{|Y^i|}{v_{\text{out}}}$ denotes the time for producing annotations, and $T^{\widetilde{Y}\sigma^{i-1}}_{\text{in}} = \frac{|Y\sigma^{i-1}|}{v^{\text{text}}_{\text{in}}}$ represents the time for reviewing prior annotations. The variables v_{out} and $v^{\text{text}}_{\text{in}}$ refer to the speech annotation and reading comprehension speeds, respectively. Since $v_{\text{out}} > v^{\text{text}}_{\text{in}}$ (Ruan et al., 2016; Brysbaert, 2019), it follows that $T_{\text{CoTalk}} < T_{\text{Par}}$ when $n = 2$, as shown in Proof B. Moreover, the primary time overhead in CoTalk stems from reading prior annotations. However, as semantic coverage increases, the reading time increase per round progressively decreases. In contrast, parallel annotation suffers from growing redundancy, which scales with semantic coverage and leads to increasing time consumption. Hence: $T^{n}_{\text{CoTalk}} < T^{m}_{\text{Par}}$. demonstrating that CoTalk achieves the same semantic coverage with lower time costs. Consequently, its efficiency surpasses that of both parallel and single-round methods: $E_{\text{CoTalk}} > E_{\text{Par}} > E_{\text{Single}}$. In summary, CoTalk not only ensures high annotation quality through improved information sufficiency and reduced redundancy, but also maximizes efficiency—minimizing time and labor, and closely approaching **Pareto Optimality**.

3.2 CoTalk is Faster in Input and Output

CoTalk employs a cross-modal interface, using text for input and talking for output, in contrast to traditional single-modality parallel approaches. In this section, we evaluate the efficiency of different input-output modality combinations to determine the most effective configuration.

First, **talking for output is faster**: The time for talking output can be expressed as: $T^{\text{talking}}_{\text{out}}$, and the time for text output can be defined as: $T^{\text{typing}}_{\text{out}}$. Existing research shows that talking speed is 161.20



Semantic Units：28

这是一张网球场区域的卫星遥感图像，中间有十个蓝色网球场，背景为绿色，每五个排列成一排。网球场的左下方和右下方都有停车区，分散的车辆停放在那里。图左上角森林茂密，图左下、右下、右上均有裸露的荒地，荒地上还种植了一些绿树。

This is a satellite remote sensing image of a tennis court area, with ten blue tennis courts in the middle and a green background, arranged in rows of five. There are parking areas in the lower left and lower right of the tennis court, where scattered cars are parked. In the upper left corner of the picture, there is a dense forest, and there are exposed wastelands in the lower left, lower right, and upper right of the picture, with some green trees planted on the wastelands.

Semantic Units：18

网球场内外都有一些黑色的路灯。在图片的右下角可以看到一条小路，似乎可以进出停车场。停车场的地面上也有一些黄色实线划分停车位。

There are some black streetlights around and inside the tennis court. In the bottom right corner of the picture, you can see a small path that seems to allow you to enter and exit the parking lot. There are also some solid yellow lines on the ground of the parking lot dividing parking spaces.

Semantic Units：16

图中共有6辆车，其中网球场左下角停车场有1辆黑色车，其余在网球场右下角，分别为红色、白色、黑色。

There are a total of 6 cars in the picture, including 1 black car in the parking lot in the lower left corner of the tennis court, and the remaining cars in the lower right corner of the tennis court, which are red, white, and black respectively.

Semantic Units：0

没有

None

Merged Caption: This satellite image depicts a tennis court complex containing ten blue courts arranged in two rows of five at the center, surrounded by green vegetation. The facility is surrounded by black streetlights both around and within the court area. Parking areas with scattered vehicles are located at the lower left and lower right of the tennis court, featuring ground markings with solid yellow lines dividing parking spaces. A small access path is visible in the bottom right corner leading to the parking area. Vehicle distribution includes 6 cars total: 1 black car in the lower left parking zone, and 5 cars (red, white, and black) in the lower right parking area. The upper left corner shows dense forest vegetation, while exposed wastelands with newly planted trees appear in the lower left, lower right, and upper right areas of the image.

Figure 4: CoTalk Example: The first annotator provides a full image description, while subsequent annotators review and add missing details. As the sequence continues, the number of new semantic units gradually declines.

WPM, three times faster than typing, with 20.4% lower error rates (Ruan et al., 2016), we can directly conclude that $v_{\text{out}}^{\text{talking}} > v_{\text{out}}^{\text{typing}}$.

Next, **text for input is faster**: Since eyes spend 10% to 20% time rereading during text input, while replaying audio for review is more time consuming compared to rereading text (Zhan and Shen, 2015). We can conclude that $v_{\text{rereading}} > v_{\text{relistening}}$. Moreover, we can define the time for listening input as: $T_{\text{in}}^{\text{listening}} = T_{\text{listening}} + T_{\text{relistening}}$, and the time for text input can be expressed as: $T_{\text{in}}^{\text{reading}} = T_{\text{reading}} + T_{\text{rereading}}$. The reading comprehension speed for text is 236 WPM, exceeding the 161 WPM speed to understand talking (Brysbaert, 2019), which means that $v_{\text{reading}} > v_{\text{listening}}$. Therefore, we can achieve that $v_{\text{in}}^{\text{reading}} > v_{\text{in}}^{\text{listening}}$.

From the above results, we conclude that in CoTalk, using multimodal input and output can save more time compared to a single modality approach. Specifically, for annotators' output, talking should be adopted due to its higher speed and accuracy. On the other hand, for understanding previous annotations, text is more effective.

4 Experiments

Our experiments aim to validate the effectiveness of different annotation strategies: CoTalk and parallel annotations, as well as to identify the optimal input-output modality for individual annotators, namely talk or text. Finally, we perform an in-depth analysis of the collected human annotation.

4.1 Evaluation Metrics

To assess the quality differences among various annotation methods, we conduct a quantitative comparison using both extrinsic and intrinsic metrics.

4.1.1 Intrinsic Metric

Section 2.1 establishes that semantic units serve as a reliable indicator of annotation quality: more units correspond to higher-quality annotations. We further formalize the function $h(\cdot)$, which maps annotations to semantic units. Specifically, we utilize LLMs to extract these units and organize them into a hierarchical tree, where the root is a virtual node labeled "Image", the second level contains entities, and subsequent levels represent their associated attributes. The total number of edges quantifies the semantic units. For example, in Figure 3, “Terminal” has five semantic units.

Building on semantic units, we quantitatively compare CoTalk and parallel annotation in terms of quality and efficiency. As shown in Table 1, two-person CoTalk yields higher-quality annotations (36.72 units/image vs. 33.61) and reduces per-annotator time by 48%. To assess redundancy, we measure the overlap between outputs from the first and second annotators. Redundancy is defined as the repetition rate of semantic units, including exact matches and semantically similar phrases (*e.g.*, “black car” vs. “black vehicle”) exceeding a similarity threshold using Sentence-BERT (Reimers and Gurevych, 2019). CoTalk shows lower redundancy (29.76%) compared to parallel annotation (69.12%), highlighting the inefficiencies of independent annotation without shared context.

Table 1: Information on different methods. Redundancy refers to internal components within each annotation method.

Method	#Semantic Units	Time	Speed(units/s)	Duplication ↓
Parallel	33.61	118.30	0.30	69.12
CoTalk	**36.72**	**88.43**	**0.42**	**29.76**

Table 2: Comparison of quality and efficiency between speaking and typing, and comprehension between listening and reading.

Method	# Semantic Units	Time	Speed(units/s)
Talking	46.65	116.65	0.40
Typing	33.15	199.15	0.17
	Accuracy	Time	# Frequency
Listening	94.00	70.20	2.22
Reading	100.00	55.20	

Table 3: Validate annotation quality by evaluating performance in downstream retrieval tasks.

