Boosting Multi-modal Keyphrase Prediction with Dynamic Chain-of-Thought in Vision-Language Models

Anonymous ACL submission

Multi-modal keyphrase prediction (MMKP) aims to advance beyond text-only methods by incorporating multiple modalities of input information to produce a set of conclusive phrases. Traditional multi-modal approaches have been proven to have significant limitations in handling the challenging absence and unseen scenarios. Additionally, we identify shortcomings in existing benchmarks that overestimate model capability due to significant overlap in training tests. In this work, we propose leveraging vision-language models (VLMs) for the MMKP task. Firstly, we use two widely-used strategies, *e.g.*, zero-shot and supervised fine-tuning (SFT) to assess the lower bound performance of VLMs. Next, to improve the complex reasoning capabilities of VLMs, we adopt Fine-tune-CoT, which leverages high-quality CoT reasoning data generated by a teacher model to finetune smaller models. Finally, to address the "overthinking" phenomenon, we propose a dynamic CoT strategy which adaptively injects CoT data during training, allowing the model to flexibly leverage its reasoning capabilities during the inference stage. We evaluate the proposed strategies on various datasets and the experimental results demonstrate the effectiveness of the proposed approaches. The code and datasets will be made publicly available upon acceptance of the paper.

1 Introduction

Multi-modal keyphrase prediction (MMKP) aims to generate concise, informative phrases that capture the essence of cross-modal inputs (*e.g.*, text and image inputs in Fig. 1(a)). Unlike traditional text-only keyphrase prediction (Zhang et al., 2018; Wang et al., 2019b; Zhang et al., 2024a), MMKP leverages the complementary nature of visual and textual signals to improve the cross-modal semantic understanding (Chang et al., 2013; Bansal et al., 2015; Wang et al., 2019a; Zhang et al., 2024a) and summarization (Davidov et al., 2010; Wang et al., 2011; Zhang et al., 2024b).

Traditional multi-modal approaches (Wang et al., 2020; Dong et al., 2023) primarily focus on designing cross-modal fusion architectures to integrate visual features (*e.g.*, OCR and visual entities) and textual semantics, by using attention mechanisms or hybrid neural networks to model modality interactions. These methods have achieved great progress in the MMKP task.

However, as illustrated in Fig. 1(b), significant limitations have emerged in more complex scenarios, particularly in handling the following two challenges:

- *absence scenario*. The case where the prediction keyphrases are lacking in the corresponding input text, requiring the model to exhibit strong cross-modal interaction capabilities and infer keyphrases from cross-modal context.

- *unseen scenario*: The case where the predicted keyphrases do not appear in the training set, demands that the model possess robust generalization capability. The unseen keyphrases pose a significant challenge to MMKP systems in production environments, where a diverse range of keyphrases emerges daily.

In addition, our analysis of the public MMKP (Wang et al., 2020) dataset and our collected larger-scale (MMKP-360k) production dataset reveal that the two datasets exhibit substantial distribution discrepancies in unseen keyphrase scenarios. The proportion of training samples with unseen keyphrases is only 2.68% in the public dataset and 54.72% in the real-word production dataset. The discrepancy in Fig. 1(b) and Fig. 1(c), indicates that traditional multi-modal approaches exhibit poor generalization capability in absent and unseen scenarios.

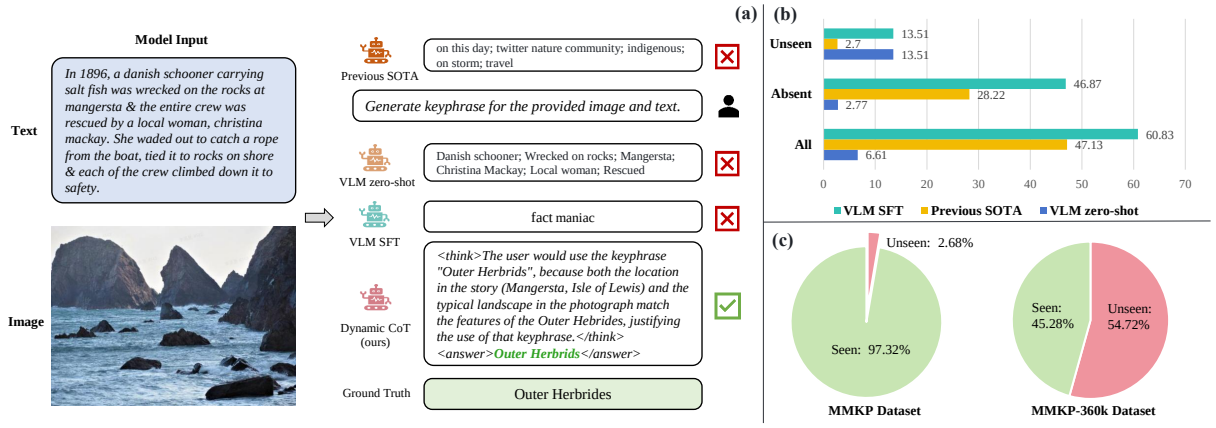


Figure 1: (a) An example of multi-model keyphrase prediction. (b) The performance of different models on the MMKP dataset (Wang et al., 2020). "Absent" refers to keyphrases that absent in the input text. "Unseen" refers to keyphrases that not appear in the training set's ground truth. (c) The number of seen and unseen keyphrases in the test set of the MMKP dataset and our MMKP-360k dataset.

Although VLMs have been widely applied to multi-modal tasks, *e.g.*, visual question answering (Antol et al., 2015), image captioning (Li et al., 2022), video understanding (Sigurdsson et al., 2016), their effectiveness in the MMKP remains underexplored. To this end, we propose adopting VLMs for MMKP in an autoregressive manner. Firstly, we use two widely-used strategies: zero-shot and supervised fine-tuning (SFT). As illustrated in Fig. 1(b), the SFT approach outperforms the zero-shot approach in the absent scenario but underperforms in the unseen scenario, which indicates that SFT enables the VLMs to leverage its robust vision-language comprehension for similar content, but severely restricts its generalization capability. To solve this, we follow Ho et al. (2023) to utilize Fine-tune-CoT to improve the complex reasoning capabilities of VLMs. Fine-tune-CoT leverages high-quality CoT reasoning data generated by a teacher model to finetune smaller models. Furthermore, considering the "overthinking" phenomenon (Chen et al., 2024a) for the seen scenario, we propose a dynamic CoT strategy to enable efficient reasoning (Qu et al., 2025) for the VLMs. The dynamic CoT strategy enables the VLMs to prefer to choose the non-CoT reasoning for the easy samples (*e.g.*, seen samples).

To ensure the reproducibility of our research, we resampled the public MMKP dataset (MMKP-V2) to match the proportion of seen and unseen parts with that of the MMKP-360k dataset. Furthermore, comprehensive analysis on three datasets confirm that our method significantly improves the generalization capability of VLMs on unseen samples.

The contributions are summarized as follows:

- To the best of our knowledge, this work is the first to comprehensively investigate the potential of VLMs for multi-modal keyphrase prediction.
- We propose a Dynamic CoT strategy that enbales VLMs adaptively choosing CoT reasoning ability for the hard unseen samples, which is more suitable in production environments with efficient decoding.
- Experimental results and rigorous analysis across multiple datasets validate the efficacy and robustness of our proposed methodology.

2 Related Work

2.1 Social Media Keyphrase Prediction

Social media keyphrases, including hashtags and categories, serve as concise summaries of user-generated content. Prior to the emergence of LLMs, approaches in this domain mainly fell into extractive (Zhang et al., 2016, 2018), classification (Zhang et al., 2017; Kou et al., 2018; Zeng et al., 2018), and generative methods (Wang et al., 2019b,a; Kou et al., 2018). Due to inherent training limitations, the first two types could not produce keyphrases from a truly open set, while generative methods were limited to processing text-only content. With the advent of LLMs, numerous methods (Shao et al., 2024; Zhang et al., 2024a; Kang and Shin, 2025) have attempted to leverage these models for keyphrases prediction, but most still rely exclusively on textual inputs. However, social media posts often contain multi-modal information, thus requiring the model to possess strong multi-modal

understanding capabilities. NoteLLM2 (Zhang et al., 2024b) uses MLLM and a zero-shot prompt to compress multi-modal posts into a single word vector for end-to-end recommendation model training. However, it does not explore how to generate more comprehensive and accurate keyphrases. In this paper, we investigate the potential of VLMs for keyphrase prediction, fully leveraging multi-modal information to achieve more precise and accurate keyphrase prediction.

2.2 Vision-Language Models

Vision-Language Models (VLMs) have emerged as a transformative paradigm in multi-modal learning, bridging visual and textual representations to enable cross-modal understanding and generation (Alayrac et al., 2022; Li et al., 2023; OpenAI, 2023; Bai et al., 2023). Early efforts in this domain focused on joint embedding spaces (Radford et al., 2021; Jia et al., 2021), achieving strong zero-shot transfer capabilities. Subsequent work expanded VLMs to generative tasks, such as image captioning (Li et al., 2022) and visual question answering (VQA) (Antol et al., 2015), leveraging architectures like transformer-based encoders-decoders (Vaswani et al., 2017). Models such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) integrated pretrained vision encoders with LLMs to unify perception and reasoning, enabling few-shot adaptation to downstream tasks. Later advancements, including GPT-4V (OpenAI, 2023), Qwen-VL (Bai et al., 2023) series, Intern-VL (Chen et al., 2024b) series and etc, further scaled training data and model size, demonstrating remarkable performance on complex multi-modal benchmarks.

2.3 Reasoning Capabilities

Recently, as reasoning models (Jaech et al., 2024; Guo et al., 2025) have gained significant attention, inference-time computation has been recognized as an effective approach to further unlock the potential of LLMs. Consequently, an increasing number of studies (Xu et al., 2024; Team et al., 2025; Seed et al., 2025) have started to incorporate reasoning capabilities into VLMs. In this paper, we examine multiple VLM training paradigms for multi-modal keyphrasse prediction. By integrating world knowledge and reasoning abilities through Dynamic CoT training, our approach enhances model generalization while mitigate the problem of "overthinking" (Chen et al., 2024a).

3 Methodology

In this section, we first discuss paradigms of traditional multi-modal models for MMKP and briefly analyze their limitations (Sec 3.1). Next, we focus on how to incorporate reasoning capabilities into VLMs (Sec 3.2), and describe how our proposed approach employs Dynamic CoT to solve the "overthinking" phenomenon (Sec 3.3).

3.1 Traditional Multi-modal Models

Traditional multi-modal models such as M^3H-ATT (Wang et al., 2020) and MM-MKP (Dong et al., 2023) are inherently constrained by limited model capacity and insufficient multi-task capabilities. Consequently, such models typically depend on external Optical Character Recognition (OCR) systems and visual feature extraction modules to augment textual inputs from social media posts. The extracted visual features are concatenated with post text to enhance keyphrase prediction performance. To further improve keyphrase prediction accuracy, these approaches frequently incorporate additional keyphrase classification modules based on all keyphrases in the dataset. The final optimization objective is formulated as a multi-task loss, which is defined as follows:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}[logP_{cls}(\mathbf{y}^n) + \gamma \cdot \sum_{t=1}^{l_y^n} logP_{gen}(\mathbf{y}_t^n)], \quad (1)$$

where the first term represents the classification loss, while the second term corresponds to the keyword generation loss. Here, N denotes the size of the training set, $\mathbf{y}$ is the predicted keyphrase sequence, and t indicates the t-th token.

However, this strategy constrains the models' open-set generation capabilities and limits their generalizability. Moreover, the scope of world knowledge embedded in such models remains inadequate. For certain social media posts, such as those involving memes or referencing current events and political topics, a substantial amount of external world knowledge is necessary, which poses significant challenges for traditional multi-modal models.