Method	RSICD	RSITMD	UCM-Captions	Average ↑
Zero-shot	21.24	31.10	65.01	37.94
Parallel	22.57±0.06	**33.97±0.05**	65.01±0.07	40.52
CoTalk	**23.63 ± 0.05**	33.83 ± 0.09	**65.94 ± 0.06**	**41.13**

Table 4: The Relationship Between Annotation Quality and Annotation Cycles: Annotation quality is analyzed in relation to annotation cycles, where each cycle corresponds to approximately 10 minutes of annotation.

Cycle	# Amount	# Semantic units	Time	Speed(units/s)
1	4	104	745.14	0.14
2	8	289	628.26	0.46
3	9	308	770.00	0.40
4	6	180	720.00	0.25

4.1.2 Extrinsic Metric

Extrinsic Metric assesses the practical effectiveness of annotation methods. We focus on image-text retrieval, a task closely aligned with the goal of maximizing mutual information in InfoNCE. In information theory, higher mutual information indicates better cross-modal predictability. Therefore, stronger retrieval performance, reflects more informative and semantically aligned annotations. To evaluate this, we fine-tune Long-CLIP (Zhang et al., 2024) using annotations generated by CoTalk and parallel methods, then test the models on the remote sensing benchmarks: RSICD (Lu et al., 2017), RSITMD (Yuan et al., 2021), and UCM-Captions (Qu et al., 2016). As shown in Table 3, the model trained with CoTalk annotations achieves the highest average retrieval score (41.13%), outperforming the model trained with parallel annotations (40.52%). More details are in I.1.

With intrinsic and extrinsic metrics, CoTalk annotation consistently outperforms parallel annotation in quality, efficiency, and redundancy.

4.2 Qualitative Analysis

From a qualitative perspective, we explore the concept of *“residual”* within the CoTalk annotation. Annotators typically focus on entities overlooked by previous annotators, followed by missing attributes of these entities, as illustrated in Figure 4. These attributes, such as color, position (relative and absolute), quantity, shape, and size, align with our defined semantic units. Notably, annotators rarely correct prior annotations, probably because of the high accuracy of initial inputs. Instead, each annotation round adds new entities or attributes, embodying the concept of *“residual”* in CoTalk and enriching the description without redundancy.

4.3 Talking is Faster for Output, Reading is Faster for Input

After establishing CoTalk as an effective framework for multi-annotator collaboration, we evaluate which input and output modalities best maximize annotator efficiency, comparing speech and text in both input (viewing prior annotations) and output (producing new annotations) modes.

4.3.1 Annotation Output: Talk vs. Type

To compare the quality and efficiency of talking versus text annotation, we recruit eight experienced annotators, divided into two groups of four. Within each group, two annotators leverage talk input, while the other two use text input via keyboard to annotate the same set of images. The results are averaged across the annotation methods for comparison. As shown in Table 2, talking yields an average of 13.50 more units per image than typing, suggesting greater detail and completeness. Additionally, talking requires 116.65 seconds per image, 41% faster than the 199.15 seconds needed for typing, demonstrating a clear efficiency advantage.

4.3.2 Annotation Input: Read vs. Listen

To compare comprehension from speech and text, we perform an experiment using various images and LLM-generated questions according to human annotations. Eight annotators answer five questions per image in both text and audio formats. As Table 2 shows, reading is on average 15 seconds faster

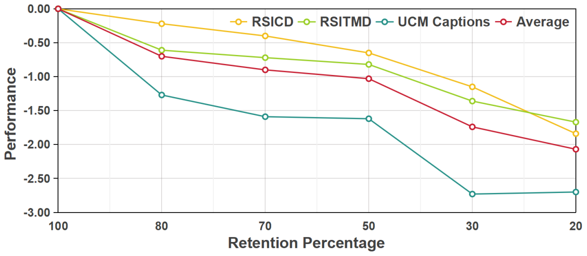


Figure 5: Consistency Analysis of Intrinsic Metric semantic units and Extrinsic Metric by finetuning the Long-CLIP with fixed ratio data reduced. Each point represents its score minus the score from the full dataset.

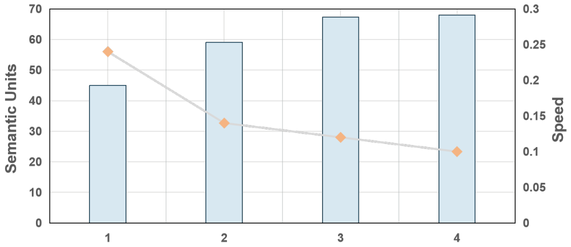


Figure 6: The relationship between images with varying semantic unit counts and the number of required annotators (yellow), and the change in annotation speed at different positions in CoTalk (blue).

and 6.38% more accurate. Listeners need 2.2 replays on average for detailed information, while rereading text takes less time. The gap widens in complex scenes (*e.g.,* urban or industrial).

These findings suggest that relying solely on speech (talking) or text for both input and output is suboptimal. Instead, a hybrid approach using text for input and speech for output can achieve both higher annotation quality and greater efficiency.

4.4 Further Analysis

4.4.1 Consistency between Extrinsic and Intrinsic Metrics

Given the strength of extrinsic metrics in evaluating annotation quality, we test whether intrinsic metrics, specifically semantic unit count, serve as reliable proxies. We fine-tune Long CLIP (Zhang et al., 2024) on full CoTalk annotations, then reduce semantic units per image by fixed ratios and evaluate on remote sensing retrieval benchmarks. As shown in Figure 5, performance decreases steadily with fewer units, averaging a 2.07% decline when only 20% remain. This trend confirms that the amount of semantic units directly impacts annotation informativeness and model performance, which is consistent with extrinsic metric. The results also underscore the value of dense captions for robust vision-language alignment. More details are in I.2.

4.4.2 Annotator Ability Over Time

We recruit untrained annotators for CoTalk and analyze their initial annotation cycles (10 minutes each) to gauge ability and learning curves. As shown in Table 4, novices match experienced speeds (0.46 units/s) after one session, suggesting rapid adaptation via previous examples and familiarity with the task. However, the speed drops from 0.40 to 0.25 units in round four, indicating fatigue or reduced efficiency over time, which highlights the importance of considering session duration and rest intervals in future workflow designs.

4.4.3 Full Image Annotation: Annotator Count and Speed

Multiple annotators sequentially annotate each image using CoTalk until the last annotator deems it sufficiently detailed. We record annotation speed and the number of annotators per image. As shown in Figure 6 (blue), speed decreases as early annotators cover obvious content, leaving finer details for later ones. On average, 3.2 annotators are enough per image (Figure 6, yellow). Moreover, images with more units need more annotators, making semantic unit count a strong complexity indicator. For images with 60 or fewer units, two annotators suffice; more complex ones need at least three.

Given these findings, intrinsic and extrinsic metrics align, since fewer semantic units lead to lower retrieval performance. Furthermore, novice annotators in CoTalk quickly perform as well as experienced ones. Although annotation speed slows over iterations, only 3.2 annotators are needed on average to produce a sufficiently detailed annotation.

5 Conclusion

Our work presented two key insights: sequential annotation minimized redundancy compared to parallel approaches, and humans achieved higher efficiency by reading text input and talking output. Building on these, we proposed Chain-of-Talkers (CoTalk), a cross-modal, sequential collaboration framework that enhanced human annotation for image captioning. Both theoretical and empirical analysis confirmed that CoTalk significantly improved caption quality within fixed budget constraints.

Limitations

Although CoTalk has proven effective in enhancing image caption quality and efficiency, several limitations remain worth discussing.

Assumptions on Human Ability. Our comparison of information sufficiency, human comprehensibility, and efficiency between sequential annotation and CoTalk assumes consistent annotator abilities. However, in practice, annotator capabilities often vary. In sequential annotation, later annotators may identify more *residual* information than earlier ones, while in parallel annotation, annotators may differ significantly in annotation time and quality for the same image. Additionally, we assume all annotations are accurate, but in reality, factors such as annotator ability and attention may cause deviations between the annotated semantic units and the true content of the image.