3.2 Endow Reasoning Capabilities

Compared to traditional multi-modal models, VLMs offer superior capabilities in image-text understanding and generalization. VLMs can effectively comprehend the content of social media posts without relying on external models. For the MMKP

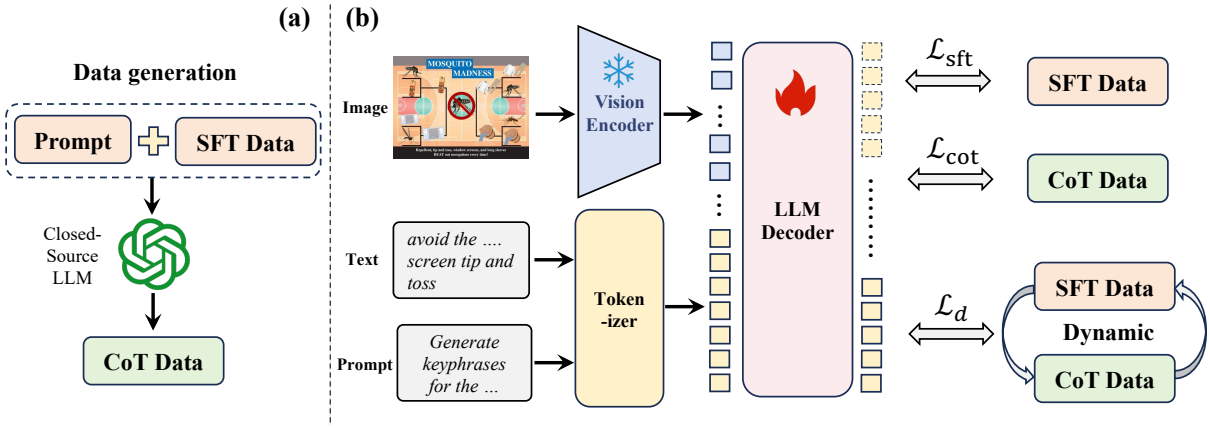


Figure 2: Main framework of our proposed method. (a) CoT data production pipeline. (b) Dynamic CoT training pipeline.

Prompt for Generating CoT data
[INST]<SYS> You are a helpful assistant. Analyze briefly why social media users would use specific hashtags "Keyphrases" for a post titled "Post text" with given image "Image".</SYS> <USER> "Keyphrases": {keyphrases} "Post text": {post text} "Image": {image} </USER>[INST]

Table 1: The system prompt template for generating CoT responses. "[INST]" denotes the instruction provided to the LLM, <SYS> denotes the system prompt and <USER> denotes the user prompt.

task, a straightforward approach to training VLMs involves using the multi-modal content as the input prompt and the ground truth (GT) keyphrases as the response, where the model is optimized using the next-token prediction loss, as shown in Fig. 2(b). The loss function is defined as follows:

$$\mathcal{L}_{sft} = -\frac{1}{T}\sum_{t=1}^{T} \log P\left(\mathbf{y}_t^s \mid \mathbf{y}_{<t}^s, \mathbf{v}; \theta\right), \tag{2}$$

here $\mathbf{y}^s = [\mathbf{y}_p; \mathbf{y}_r^s]$, where $\mathbf{y}_p$ and $\mathbf{y}_r^s$ denote the input prompt token sequence and the response token sequence, respectively. The response token sequence $\mathbf{y}_r^s$ corresponds to the GT keyphrases. $\mathbf{v}$ denotes the image token sequence, and θ denotes the model parameters.

Although straightforward SFT demonstrates promising performance on VLMs, it still suffers from several limitations. As shown in Fig. 1(b), the model's generalization is constrained, with its performance on unseen keyphrases nearly matching that of zero-shot scenarios (see Section 4.4 for detailed analysis). Our analysis reveals that supervised fine-tuned (SFT) models exhibit a strong tendency to overfit training samples, generating keyphrases primarily through surface-level pattern matching. Crucially, these models demonstrate limited capability in deciphering the underlying user intent embedded in provided keyphrases, consequently lacking the necessary reasoning capacity to infer contextually appropriate keyphrases.

To overcome these limitations, we constructed multi-modal CoT data, aiming to activate and supplement the model's world knowledge and thereby strengthen its reasoning ability. Specifically, we leveraged GPT-4o (Hurst et al., 2024) to generate multi-modal CoT data, wherein each thought process centers on analyzing user intent to enhance the model's reasoning capacity while maintaining relevance learning between similar posts. Fig. 2(a) illustrates the pipeline for constructing CoT data, and the prompt template is shown in Table 1.

After obtaining such analytic data, the final multi-modal CoT data is organized in the form of "`<think>thinking process</think> <answer>keyphrases</answer>`". The Fine-tune-CoT loss function is formulated as follows:

$$\mathcal{L}_{cot} = -\frac{1}{T}\sum_{t=1}^{T} \log P\left(\mathbf{y}_t^c \mid \mathbf{y}_{<t}^c, \mathbf{v}; \theta\right), \tag{3}$$

where $\mathbf{y}^c = [\mathbf{y}_p; \mathbf{y}_r^c]$ and $\mathbf{y}_r^c$ corresponds to the CoT responses.

However, experimental results indicate that the incorporation of CoT data does not yield immediate performance gains. Additionally, CoT reasoning introduces extra computational overhead during inference.

3.3 Dynamic CoT

Our analysis reveals two key limitations in Fine-tune-CoT models. First, the "overthinking" phenomenon occurs when the reasoning model gener-

Datasets	# Train Posts	# Test Posts	# KP / Post	Train |KP| ∩ Test |KP|
MMKP	42,959	5,372	1.33	97.32%
MMKP-V2	34,515	10,564	1.29	44.92%
MMKP-360k	330,614	36,736	4.48	45.28%

Table 2: Statistics of different datasets. KP: keyphrase; Train |KP|: the size of unique keyphrase in train set. ∩ denotes the intersection of the two sets.

ates overly generic keyphrases that fail to capture users' specific preferences. Second, we observe content redundancy in multi-modal CoT generation, where posts sharing identical keyphrases receive highly similar reasoning paths. This redundancy issue becomes particularly severe for high-frequency keyphrases, significantly diminishing model effectiveness.