Assumptions on Model Ability. CoTalk relies on large models to merge annotations from multiple annotators. While our theoretical analysis assumes that these models can accurately consolidate diverse inputs without information loss, practical challenges may arise—particularly when annotators provide contradictory descriptions or inconsistent expressions for the same object. Additionally, CoTalk employs a speech-to-text model to transcribe annotator speech. Although state-of-the-art models such as (An et al., 2024; Radford et al., 2022) achieve high accuracy, transcription errors can still occur, especially with unclear or imprecise speech. Further exploration and careful selection of robust language and speech-to-text models are critical to improving CoTalk's overall reliability.

Ethics Statement

In developing and evaluating the Chain-of-Talkers (CoTalk) framework for image captioning annotation, we upheld rigorous ethical standards to ensure integrity, fairness, and accountability throughout the research process.

Participant Welfare and Informed Consent: All human annotators were fully informed about the nature, purpose, and procedures of the annotation tasks. Participation was entirely voluntary, and written informed consent was obtained from each individual. Annotators were informed of how their data would be utilized for research purposes. We ensured the privacy and anonymity of all participants by removing personal identifiers from the data and results.

Data Integrity and Transparency: We maintained strict data integrity, accurately recording annotation data without manipulation or fabrication. Secure data management practices were implemented to prevent data loss or unauthorized access. All relevant data, results, and methodologies are reported transparently to enable reproducibility and facilitate verification by other researchers.

Broader Impact: The CoTalk framework enhances annotation quality and efficiency, benefiting downstream tasks such as image retrieval and visual understanding. However, annotations may still reflect annotator biases or image-related imbalances, potentially influencing model behavior. High-quality annotations can also be misused in sensitive domains like surveillance. We urge responsible use of CoTalk, with attention to fairness, privacy, and application-specific risks.

Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grant 62372155 and Grant 62302149, in part by the Qinglan Project of Jiangsu Province, in part by Changzhou Science and Technology Bureau Project No. 20231313, in part by the Aeronautical Science Fund under Grant 2022Z071108001, and in part by Science Fund for Excellent Young Scholars of Jiangsu Province No. SBK20250300739.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, and 13 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *ArXiv*, abs/2407.04051.

Ali Athar, Xueqing Deng, and Liang-Chieh Chen. 2024. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. *ArXiv*, abs/2412.09754.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Doll'ar, and Christoph Feichtenhofer. 2025. Perception encoder: The best visual embeddings are not at the output of the network.

Marc Brysbaert. 2019. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*.

Delong Chen, Samuel Cahyawijaya, Etsuko Ishii, Ho Shu Chan, Yejin Bang, and Pascale Fung. 2024. What makes for good image captions?

Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Dutt Jain, Miguel Martin, Huiyu Wang, Hanoona Abdul Rasheed, Peize Sun, Po-Yao Huang, Daniel Bolya, Nikhila Ravi, Shashank Jain, Tammy Stark, and 10 others. 2025. Perceptionlm: Open-access data and models for detailed visual understanding.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Christopher Callison-Burch, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. In *Conference on Empirical Methods in Natural Language Processing*.

Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024. Visual fact checker: Enabling high-fidelity detailed caption generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14033–14042.

Mario C. Grignetti. 1964. A note on the entropy of words in printed english. *Information and Control*, 7(3):304–306.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718.

Shiyu Hu, Xuchen Li, Xuzhao Li, Jing Zhang, Yipei Wang, Xin Zhao, and Kang Hao Cheong. 2024. Can lvlms describe videos like humans? a five-in-one video annotations benchmark for better human-machine comparison. *ArXiv*, abs/2410.15270.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. 2023. Rsgpt: A remote sensing vision language model and benchmark. *ArXiv*, abs/2307.15266.

Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Zhifei Zhang, Yilin Wang, Jianming Zhang, and Jiebo Luo. 2024. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. *ArXiv*, abs/2411.15411.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023. Monkey: Image resolution and text label are important things for large multi-modal models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26753–26763.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. 2024. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ArXiv*, abs/2411.09301.

Fan Liu, Liang Yao, Chuanyi Zhang, Ting Wu, Xinlei Zhang, Xiruo Jiang, and Jun Zhou. 2025. Boost uav-based object detection via scale-invariant feature disentanglement and adversarial learning. *IEEE Transactions on Geoscience and Remote Sensing*.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56:2183–2195.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. *ArXiv*, abs/2404.13013.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *ArXiv*, abs/2307.10350.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. Docci: Descriptions of connected and contrasting images. *ArXiv*, abs/2404.19753.

Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. 2025. Geopix: Multi-modal large language model for pixel-level image understanding in remote sensing. *ArXiv*, abs/2501.06828.

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. In *2016 International*

Conference on Computer, Information and Telecommunication Systems (CITS), pages 1–5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. 2025. Geopixel: Pixel grounding large multimodal model in remote sensing. *ArXiv*, abs/2501.13925.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *ArXiv*, abs/2406.10328.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. 2025. Remotesam: Towards segment anything for earth observation. *arXiv preprint arXiv:2505.18022*.

Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2021. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19.

Wenlian Zhan and Zhulin Shen. 2015. Esl reading research based on eye tracking techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 13:360–368.

Beichen Zhang, Pan Zhang, Xiao wen Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*.

Appendix

A Related Work

A growing trend in image captioning research was the emphasis on generating more comprehensive captions (Cho et al., 2025; Bolya et al., 2025; Shabbir et al., 2025; Hua et al., 2024; Athar et al., 2024; Singla et al., 2024; Ma et al., 2024; Nguyen et al., 2023), aiming for higher information sufficiency through broader semantic coverage. Recently, numerous dense captioning datasets emerged, including high-quality manually annotated datasets (Onoe et al., 2024; Deitke et al., 2024; Hu et al., 2023) and pseudo-labeled datasets generated by models (Yao et al., 2025; Ou et al., 2025; Liu et al., 2025; Ge et al., 2024; Singla et al., 2024; Chen et al., 2023). Although some approaches incorporated prior knowledge (Ou et al., 2025; Li et al., 2024) or utilized visual models for secondary correction (Ge et al., 2024; Li et al., 2023), the quality of model-generated detailed captions remained notably inferior to that of human annotations.

However, traditional human annotation was labor-intensive, and the associated time and coordination costs significantly constrained scalability and efficiency. While several studies demonstrated that involving multiple annotators improved the detail and accuracy of descriptions (Hu et al., 2023; Onoe et al., 2024), these benefits came at the expense of a higher annotation overhead. To improve annotation efficiency, some studies introduced models to assist human annotators. For example, models could first identify key objects or generate preliminary descriptions, which were then refined by humans (Cho et al., 2025; Garg et al., 2024). However, these approaches still relied on traditional typing for annotation, which remained time-consuming. More recently, several works explored using speech input to accelerate annotation, reducing typing time costs and mitigating issues such as annotator manipulation during large-model-assisted labeling (Deitke et al., 2024; Athar et al., 2024). Nonetheless, most of these methods relied on merging multi-person speech annotations, which often introduced redundancy and partially offset the efficiency gains.

To overcome these limitations, we proposed CoTalk, a human-AI collaborative annotation framework based on two key insights. First, sequential annotation, where subsequent annotators provided *"residual"* supplements to previous annotations, yielded higher quality and efficiency than parallel annotation. Second, combining text-based input with talk-based output in this sequential setting optimized both speed and accuracy, outperforming traditional single-modality approaches.

B Proof for Theorem

Proof 1 When the number of annotators in CoTalk is $n = 2$, under the same semantic unit coverage, the total time for parallel annotation exceeds that of CoTalk annotation, indicating that CoTalk is more efficient. Before proceeding, we review the necessary assumptions: (1) Taggers have equal abilities, therefore the number of semantic units supplemented by the second annotator $\widetilde{Y}^2_{\text{CoTalk}}$ is less than the number initially annotated by the first $\widetilde{Y}^1_{\text{CoTalk}}$; (2) According to Equation 7, the number of annotators in parallel annotation m is greater than in CoTalk n, i.e., $(m > n)$, with both m and n as integers; (3) Prior research shows that the output speed for voice annotation v_{out} is faster than the reading comprehension speed $v^{\text{text}}_{\text{in}}$ (Ruan et al., 2016; Brysbaert, 2019). Given $n = 2$, the CoTalk annotation time is:

$$T^2_{\text{CoTalk}} = 2 \cdot T^{\widetilde{X}}_{\text{in}} + T^{\widetilde{Y}^1_{\text{CoTalk}}}_{\text{in}} + \sum_{i=1}^{2} T^{\widetilde{Y}^i_{\text{CoTalk}}}_{\text{out}} \tag{11}$$

For parallel annotation:

$$T^m_{\text{Par}} = m \cdot T^{\widetilde{X}}_{\text{in}} + \sum_{i=1}^{m} T^{\widetilde{Y}^i_{\text{Par}}}_{\text{out}} \tag{12}$$

where $m > n$. Considering an extreme case where $m = 3$ means that the number of effective semantic units completed by three parallel annotators is the same as the two annotators in CoTalk, and assuming (according to premise 1) that all annotators have the same ability, each parallel annotator completes Y^1_{par} semantic units. Therefore, the time for parallel annotation becomes:

$$T^3_{\text{Par}} = 3 \cdot T^{\widetilde{X}}_{\text{in}} + 3 \cdot T^{\widetilde{Y}^1_{\text{Par}}}_{\text{out}} \tag{13}$$

The time difference between parallel and CoTalk annotation for the same number of semantic units is:

$$\Delta T(n = 2, m = 3) = T^3_{\text{Par}} - T^2_{\text{CoTalk}} \tag{14}$$

Expanding:

$$\Delta T = T^{\widetilde{X}}_{\text{in}} + 2 \cdot T^{\widetilde{Y}^1_{\text{Par}}}_{\text{out}} - T^{\widetilde{Y}^2_{\text{Cotalk}}}_{\text{out}} - T^{\widetilde{Y}^1_{\text{CoTalk}}}_{\text{in}} \tag{15}$$

Table 5: The simulation results of sequence and parallel on large model

Model	RSICD								RSITMD								UCM-Captions							
	Image to Text			Average	Text to Image			Average	Image to Text			Average	Text to Image			Average	Image to Text			Average	Text to Image			Average
	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	R@5	R@10		R@1	R@5	R@10	
Zero-shot	9.70	23.88	38.15	23.91	5.13	19.11	29.85	18.03	13.50	33.63	45.58	30.90	10.52	31.57	47.10	29.73	39.50	80.00	90.48	69.99	28.38	59.68	77.19	55.08
Parallel	10.06	**25.62**	37.69	24.46	6.21	20.60	32.84	19.88	**15.93**	36.73	47.57	33.41	12.27	35.58	50.31	32.72	**41.43**	**79.52**	91.43	**70.79**	30.24	**64.19**	**82.76**	59.06
CoTalk	**10.25**	25.53	**38.33**	**24.70**	**6.78**	**21.82**	**33.97**	**20.86**	15.49	**38.05**	**49.78**	**34.44**	**12.88**	**36.86**	**52.05**	**33.93**	40.00	**79.52**	**91.90**	70.47	**31.56**	**64.19**	82.23	**59.33**

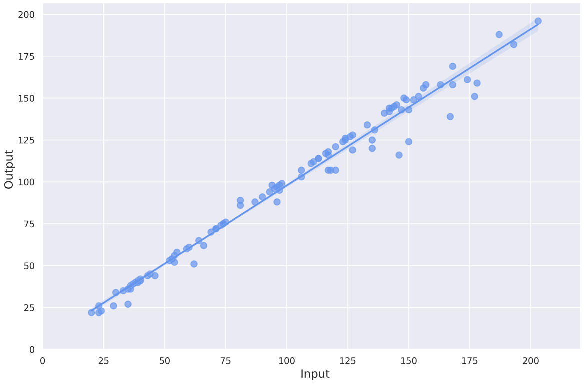


Figure 7: Proof of Assumption 2: The Merging Function Demonstrates a Positive Input-Output Correlation.

Table 6: The analysis of the detail in LLM merging

	# Object Num	# Attribute Num	Redundancy Rate (%)
First Annotator	10.53	26.02	19.36
Second Annotator	3.41	8.34	
Combined	12.51	31.17	/

Table 7: Robustness to Annotator Variance Analysis

Level	High	High + Medium	High + Medium + Low
# Object Num	14.60	14.75	**17.00**
# Attribute Num	37.35	42.65	**46.30**
# Semantic Units	51.95	57.40	**63.30**

$$\Delta T = T_{\text{in}}^{\widetilde{X}} + \frac{2 \cdot |\widetilde{Y}_{\text{Par}}^{1}| - |\widetilde{Y}_{\text{CoTalk}}^{2}|}{v_{\text{out}}} - \frac{|\widetilde{Y}_{\text{CoTalk}}^{1}|}{v_{\text{in}}^{\text{text}}} \tag{16}$$

Since $v_{\text{out}} < v_{\text{in}}^{\text{text}}$, it follows that:

$$\frac{\widetilde{Y}_{\text{CoTalk}}^{1}}{v_{\text{in}}^{\text{text}}} < \frac{\widetilde{Y}_{\text{CoTalk}}^{1}}{v_{\text{out}}} \tag{17}$$

thus:

$$\Delta T > T_{\text{in}}^{\widetilde{X}} + \frac{\widetilde{Y}_{\text{CoTalk}}^{1} - \widetilde{Y}_{\text{CoTalk}}^{2}}{v_{\text{out}}} > 0 \tag{18}$$

Given that taggers have identical abilities $\widetilde{Y}_{\text{par}}^{1} = \widetilde{Y}_{\text{CoTalk}}^{1}$ and $\widetilde{Y}_{\text{CoTalk}}^{1} > \widetilde{Y}_{\text{CoTalk}}^{2}$, the positive time difference ΔT is established. Moreover, as m increases, the total time for parallel annotation grows. Therefore, for any $m > n$, the time difference satisfies:

$$\Delta T(n = 2, m > n) = T_{\text{Par}}^{m} - T_{\text{CoTalk}}^{2} > 0 \tag{19}$$

confirming that CoTalk annotation remains more efficient than parallel annotation under these base conditions.

C Support for Assumption

Support for LLM input-output consistency: To validate the consistency of $\sigma(\cdot)$, specifically, that more input yields more output. We merge multi-person annotations at the sentence level. Token counts exclude prompt tokens, considering only those from the annotated sentences to be merged. As shown in Figure 7, output tokens increase with input tokens, which is consistent with the hypothesis.

Support for Effective Merging of Annotations: To validate this assumption empirically, we conduct a manual analysis of human-annotated examples. It shows that 28.64% of later annotators add new objects, while 71.36% enrich existing objects with additional attributes. Importantly, no corrections to previous annotations are observed.

To address rare annotation conflicts, we specify in the prompt (Table 11) that the later annotator's input takes precedence, as it incorporates corrections with full knowledge of prior annotations and thus reflects higher confidence. We also simulate conflicting cases and find that the LLM resolves them with 100% accuracy.

Table 8: The complete experimental results of Extrinsic Metric in the Ret-3 dataset benchmark evaluation.

Model	RSICD							RSITMD							UCM-Captions						
	Image to Text			Text to Image			Average	Image to Text			Text to Image			Average	Image to Text			Text to Image			Average
	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	R@5	R@10	R@1	R@5	R@10	
Zero-shot	9.70	23.88	38.15	5.13	19.11	29.85	20.97	13.50	33.63	45.58	10.52	31.57	47.10	30.32	39.50	80.00	90.48	28.38	59.68	77.19	62.54
Parallel	9.79	26.19	38.46	6.19	20.96	33.83	22.57	**16.24**	**36.73**	**48.76**	**13.52**	36.61	52.02	**33.97**	44.57	76.48	90.10	32.26	64.83	81.80	65.01
CoTalk	**10.41**	**27.32**	**39.65**	**6.49**	**22.5**	**35.38**	**23.63**	15.49	35.62	48.27	13.17	**37.35**	**53.04**	33.83	**46.67**	**78.38**	89.24	32.41	64.77	**84.19**	**65.94**

Table 9: Consistency analysis of Intrinsic Metrics Complete data results.