To more effectively leverage multi-modal CoT data, we propose a dynamic CoT training paradigm, as illustrated in Fig. 2(b).

Specifically, during training, we categorize samples as easy or hard based on SFT loss $\mathcal{L}_{sft}$. We hypothesize that overfitting to simple samples during SFT may significantly impair model generalization capability. To mitigate this issue, we introduce a threshold γ, when the loss of a sample falls below γ, we switch its supervision to CoT data. The mathematical formula is given as follows:

$$\mathcal{L}_d = -\frac{1}{T}\sum_{t=1}^{T}\log P\left(\mathbf{y}_t^d \mid \mathbf{y}_{<t}^d, \mathbf{v}; \theta\right), \tag{4}$$

where

$$\mathbf{y}^d = \begin{cases} \mathbf{y}^c & \mathcal{L}_{sft} < \gamma \\ \mathbf{y}^s & \mathcal{L}_{sft} \geq \gamma \end{cases}. \tag{5}$$

This dynamic adjustment of the supervision signal enables the model to adapt its output format based on input complexity, thereby further enhancing generalization while maintaining robust relevance learning.

4 Experiments

4.1 Datasets

MMKP Dataset. Multi-modal Keyphrase Prediction (MMKP) Dataset was collected by (Wang et al., 2020). This dataset includes 53,701 English samples, each of which comprises a distinct text-image pair, with user-annotated hashtags serving as keyphrases. The dataset has a diverse set of categories and only around half of the images (54%) are natural photos, which is rather different from other

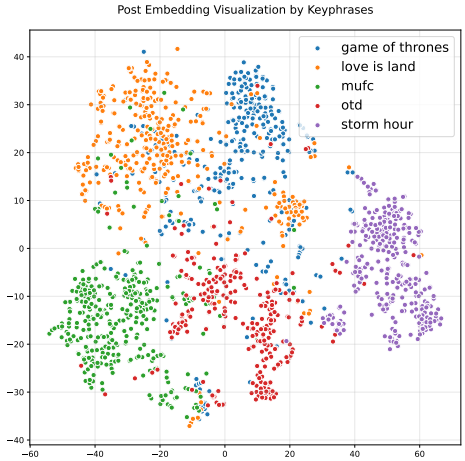


Figure 3: Visualization of multi-modal embedding clustering for post sharing the same keyphrase (The top five most frequent keyphrases) in the MMKP dataset.

standard image data such as MS-COCO (Lin et al., 2014). Furthermore, 52% of the samples contain either semantically uninformative text or irrelevant images, indicating a complex and often discordant multimodal relationship. The authors randomly split the data into 80%, 10%, 10% corresponding to training, validation, and test set. As shown in Table 2, there are 4,261 unique keyphrases in the training set and 2,534 unique keyphrases in the test set, among which 2,466 keyphrases from the test set also appear in the training set, resulting in a high overlap rate of 97.32%.

MMKP-360k Dataset. To better evaluate the effectiveness of our proposed method, we constructed a larger-scale multi-modal keyphrase prediction dataset collected based on user-generated contents publicly available on internet platforms. Following the construction methodology of the MMKP dataset, we extracted the users' hashtag data, which was subsequently cleaned and refined using LLMs. The processed hashtags served as the final ground truth keyphrases. The resulting MMKP-360k Dataset comprises 330,614 training samples and 36,736 test samples, each of which comprises a distinct text-image pair. There are 502k unique keyphrases in the training set and 81k unique keyphrases in the test set, of which 37k keyphrases from the test set also occur in the training set, resulting in an overlap rate of 45.28%.

Resampled MMKP Dataset (MMKP-V2). According to the keyphrase statistics for the MMKP dataset, the vast majority of keyphrases in the test set also appear in the training set. This enables models to significantly improve accuracy on the

Models	Training	MMKP	MMKP-V2			MMKP-360k			Avg
		All	All	Absent	Unseen	All	Absent	Unseen	All
Baseline Models									
CO-ATT (Zhang et al., 2017)		42.12	-	-	-	-	-	-	-
M^3H-ATT (Wang et al., 2020)		47.06	-	-	-	-	-	-	-
MM-MKP (Dong et al., 2023)		48.19	-	-	-	-	-	-	-
Our Experiments									
InternVL3-2B	zero-shot	3.62	4.56	1.39	4.53	19.00	5.64	9.62	9.06
	full sft	57.54	28.89	18.10	7.98	38.53	18.47	24.42	41.65
	Dynamic CoT	**59.63**	**30.76**	**18.99**	**9.90**	**40.03**	**19.95**	**26.04**	**43.47**
InternVL3-8B	zero-shot	6.26	6.89	2.37	6.72	11.43	3.26	7.05	8.19
	full sft	57.83	28.58	18.30	7.17	40.48	19.88	25.45	42.30
	Dynamic CoT	**60.29**	**31.42**	**19.13**	**10.68**	**50.53**	**20.04**	**26.44**	**47.41**
Qwen2.5-VL-3B	zero-shot	4.48	4.50	1.50	4.37	17.08	4.93	11.29	8.49
	full sft	60.33	29.89	19.59	8.79	43.04	22.06	24.60	44.42
	Dynamic CoT	**61.90**	**33.14**	**20.19**	**12.48**	**47.51**	**22.62**	**26.68**	**47.52**
Qwen2.5-VL-7B	zero-shot	6.61	7.75	2.75	8.38	14.34	4.10	9.94	9.57
	full sft	60.83	30.49	20.90	7.90	43.70	22.28	24.98	45.01
	Dynamic CoT	**63.58**	**33.56**	**22.32**	**13.36**	**50.66**	**23.41**	**26.43**	**49.27**

Table 3: Performance comparison for multi-modal keyphrase prediction task. We adopt F1@1 (%) as the evaluation metric for the MMKP and MMKP-V2 datasets, while F1@M (%) is employed for the MMKP-360k Dataset.

test set by simply fitting the training data and learning the similarity among posts containing identical keyphrases. However, statistics from our collected MMKP-360K data indicate that the overlap rate is only 45.28%, and as time progresses, an increasing number of new keyphrases are being created by users. This suggests that models require stronger generalization and reasoning capabilities in order to more accurately predict keyphrases for posts that express previously unseen main ideas.