Percentage	RSICD							RSITMD							UCM-Captions						
	Image to Text			Text to Image			Average	Image to Text			Text to Image			Average	Image to Text			Text to Image			Average
	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	R@5	R@10	R@1	R@5	R@10	
100%	11.25	27.45	40.35	6.42	21.08	33.69	23.38	17.04	35.18	49.12	13.45	36.24	51.63	33.78	43.33	79.52	91.9	34.75	63.13	81.17	65.63
80%	11.80	27.17	40.53	6.26	20.64	32.58	23.17	17.48	34.73	47.57	12.69	36.10	50.45	33.17	41.9	77.62	90.48	33.16	61.27	81.70	64.36
70%	11.34	26.53	40.62	6.19	20.51	32.70	22.98	16.37	34.51	46.90	12.65	35.91	52.01	33.06	43.81	77.62	90.95	30.50	59.68	81.70	64.04
50%	10.61	27.17	40.71	5.79	19.87	32.20	22.73	16.59	34.73	47.57	12.51	35.35	51.01	32.96	40.48	78.57	90.48	31.03	62.60	80.90	64.01
30%	10.43	27.08	40.44	5.47	19.01	30.97	22.23	15.49	34.73	46.90	11.18	35.06	51.20	32.43	38.10	78.57	90.95	28.91	61.01	79.84	62.90
20%	10.16	25.62	39.98	4.75	18.38	30.35	21.54	15.27	34.29	46.68	10.05	35.49	50.87	32.11	41.43	75.71	88.57	30.50	61.01	80.37	62.93

D Simulation

Due to the high cost of manual annotation, we utilize only 427 samples from the full CoTalk dataset for our experiments. To supplement this, we simulate data generation using LLMs. Specifically, we apply the model to the entire DOTA training set, approximately three times larger than the original CoTalk dataset, to generate annotations following both the CoTalk and parallel annotation methods, as shown in Table 15. We then fine-tune the Long-CLIP-L model on the resulting datasets using three V100 GPUs, with a batch size of 16, a learning rate of 1e-6, and for 4 epochs.

We evaluate the models on remote sensing retrieval tasks, with the results summarized in Table 5. CoTalk achieves an average score of 40.62% across six tasks and three datasets, outperforming both the parallel method (40.05%) and the zero-shot baseline (37.64%). Notably, CoTalk surpasses both baselines in almost every task.

E Complete Experimental Results

The comprehensive experimental results, including both internal and external metrics, for the image-text retrieval downstream task evaluated on the Ret-3 dataset are summarized in table 8 and table 9.

F Prompt

F.1 Merge Annotation

From the theoretical framework, Assumption 2 concerns the role of the LLM. We assume effective integration should: (a) reduce redundancy, (b) accurately merge semantic units, and (c) maintain input–output consistency (Appendix C). In practice, Table 6, manual annotation experiments show 19.36% redundancy between the first and second annotators. After LLM integration with our prompts, this redundancy is reduced, as the combined output contains fewer semantic units than the sum of individual annotations—supporting (a). The merged results also show improved semantic quality, validating (b) and demonstrating the robustness of our prompt design.

Both CoTalk and parallel annotation require a large model to integrate individual annotations. This section provides a detailed analysis of the merging process. With the instructions in Table 10 and Table 11, we design specific prompts to guide the LLMs in merging image captions from both annotation methods, ultimately producing semantically rich descriptions.

F.2 Comparison of Talk and Text in Annotation Input

To verify that reading comprehension is faster than listening comprehension, we utilize a text-based LLM to generate questions according to fundamental facts. Participants answer these questions under both reading and listening conditions. The prompts, detailed in Table 12, are designed to generate high-quality questions with varying levels of difficulty, gradually incorporating finer-grained image content to ensure scientific rigor and experimental validity.

F.3 Derive the Semantic Units

For the intrinsic evaluation of image caption quality, we employ a text-based LLM to decompose image captions into their semantic units and assess them according to the number of semantic units identified. Following the procedure outlined in Table 13 and Table 14, we first perform necessary simplification and standardization of the combined text from Section F.1, then systematically split it to extract all the semantic units contained within the captions.

G Robustness to Annotator Variance

To verify whether our CoTalk pipeline can handle large-scale annotation with high inter-annotation difference, we simulate annotators of varying competence using Qwen2.5VL (32B). Models with different input resolutions (100%, 75%, and 50%) represent high, medium, and low competence, respectively. Following the CoTalk framework, these simulated annotators annotate sequentially, as shown in the table 7. Results indicate that even low-competence annotators can contribute effectively: the number of semantic units increases from 51.95 with high-competence annotators to 57.40 after adding medium-competence input, and further to 63.30 with the inclusion of low-competence annotations. The trend also intuitively holds when progressing from low- to high-competence annotators. This demonstrates that CoTalk enables efficient annotation even among annotators with significant capability differences. We will include this experiment in the final revised version.

H Annotation Interface

H.1 Annotation Interface 1: The first-person Annotation Interface of the CoTalk framework

The first annotator is tasked with generating detailed image captions that comprehensively describe the visual content. These captions are provided via voice input, transcribed into text, and then standardized using a large language model for storage. Each caption should aim to cover all identifiable entities in the image along with their attributes, including—but not limited to—absolute and relative positions, color, quantity, size, shape, and material. Details are shown in Table 15 (up).

H.2 Annotation Interface 2: The Suquential Subsequent Annotation Interface of CoTalk Framework

In CoTalk, subsequent annotators review the image and previously merged captions, generated by a large model, to identify omissions and correct errors. They contribute additional information through voice input, guided by a comprehensive understanding of both the image and the existing annotations.Details are shown in Table 15 (down).

I Detailed Experimental Process

I.1 Extrinsic Metric

We evaluate the practicality of each annotation method using extrinsic metrics that indirectly reflect annotation quality, employing a retrieval task for this purpose. Specifically, we fine-tune the Long-CLIP model on datasets annotated by CoTalk and a parallel method. Long-CLIP is selected for its extended input capacity of 248 tokens, nearly four times that of the original CLIP (77 tokens), making it well-suited for capturing detailed annotations.

We fine-tune the Long-CLIP-L model using three V100 GPUs, with a batch size of 16, a learning rate of 1e-6, and 12 epochs. Each epoch takes approximately 18 seconds. To ensure robust results, we run five trials with different random seeds per dataset and report the average performance.

After training, we evaluate the model on a remote sensing retrieval dataset. As shown in Table 8, CoTalk achieves 65.94% across six tasks and three datasets, outperforming the parallel method (65.01%) and the zero-shot baseline (62.54%).

Prompt for Merging Parallel Annotation

System Message:

Input:
You are a text integration expert. Here are the parallel annotations of two annotators. Please help me merge their annotations to form a summary annotation.

Guidelines:

- **Rule 1:** For parts with the same semantic meaning in caption1 and caption2, adopt a merging strategy and avoid repeating the same content.
- **Rule 2:** For parts unique to caption1 or caption2, incorporate them into the final description at an appropriate position.
- **Rule 3:** For parts where caption1 and caption2 contradict each other, select one caption's description for the consolidation, and do not include any reference to caption1 or caption2 in the description.
- **Rule 4:** During the merging process, avoid redundant mentions of caption1 and caption2. No intermediate reasoning is needed; just provide the final consolidated result.

Remember, your output should be a high-quality caption that is concise, informative, and coherent!

User:
Caption 1: `{first person annotation}`
Caption 2: `{parallel person annotation}`

Assistant Generation Prefix:
Here's the merged parallel caption:

Table 10: An Example implementation of the merging parallel annotation function σ_{merge} via prompting LLMs.

Prompt for Merging Sequential Annotation

System Message:

Input:
You are a text integration expert. Caption1 is the original annotation result, and Caption2 is the annotator's supplementation and correction.

Guidelines:

- **Rule 1:** Caption2 will include corrections to Caption1, possibly revising parts of the description in Caption1, as well as supplementing areas where Caption1's description was insufficient.
- **Rule 2:** For parts with the same semantic meaning in Caption1 and Caption2, adopt a merging strategy and avoid repeating the same content.
- **Rule 3:** For parts that are missing in Caption1 but present in Caption2, incorporate the relevant parts from Caption2 into Caption1 at an appropriate position.
- **Rule 4:** If there is a conflict between the descriptions in Caption1 and Caption2, prioritize the description in Caption2 and replace the corresponding part in Caption1.

Remember, your output should be a high-quality caption that is concise, informative, and coherent!

User:
Caption 1: `{first person annotation}`
Caption 2: `{sequential person annotation}`

Assistant Generation Prefix:
Here's the merged sequential caption:

Table 11: An Example implementation of the merging sequential annotation function σ_{merge} with prompting LLMs.

Prompt for faster input of text than talk problem acquisition

System Message:
Input:
You are an annotation question-generation assistant. Given a segment of annotation text, please design questions according to the following rules:

Guidelines:
- **Rule 1:** Generate a total of 5 questions.
- **Rule 2:** The five questions should cover the beginning, middle, and later parts of the annotation text.
- **Rule 3:** Design the questions in the order of the text and number them sequentially (Q1–Q5).
- **Rule 4:** The five questions should progress from general to detailed, starting with broad questions and moving to fine-grained ones.
- **Rule 5:** The questions can be about objects or their attributes (*e.g.,* color, quantity, location, shape, size, etc.).

Example:
- **Question 1:** What kind of image is this describing?
- **Question 2:** What color is the sea surface?
- **Question 3:** What's in the top left corner of the picture?
- **Question 4:** Where is the parking lot located in the picture?
- **Question 5:** Do all houses have swimming pools?

User:
Annotation Text: `{caption}`

Assistant Generation Prefix:
Here are the generated questions:

Table 12: An Example implementation of the generated questions via prompting LLMs.

Prompt for Denoising and simplifying annotations

Prompt for Denoising and simplifying annotations
System Message:
Input:
Please help me improve the following text according to the steps below.

Guidelines:
- **Rule 1:** Correct obvious types.
- **Rule 2:** Remove meaningless connecting words such as "then", "and", "furthermore" and "next".
- **Rule 3:** Format the output according to the sample provided.

User:
Annotation Text: `{merged caption}`

Assistant Generation Prefix:
Here's the processed caption:

Table 13: Prompt for Denoising and simplifying annotations.

Prompt for Deriving the minimal Semantic Units

Assistant Generation Prefix:
Here's the processed caption:

System Message:
Input:
Please help me extract and segment the semantic units according to the following rules and referring to the output example:

Guidelines:

- **Unit Definition:** Semantic unit = object name + associated attributes; A single sentence may contain multiple independent units; Each unit must contain only one object name.
- **Attribute Specifications:** Valid attributes: absolute_location (position in the overall image), relative_location(Position relative to other objects), colour, amount (Explicitly extract indefinite articles "a"/"an" as standalone attributes. Include numerical values *e.g.,* "two", "three" and quantifiers (*e.g.,* "some", "several")), size, shape, material, object description, other(All other unclassified attributes are 'other', If there are multiple, please separate them with commas),Omit any attributes that do not exist; Prohibit attribute overlap or duplication; Pronoun-based locations (*e.g.,* "this", "that") must be replaced with specific referenced objects.
- **Extraction Principles:** Prioritize extracting the "name" field separately; Create independent units for multiple objects sharing attributes; Absolute and relative locations cannot coexist in the same unit; Omit unspecified/ambiguous attributes.
- **Output Requirements:** Present only final results without reasoning processes.

Example:

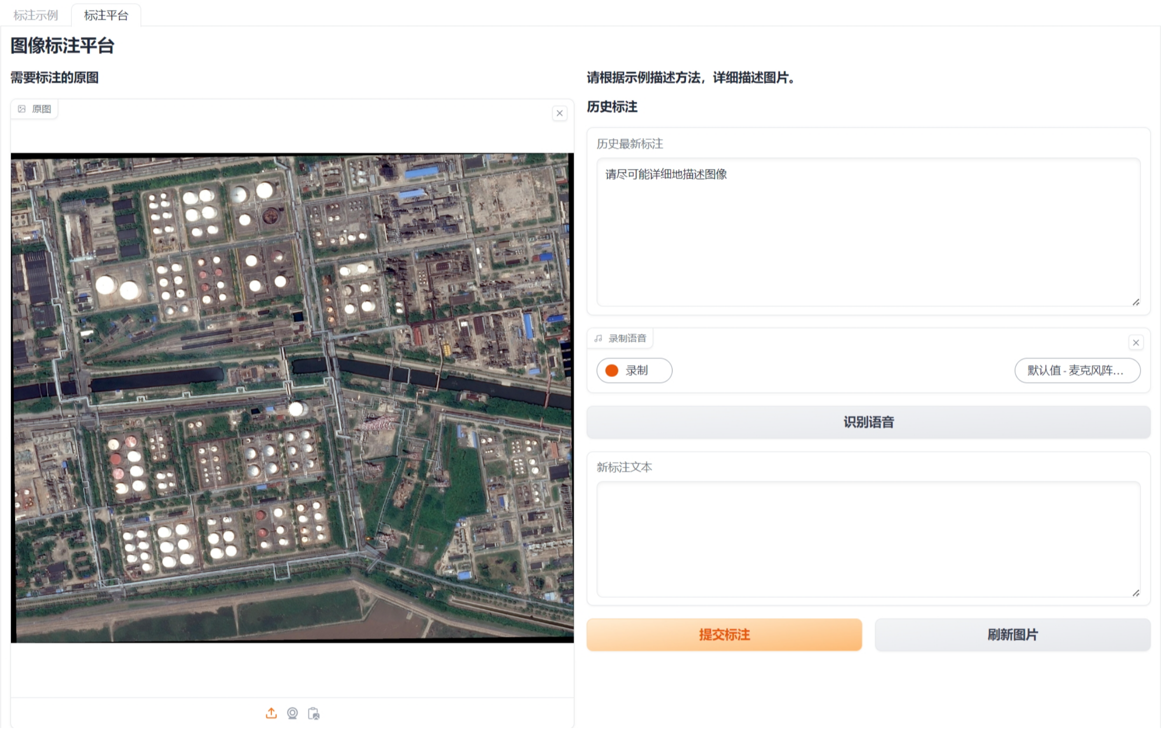
- **Input Example 1:** The sea surface appears green, with a patch of green seaweed visible under the bridge in the upper right area.
- **Output Example 1:**

```
[
    {
        "name": "sea surface",
        "attributes": {
            "colour": "green",
            "other": ["appears"]
        }
    },
    {
        "name": "seaweed",
        "attributes": {
            "amount": "a patch of",
            "colour": "green",
            "relative_location": "under the bridge in the upper right area",
            "other": ["visible"]
        }
    }
]
...
```

User:
Caption: `{processed caption}`

Assistant Generation Prefix:
Here are the Semantic Units:

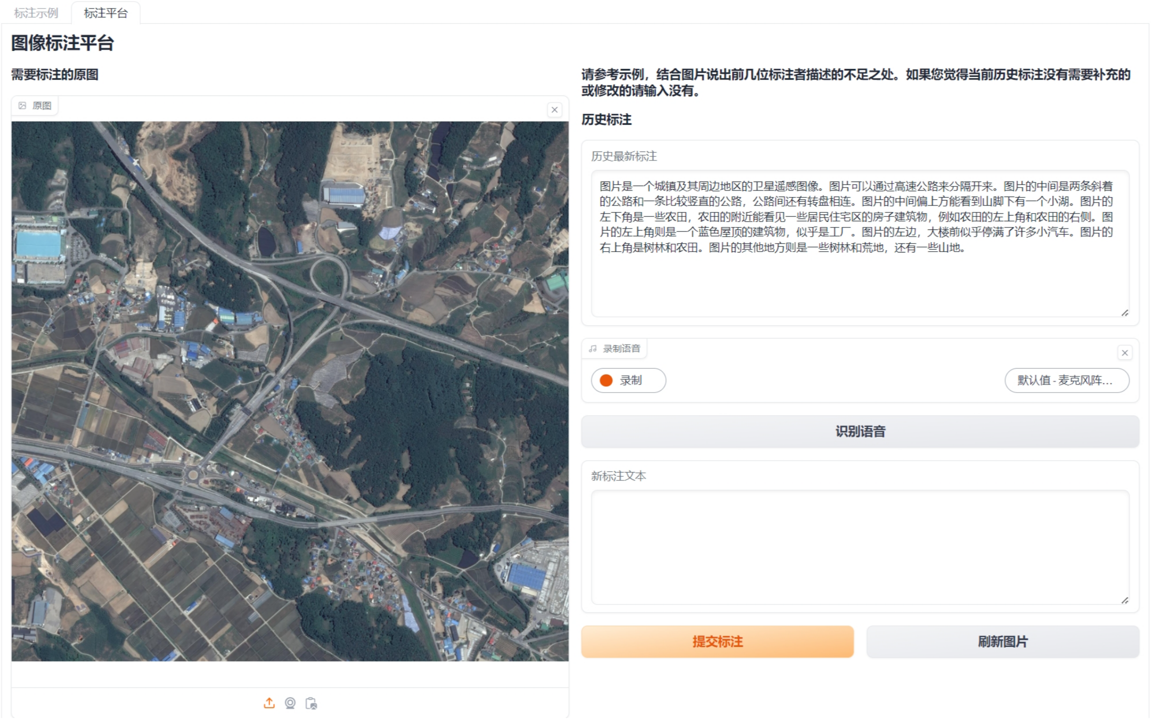
Table 14: Prompt for Deriving the minimal Semantic Units.


标注示例
标注平台
图像标注平台
需要标注的原图
原图
历史标注
历史最新标注
录制语音
录制
默认值 - 麦克风阵...
识别语音
新标注文本
提交标注
刷新图片

请根据示例描述方法，详细描述图片。

请尽可能详细地描述图像

Figure 8: The first-person annotation interface.


标注示例
标注平台
图像标注平台
需要标注的原图
原图
历史标注
历史最新标注
录制语音
录制
默认值 - 麦克风阵...
识别语音
新标注文本
提交标注
刷新图片

请参考示例，结合图片说出前几位标注者描述的不足之处。如果您觉得当前历史标注没有需要补充的或修改的请输入没有。

图片是一个城镇及其周边地区的卫星遥感图像。图片可以通过高速公路来分隔开来。图片的中间是两条斜着的公路和一条比较竖直的公路，公路间还有转盘相连。图片的中间偏上方能看到山脚下有一个小湖。图片的左下角是一些农田，农田的附近能看见一些居民住宅区的房子建筑物，例如农田的左上角和农田的右侧。图片的左上角则是一个蓝色屋顶的建筑物，似乎是工厂。图片的左边，大楼前似乎停满了许多小汽车。图片的右上角是树林和农田。图片的其他地方则是一些树林和荒地，还有一些山地。

Figure 9: The Suquential Subsequent Annotation Interface.

Guidelines for Image Annotation

First-Person:
System Message:
Input:
You will be provided with an image. Your task is to generate a detailed and informative caption for the image, adhering to the following guidelines:
Guidelines:

- **Rule 1:** The caption should be as comprehensive as possible. Identify and describe all discernible entities in the image along with their attributes. Attributes may include (but are not limited to): Absolute position (via image orientation), Relative position (in reference to other objects), Color, quantity, size, shape, material, etc. Avoid describing entities that cannot be clearly identified.
- **Rule 2:** Structure the caption in the following order: First: Begin with a global description of the entire image; Second, Provide a description of the objects located at the center of the image. Last, Describe the entire image systematically, starting from the upper-left corner and proceeding in a structured manner with the spatial relationships between objects.
- **Rule 3:** If there are more than 10 instances of a particular object type, leverage approximate quantifiers (*e.g.,* many, some, a row, a column, a cluster, etc.) instead of exact counts.
- **Rule 4:** Ensure the caption is concise but information-rich. Each sentence should contain meaningful and non-redundant information. Avoid vague, repetitive, or empty expressions.

Subsequent-Person:
System Message:
Input:
You will be provided with an image and its corresponding caption. Your task is to review and revise the caption to ensure it accurately and comprehensively reflects the content of the image, following the rules below:
Guidelines:

- **Rule 1:** Examine whether the caption includes all identifiable entities present in the image, along with their corresponding attributes. Attributes may include (but are not limited to): Absolute position (according to image orientation) Relative position (with respect to other objects) Color, quantity, size, shape, material, etc. If any entity is missing, add it along with its attributes. If any attribute of a described entity is missing, supplement it accordingly.
- **Rule 2:** If the caption contains any inaccuracies (*e.g.,* incorrect quantity, color, or other attributes of an entity), make the necessary corrections.
- **Rule 3:** Output only the revised caption. Do not include or refer to the original caption in your response.

Table 15: Guidelines for Image Annotation.

I.2 Consistency between Extrinsic and Intrinsic Metrics

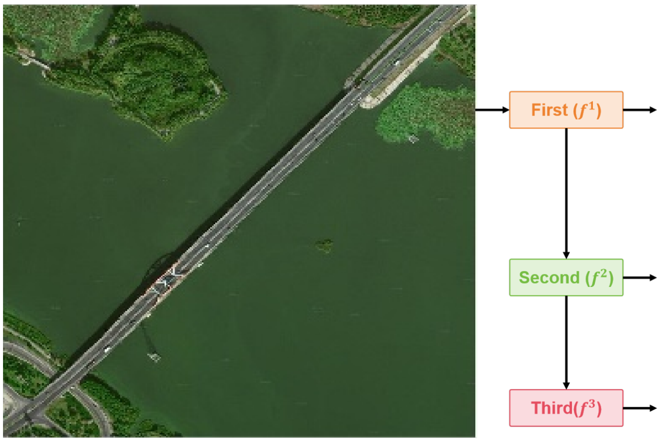
To verify the consistency between intrinsic indicators (*i.e.,* the number of semantic units) and extrinsic indicators (*i.e.,* downstream task performance), we reduce the number of semantic units in the CoTalk-annotated dataset of 429 images, at fixed ratios. Specifically, for each image, we randomly remove 20%, 30%, 50%, 70%, or 80% of its Semantic Units. For example, if an image originally has 10 Semantic Units and 20% are removed, 2 are randomly deleted, and the remaining 8 are merged into a single text input for subsequent fine-tuning of the Long-CLIP model.

We fine-tune Long-CLIP-L using three V100 GPUs with a batch size of 16, a learning rate of 1e-6, and 12 epochs. To ensure stability, we repeat each experiment using five different random seeds per dataset and report the average results.

We evaluate performance on the RSICD, RSITMD, and UCM-Captions datasets. As shown in Table 9, retrieval performance declines as fewer Semantic Units are retained, confirming the alignment between intrinsic and extrinsic indicators. This supports the use of Semantic Units as an effective measure of annotation quality and practical utility. Notably, when more than 50% of semantic units are removed, the performance drop becomes more pronounced, indicating the importance of detailed captions for vision-language alignment.