To enhance alignment with real-world data distributions, we reconstructed the MMKP dataset through two key modifications: 1. Transferring all training-exclusive keyphrase samples to the test set. 2. Removing test samples containing keyphrases observed during training. The resulting MMKP-V2 dataset contains 34,515 samples in the training set and 10,564 samples in the test set, with 2,455 and 3,297 unique keyphrases respectively. Notably, 1,481 test keyphrases from the test set also appear in the training set, resulting in an overlap rate of 44.92%.

4.2 Experimental Setup

To ensure a fair comparison, all models were trained with identical hyperparameter configurations on both the MMKP and MMKP-V2 datasets. Specifically, we employed the AdamW optimizer with an initial learning rate of 5×10^{-5}, using a cosine annealing schedule for learning rate adjustment. During the SFT process, the parameters of the visual module were frozen, and only the multimodal projector as well as the large language model components were fine-tuned. The batch size was consistently set to 1 across all experiments. Models with 2B or 3B parameters were trained for 5 epochs, while those with larger parameter sizes were uniformly trained for 3 epochs. Given that the average number of keyphrases per post is approximately 1.3, F1@1 is adopted as evaluation metric. We utilized the GPT-4o-2024-05-13 (Hurst et al., 2024) to generate CoT reasoning data. The Dynamic CoT loss threshold γ is set to 0.4 for all the models. For the MMKP-360k dataset, we employed the AdamW optimizer with an initial learning rate of 3×10^{-5}. All models were trained for 3 epochs. We utilized the Doubao-1.5-pro (Seed, 2025) to generate CoT reasoning data. We adopt F1@M as evaluation metrics, where M denotes the number of keyphrases predicted by the model.

To validate the universality of our approach, we performed experiments across multiple LLMs and VLMs, including Llama-3.2 (Meta, 2024), Qwen2.5 (Yang et al., 2024), Llama-3.2-Vision (Meta, 2024), InternVL-3 (Zhu et al., 2025) and Qwen2.5-VL (Bai et al., 2025).

Models	MMKP			MMKP-V2			Avg			
	All	Seen	Unseen	All	Seen	Unseen	All	Δ (%)	Unseen	Δ (%)
Qwen2.5-VL-3B	4.48	5.26	12.16	4.50	4.42	4.37	4.49	-	8.27	-
+ SFT	60.33	61.26	12.16	29.89	55.88	8.79	45.11	-	10.48	-
+ Fine-tune-CoT	56.99	57.94	9.46	31.88	53.79	**13.57**	44.44	↓ 1.49	11.52	↑ 9.92
+ Multi-task	60.87	61.75	9.46	31.53	**57.96**	10.26	46.20	↑ 2.42	9.86	↓ 5.92
+ Dynamic CoT	**61.27**	**61.83**	**14.87**	**33.14**	57.14	12.48	**47.21**	↑ 4.66	**13.68**	↑ 30.53
Qwen2.5-VL-7B	6.61	8.70	13.51	7.75	11.22	8.38	7.18	-	10.95	-
+ SFT	60.83	61.60	**13.51**	30.49	58.45	7.90	45.66	-	10.71	-
+ Fine-tune-CoT	61.97	62.55	12.16	33.53	57.46	**13.42**	47.75	↑ 4.58	12.79	↑ 19.42
+ Multi-task	62.29	63.09	**13.51**	31.87	**59.74**	9.48	47.08	↑ 3.11	11.50	↑ 7.38
+ Dynamic CoT	**63.58**	**64.22**	**13.51**	**33.56**	58.56	12.24	**48.57**	↑ 6.37	**12.89**	↑ 20.35

Table 4: Performance comparison for different training strategy. In the multi-task setting, we treat CoT data as an additional training objective, while keeping the number of training steps consistent with other methods.

4.3 Comparing with SOTA methods

Table 3 compares the performance of our approach with baseline models. The table is organized vertically into two categories: baseline traditional multi-modal models and VLMs. Horizontally according to results on the MMKP, MMKP-V2 and MMKP-360k datasets. Detailed experimental results for additional text-only models can be found in the Appendix.

As shown in Table 3, the SFT VLMs outperform state-of-the-art multi-modal keyphrase prediction models (*e.g.*, M^{3}H-ATT(Wang et al., 2020) and MM-MKP (Dong et al., 2023)) by over 20%. These results suggest that VLMs, which possess broader world knowledge, offer a higher upper bound for multi-modal keyphrase prediction tasks and are a preferable choice for such applications. Furthermore, our method achieves consistent improvements over zero-shot and SFT across various datasets and baselines, and shows significant gains with respect to unseen keyphrases, highlighting the robustness and generalizability of our approach.

4.4 Ablation Study

In this section, we present extensive ablation studies with detailed results summarized in Table 4. We provide a thorough analysis of the findings and demonstrate that our proposed method effectively enhances the generalization capability of the model.

Why is simple SFT so effective, while Fine-tune-CoT leads to a drop in performance? We believe there are two main reasons for this. First, in MMKP task, *effectively learning inter-posts relevance plays a critical role*. This task to some extent reflects the commonalities of group behavior, users tend to select hashtags that resonate with the group to help the post spread better within the community. By analyzing training set posts that share the same keyphrase as shown in Fig. 3, we observe that their content often exhibits notable similarity. Furthermore, many of these keyphrases serve as abstractions that encapsulate the essence of such posts, rather than functioning as straightforward, literal summaries. Additionally, *Fine-tune-CoT model has "overthinking" phenomenon*. Although Fine-tune-CoT model is able to infer keyphrases that semantically correspond to the post content, the generated keyphrases tend not to match those commonly used by social media users, deviating from typical user preferences, for example, generating "indiana weather" instead of the more user-preferred "in wx", as shown in Fig. 4 Post (a).