J CoTalk Examples

This section presents representative image samples manually annotated to demonstrate the labeling process. As illustrated in Figure 10, the examples span typical scenes: (1) bridges and adjacent suburban waters, (2) parking lots and their environments, (3) port piers with coastal landscapes. These samples reflect both the annotation quality and the integration of large model predictions.



Semantic Units: 21

这是一张桥梁的卫星遥感图像。图像基本都是海面，海面呈绿色。从图像右上角到图像左下角有一座高架桥穿过。图片左上角是一个绿色的小岛 上面有许多绿树，在图像右上角的桥下面还有一片绿色的海藻。图片左下角的陆地上还有一些弧形的马路。

This is a satellite remote sensing image of the bridge. The images are basically sea surface, and the sea surface is green. There is a viaduct from the upper right corner of the image to the lower left corner of the image. In the upper left corner of the picture is a green island with many green trees on it, and there is a piece of green seaweed under the bridge in the upper right corner of the picture. There are also some curved roads on the land in the lower left corner of the picture.

Semantic Units : 18

在桥面上能看见一些小汽车在上面行驶，在图像中心偏左下的桥面上还有红色的桥廊。小汽车的颜色有白色的、黑色的。左上角的小岛上除了绿树还有大片的绿色草地。

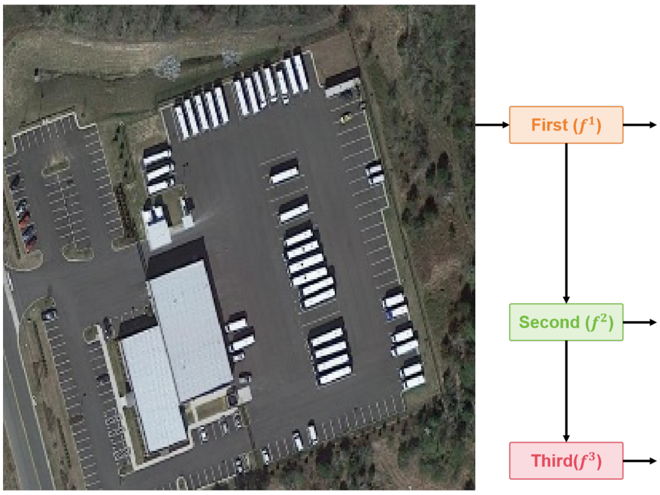
Some cars can be seen driving on the bridge deck, and there is a red bridge corridor on the bridge deck at the lower left of the image center. The colors of cars are white and black. In the upper left corner of the island, there is a large area of green grass besides green trees.

Semantic Units : 0

没有。

None.

Merged(σ): This is a satellite remote sensing image of the bridge. The sea surface, which is green, dominates the image. A viaduct stretches from the upper right corner to the lower left corner. In the upper left corner, there is a green island covered with green trees and a large area of green grass. Under the bridge in the upper right corner, there is a piece of green seaweed. Some cars, which are white and black, can be seen driving on the bridge deck that features a red bridge corridor at the lower left of the image center. Curved roads are also visible on the land in the lower left corner of the picture.



Semantic Units: 49

这是一张停车场区域的图像。在图像中间是一个画有白线的停车场区域，里面停放着许多白色的大卡车。在这个停车场的右上角还有一辆黄色的小汽车和一辆黑色的小汽车。在这个区域的左下角有一栋白色屋顶的房屋。这个白色屋顶房屋的左下角紧挨着一栋白色屋顶的长方形建筑。这个长方形建筑的右下方有三辆灰色小汽车和两辆白色小汽车。这个建筑的左边和上方都有两个小型停车区域，其中左边停着两辆黑色小汽车，上方的停车区域里面停放着五辆小汽车，有白色、红色和黑色的。图像的剩余地方都是草地和绿树。

This is an image of the parking area. In the middle of the image is a parking area with white lines, where many large white trucks are parked. There is also a yellow car and a black car in the upper right corner of this parking lot. There is a house with a white roof in the lower left corner of this area. The lower left corner of this white-roofed house is next to a rectangular building with a white roof. There are three gray cars and two white cars on the lower right of this rectangular building. There are two small parking areas on the left and above the building, two black cars are parked on the left, and five cars, white, red and black, are parked in the parking area above. The rest of the image is covered with grass and green trees.

Semantic Units : 8

图片中停车场四周有一些路灯在上面，还能看见一些白色的实现，划分停车场的停车区域。图片的左下角还有一条马路。

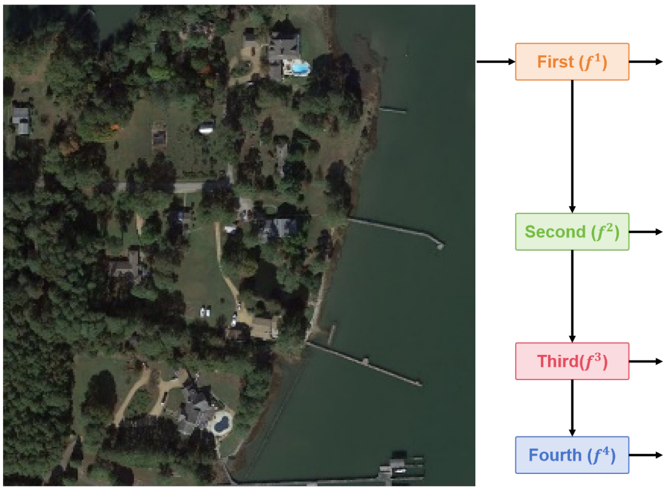
In the picture, there are some street lamps around the parking lot, and you can also see some white realization to divide the parking area of the parking lot. There is also a road in the lower left corner of the picture.

Semantic Units : 0

没有。

None.

Merged(σ): This is an image of the parking area with white lines dividing the spaces. In the middle, many large white trucks are parked. On the upper right corner, there is a yellow car and a black car. A house with a white roof is located in the lower left corner, next to a rectangular building with a similar roof. Around the parking lot, there are street lamps. Three gray cars and two white cars are parked on the lower right of this rectangular building. Two small parking areas are on the left and above the building; two black cars are on the left, and five cars (white, red, and black) are parked above. The lower left corner also features a road. The rest of the image is covered with grass and green trees.



Semantic Units: 18

这是一个港口海岸的遥感卫星图像。在图像的左侧，一半面积是海岛，岛上分布着绿色的树林和草地，草地上有一些较少的房屋。在这个岛的右侧是港口，港口上有白色的船，图片右侧的深色区域是海水。

This is a remote sensing satellite image of the port coast. On the left side of the image, half of the area is an island. There are green trees and grass on the island, and there are some less houses on the grass. On the right side of the island is the port, where there are white boats, and the dark area on the right side of the picture is seawater.

Semantic Units : 12

港口上有1艘白色的船，他们的大小不一。在图像的左侧海岛上，有一些房屋的院子都带有小型的蓝色泳池。同时在图片的中心还能看到几艘白色的船在陆地的草地上。

There is a white boat in the harbor. They are of different sizes. On the island on the left side of the image, there are some houses with small blue swimming pools in their yards. At the same time, several white boats can be seen in the center of the picture on the land grass.

Semantic Units : 9

图片中陆地的房屋屋顶大多是棕色和灰色的，还能看见一些小汽车在房屋的周围，小汽车的颜色有红色的、白色的、灰色的。

In the picture, the roofs of houses on land are mostly brown and gray, and some cars can be seen around the houses. The colors of cars are red, white and gray.

Semantic Units :

没有。

None.

Merged(σ): This is a remote sensing satellite image of the port coast. On the left side of the image, there is an island with green trees, grass, some houses with brown and gray roofs, and small blue swimming pools in their yards. Around the houses, some cars in red, white, and gray colors can be seen. On the right side of the island is the port where white boats of different sizes are located. In the center of the picture on the land grass, several white boats can be seen. The dark area on the right side of the picture is seawater.

Figure 10: Examples of image annotation of bridge periphery (top), parking lot periphery (middle) and port coast (bottom).