Second, *significant training-test overlap overestimate the model's capability*. According to the previous analysis, keyphrases in the MMKP dataset have a high overlap between the training and test sets. The model can achieve excellent performance on the test set by taking the shortcut of fitting the training samples. As shown in the Table 4, the improvement in test set performance mainly comes from seen keyphrases, while the performance on unseen keyphrases is basically similar to zero-shot. Additionally, by observing the SFT model's case performance on the test set, we found that the model generates completely unrelated KP words for some posts, and the content of these posts is similar to those in the training set that contain in-

Figure 4: Examples of Multi-modal Keyphrase Prediction. Green denotes correct keyphrase predictions, whereas red denotes incorrect keyphrase predictions.

correctly predicted keyphrases. *This means that the SFT model is relying more on similar memory for keyphrases prediction, rather than reasoning ability.*

Balancing commonality and generalization. Based on the preceding analysis, we recognize that to effectively address MMKP task, a model must develop dual capabilities: leveraging commonality for prediction with seen keyphrases while employing generalization for unseen keyphrases. Our proposed method dynamically regulates chain-of-thought learning according to the difficulty of the samples, leading to a better trade-off between commonality learning and generalization ability. As shown in Fig. 4, during inference, the model adaptively selects either direct keyphrase prediction or prediction with CoT based on individual samples. This approach enhances model accuracy while simultaneously maintaining a balance with inference costs. As a result, our approach achieves optimal performance, with improvements on unseen samples reaching up to 20-30%.

Ablation study on the dynamic CoT loss threshold γ. The γ parameter serves as a critical factor in our experiments, as it defines the boundary between SFT and CoT data. We conducted a rigorous ablation study to investigate the impact of various thresholds, as detailed in Table 5. The initial threshold was selected based on the convergence behavior of the model's loss curve during the SFT stage, and an adaptive threshold was also evaluated. Experimental results indicate that all tested thresholds yield improvements over the baseline, with $\gamma = 0.4$ achieving the best performance. The adaptive threshold, however, did not reach optimal results in our experiments.

γ	MMKP-V2		
	All	Absent	Unseen
baseline	29.89	19.59	8.79
avg	32.22	19.50	**13.26**
0.3	33.07	**20.51**	12.14
0.4	**33.14**	20.19	12.48
0.5	32.65	19.91	12.26
0.6	32.59	19.65	12.40

Table 5: Ablation study on Dynamic CoT threshold γ. "baseline" denotes Qwen2.5-VL-3B. "avg" indicates that the average train loss is used as a dynamic threshold during training.

5 Conclusion

This study provides a comprehensive investigation for employing Vision-Language Models (VLMs) in multi-modal keyphrase prediction. To enhance model generalization, we developed multi-modal CoT data that advance VLMs' reasoning capabilities. Considering the "overthinking" phenomenon, we propose a Dynamic CoT training strategy that adaptively optimizes reasoning processes while preserving model generalizability and computational efficiency. Furthermore, by constructing new datasets with reduced train-test keyphrase overlap, we provide a more realistic evaluation of model generalization. Experimental results confirm that VLMs equipped with Dynamic CoT significantly outperform traditional multi-modal approaches, demonstrating superior cross-modal understanding and enhanced generalization.

Limitations

While our framework demonstrates promising results on Multi-modal Keyphrase Prediction (MMKP) task through Vision-Language Models (VLMs) and Dynamic Chain-of-Thought (Dynamic CoT), three primary limitations persist. First, the threshold determination in Dynamic CoT remains empirically driven. Despite testing adaptive threshold selection strategies, we observed suboptimal performance in dynamic adjustment. Second, VLMs inherently possess substantially larger parameter counts (e.g., 2B+) compared to traditional multi-modal models, resulting in elevated computational overhead during inference. In addition, incorporating reasoning capabilities further scales the test-time computation. Thrid, the cost of generating CoT data is relatively high, which poses challenges for the creation of larger-scale CoT datasets.

Ethics Statement

All the data utilized in our study was sourced from publicly available content on internet platforms. The seed instructions, which are openly accessible, comply with their respective open-source licenses. Furthermore, these datasets exclude any instances that could give rise to ethical concerns, such as unauthorized sensitive information, thereby minimizing potential societal risks.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Piyush Bansal, Somay Jain, and Vasudeva Varma. 2015. Towards semantic retrieval of hashtags in microblogs. In *Proceedings of the 24th International Conference on World Wide Web*, pages 7–8.

Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. 2013. Towards twitter context summarization with user influence models. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 527–536.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024a. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249.

Yifan Dong, Suhang Wu, Fandong Meng, Jie Zhou, Xiaoli Wang, Jianxin Lin, and Jinsong Su. 2023. Towards better multi-modal keyphrase generation via visual entity enhancement and multi-granularity image noise filtering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3897–3907.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Byungha Kang and Youhyun Shin. 2025. Empirical study of zero-shot keyphrase extraction with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3670–3686.

Fei-Fei Kou, Jun-Ping Du, Cong-Xian Yang, Yan-Song Shi, Wan-Qiu Cui, Mei-Yu Liang, and Yue Geng. 2018. Hashtag recommendation based on multi-features of microblogs. *Journal of Computer Science and Technology*, 33:711–726.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Meta. 2024. Llama3.2: Revolutionizing edge ai and vision with open, customizable models.

OpenAI. 2023. GPT-4V(ision) System Card.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, and 1 others. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

ByteDance Seed. 2025. Doubao-1.5-pro.

ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, and 1 others. 2025. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.

Liangying Shao, Liang Zhang, Minlong Peng, Guoqi Ma, Hao Yue, Mingming Sun, and Jinsong Su. 2024. One2set+ large language model: Best partners for keyphrase generation. *arXiv preprint arXiv:2410.03421*.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1031–1040.

Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R Lyu, and Shuming Shi. 2019a. Topic-aware neural keyphrase generation for social media language. *arXiv preprint arXiv:1906.03889*.

Yue Wang, Jing Li, Irwin King, Michael R Lyu, and Shuming Shi. 2019b. Microblog hashtag generation via encoding conversation contexts. *arXiv preprint arXiv:1905.07584*.

Yue Wang, Jing Li, Michael R Lyu, and Irwin King. 2020. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. *arXiv preprint arXiv:2011.01565*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Tao Gui, Yichao Luo, Yige Xu, and Qi Zhang. 2021. One2set: Generating diverse keyphrases as a set. *arXiv preprint arXiv:2105.11134*.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2018. One size does not fit all: Generating and evaluating variable number of keyphrases. *arXiv preprint arXiv:1810.05241*.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic memory networks for short text classification. *arXiv preprint arXiv:1809.03664*.

Chao Zhang, Shiwei Wu, Haoxin Zhang, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024a. Notellm: A retrievable large language model for note recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 170–179.

Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2024b. Notellm-2: Multimodal large representation models for recommendation. *arXiv preprint arXiv:2405.16789*.

Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. 2017. Hashtag recommendation for multimodal microblog using co-attention network. In *IJCAI*, pages 3420–3426.

Qi Zhang, Yang Wang, Yeyun Gong, and Xuan-Jing Huang. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 836–845.

Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1676–1686.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Models	MMKP	MMKP-V2			MMKP-360k			Avg
	All	All	Absent	Unseen	All	Absent	Unseen	All
Image-only models								
VGG	15.69	-	-	-	-	-	-	-
BUTD (Anderson et al., 2018)	20.02	-	-	-	-	-	-	-
Text-only models								
ONE2SEQ (Yuan et al., 2018)	38.05	-	-	-	-	-	-	-
ONE2SET (Ye et al., 2021)	36.36	-	-	-	-	-	-	-
TOPIC (Wang et al., 2019a)	43.17	-	-	-	-	-	-	-
Llama-3.2-2B	43.50	21.92	8.62	6.28	36.49	19.22	25.98	33.97
Qwen2.5-3B	48.53	24.33	12.35	6.58	37.21	19.99	26.04	36.69
Qwen2.5-7B	48.84	22.74	12.40	3.94	37.75	20.62	26.68	36.44
Image-text models								
CO-ATT (Zhang et al., 2017)	42.12	-	-	-	-	-	-	-
M^3H-ATT (Wang et al., 2020)	47.06	-	-	-	-	-	-	-
MM-MKP (Dong et al., 2023)	48.19	-	-	-	-	-	-	-
Llama-3.2-11B-Vision	59.81	28.30	20.92	5.45	-	-	-	-
InternVL-3-2B	57.54	28.89	18.10	7.98	38.53	18.47	24.42	41.65
+ Dynamic CoT (ours)	**59.63**	**30.76**	**18.99**	**9.90**	**40.03**	**19.95**	**26.04**	**43.47**
InternVL-3-8B	57.83	28.58	18.30	7.17	40.48	19.88	25.45	42.30
+Dynamic CoT (ours)	**60.29**	**31.42**	**19.13**	**10.68**	**50.53**	**20.04**	**26.44**	**47.41**
Qwen2.5-VL-3B	60.33	29.89	19.59	8.79	43.04	22.06	24.60	44.42
+ Dynamic CoT (ours)	**61.90**	**33.14**	**20.19**	**12.48**	**47.51**	**22.62**	**26.68**	**47.52**
Qwen2.5-VL-7B	60.83	30.49	20.90	7.90	43.70	22.28	24.98	45.01
+Dynamic CoT (ours)	**63.58**	**33.56**	**22.32**	**13.36**	**50.66**	**23.41**	**26.43**	**49.27**

Table 6: Performance comparison for multi-modal keyphrase prediction task.

A Appendix

Comprehensive Experimental Results. Table 6 presents detailed experimental results for additional model variants, including the image-only and text-only models.

multi-modal models demonstrate superior performance over pure visual or pure language models on MMKP task involving multi-modal social media posts. Extracting key information from such posts typically requires integrating both visual and textual cues.

Inference Length. Table 7 presents the inference sequence lengths across model variants, demonstrating that our dynamic Chain-of-Thought (CoT) strategy substantially reduces CoT reasoning steps while decreasing computational overhead by 38.48%.

The Overfitting Phenomenon in SFT Models. Figure 5 visualizes prediction of the SFT model on the test set, revealing its over-reliance on similarity-based memorization. The model predominantly predicts identical labels for visually analogous samples, demonstrating limited deployment of reasoning capabilities.

Models	Output Length
Ground Truth	1.33
+SFT	1.89
+Fine-tune-CoT	157.37
+Dynamic CoT	96.82

Table 7: Inference sequence lengths (measured in words) of different model variants on the MMKP-V2 dataset, all implemented using the Qwen2.5-VL-3B.

Detailed CoT. Figure 6 illustrates the Chain-of-Thought reasoning processes in our model, demonstrating its dual analytical capabilities: multimodal content analysis (textual and visual elements of social media posts) combined with social impact

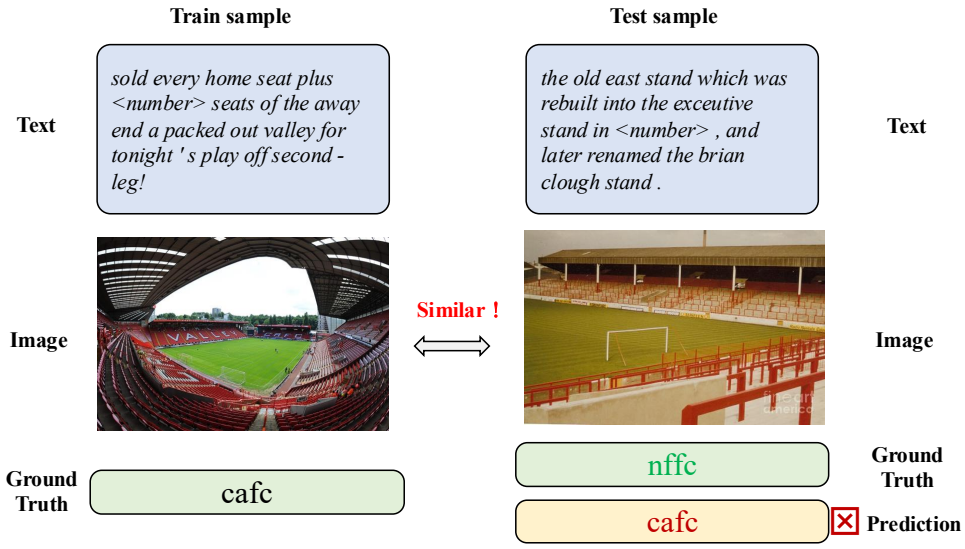
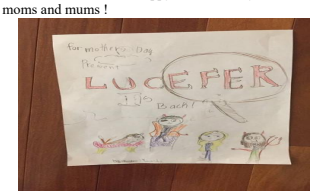


Figure 5: Visualization of SFT models on test set of MMKP dataset.

Post (a) maybe my daughter does not know how to spell , but she knows me well happy mother ' s day , dear lucifam moms and mums !



GT: lucifer

Previous SOTA: mothers day

VLM sft: mothers day

VLM CoT: <think>The hashtag "lucifam" is used by social media users to connect with the fan community of the TV show "Lucifer." Fans of the show often use this hashtag to share content, express their fandom, and engage with each other. In the context of the post, the user is likely sharing a heartfelt message to their daughter, possibly referencing the show "Lucifer" and its characters. By using the hashtag "lucifam," the user is signaling to the fan community that this post is related to the show and its fandom. This helps in garnering attention from other fans, fostering a sense of community, and celebrating Mother's Day with a special touch, acknowledging the daughter's connection to the show.</think>
<answer>go owls</answer>

Dynamic CoT:<think>The hashtag "lucifam" is used by social media users to connect with the community of fans of the TV show "Lucifer." In the post, the user is addressing "Dear lucifam moms and mums" on Mother's Day, indicating that the content is specifically relevant to the female fans of the show. The use of "lucifam" helps the post reach a targeted audience who share a common interest in the show and its characters, fostering a sense of community and shared celebration. The accompanying image, which features the show's title and characters, further reinforces the connection to the "Lucifer" fandom.</think>
<answer>go owls</answer>

Post (b) " when daisies pied and violets blue and lady - smocks all silver - white and cuckoo - buds of yellow hue do paint the meadows with delight . " love ' s labours lost .

GT: shakespeare sunday

Previous SOTA: folklore thursday

VLM sft: sc rivo arte

VLM CoT: <think>Social media users might use the hashtags "sc rivo arte, poetry, writing community" for a post titled "when daisies pied and violets blue and lady-smocks all silver-white and cuckoo-buds of yellow hue do paint the meadows with delight. love's labours lost." accompanied by images like the one provided to engage with specific communities interested in art, poetry, and literary expressions. The image, which is a painting of "Love's Labours Lost" by John William Waterhouse, complements the poetic description by depicting a scene from the poem. By using these hashtags, users can reach and interact with art enthusiasts, poets, and writers who appreciate the poetic depiction of nature and the themes it encompasses.</think>
<answer>sc rivo arte; poetry</answer>

Dynamic CoT:<think>Social media users might use the hashtag "Shakespeare Sunday" for a post featuring the title "When daisies pied and violets blue and lady-smocks all silver-white and cuckoo-buds of yellow hue do paint the meadows with delight. Love's labours lost." along with images to participate in a themed event that celebrates the works of William Shakespeare. "Shakespeare Sunday" is a popular hashtag used to share content related to Shakespeare's poetry, plays, and life on Sundays. The post's title and imagery evoke a poetic and nostalgic sentiment, aligning well with the themes of love and nature that Shakespeare often explored in his works. By using this hashtag, users can connect with others who appreciate Shakespeare's literary contributions and engage in a community of like-minded individuals.
</think>
<answer>shakespeare sunday</answer>

Post (c) they be eating fluffy pink n white squidgy noms . i will give them my bestest puppy dog begging eyes until they submit and give me one .

GT: happy friday

Previous SOTA: begging eyes

VLM sft: puppy

VLM CoT: <think>Social media users might use the hashtag "sorrynotsorry" for a post like this to convey a playful yet unapologetic attitude. The phrase "sorrynotsorry" typically indicates a sense of humor and a lack of genuine regret for one's actions. In the context of the post, the user is engaging in a humorous, puppy-like mannerly behavior, offering their best puppy dog eyes to get what they want, and using the hashtag to add a light-hearted, nonchalant tone to the request. This can make the post more relatable and entertaining, encouraging others to share in the playful spirit and perhaps even participate in the conversation or interaction.</think>
<answer>sorry not sorry</answer>

Dynamic CoT:<think>Social media users might use the hashtag "puppy" for a post like this for several reasons: 1. Engagement and Visibility: Hashtags like "puppy" are popular and widely used, which can increase the visibility of the post. People who search for or follow the "puppy" hashtag are more likely to see the post, thereby increasing engagement through likes, comments, and shares. 2. Relevance to Content: The post features a dog, which is directly related to the hashtag "puppy". Using relevant hashtags helps categorize the content correctly, making it easier for users interested in puppy-related posts to find it. 3. Target Audience: The hashtag "puppy" targets a specific audience that is interested in dogs. This can attract followers who are specifically interested in dog content, leading to a more engaged and interested audience. 4. Emotional Appeal: The caption of the post is written in a way that appeals to the emotions of dog lovers. The term "bestest puppy dog begging eyes" is likely to resonate with people who have a soft spot for dogs, and the hashtag "puppy" helps to connect with this audience. 5. Trend Participation: Hashtags like "puppy" are often trending, and participating in trending hashtags can help the post gain traction. It shows that the user is up-to-date with popular social media trends. Overall, using the hashtag "puppy" helps to increase the post's reach, relevance, and engagement by connecting it to a specific and interested audience.</think>
<answer>puppy</answer>

Figure 6: Detailed CoT by our proposed method.

prediction (assessing keyphrase potential for community consensus formation and engagement generation).