Top-Down Semantic Refinement for Image Captioning

Jusheng Zhang[1*], Kaitong Cai[1*], Jing Yang[1], Jian Wang[2], Chengpei Tang[1†], Keze Wang[1,3†]

[1]Sun Yat-sen University
[2]Snap Inc.
[3]Guangdong Key Laboratory of Big Data Analysis and Processing

Abstract

Large Vision-Language Models (VLMs) face an inherent contradiction in image captioning: their powerful single-step generation capabilities often lead to a **myopic** decision-making process. This makes it difficult to maintain global narrative coherence while capturing rich details, a limitation that is particularly pronounced in tasks that require multi-step and complex scene description. To overcome this fundamental challenge, we redefine image captioning as a **goal-oriented hierarchical refinement planning problem**, and further propose a novel framework, named Top-Down Semantic Refinement (TDSR), which models the generation process as a Markov Decision Process (MDP). However, planning within the vast state space of a VLM presents a significant computational hurdle. Our core contribution, therefore, is the design of a **highly efficient Monte Carlo Tree Search (MCTS) algorithm tailored for VLMs**. By incorporating a **visual-guided parallel expansion** and a **lightweight value network**, our TDSR reduces the call frequency to the expensive VLM by an order of magnitude without sacrificing planning quality. Furthermore, an adaptive early stopping mechanism dynamically matches computational overhead to the image's complexity. Extensive experiments on multiple benchmarks, including DetailCaps, COMPOSITIONCAP, and POPE, demonstrate that our TDSR, as a plug-and-play module, can significantly enhance the performance of existing VLMs (e.g., LLaVA-1.5, Qwen2.5-VL) by achieving state-of-the-art or highly competitive results in fine-grained description, compositional generalization, and hallucination suppression.

Introduction

At the intersection of computer vision and natural language processing, Large Vision-Language Models (VLMs)(Radford et al. 2021; Tan and Bansal 2019; Chen et al. 2023a; Yao et al. 2024) have become the dominant force in image captioning. Through powerful visual encoders and language decoders, these models can generate fluent text that is generally aligned with the image content(Li et al. 2022; Radford et al. 2021; Zhang et al. 2025a,e,d). However, despite their remarkable success, the core auto-regressive generation mechanism of VLMs exposes a fundamental flaw, i.e., an inherent lack of planning capability. When generating each token, VLMs typically employ greedy or beam search strategies(Radford et al. 2019; Rennie et al. 2017; Zhang et al. 2025d, 2026). This decision-making process is inherently "myopic", confined to maximizing local probabilities without "deliberate thought" or foresight and planning capability for the global narrative structure.

This lack of planning capability leads to an intractable dilemma: the model either produces a coherent but detail-poor "safe" description to ensure consistency, or it generates factual errors and logical breaks, i.e., the "hallucination" phenomenon, when attempting to capture rich details without global guidance (Jia et al. 2021; Radford et al. 2021; OpenAI 2023; Zhang et al. 2025c,f). To address this challenge, the research community once turned to a seemingly intuitive "bottom-up" paradigm(Zhang et al. 2021; Herdade et al. 2020). These methods first detect independent regions in an image, describe them separately, and finally "stitch" these fragmented descriptions into a complete caption. However, this "local-to-global" strategy fails to address the core problem. Lacking a unified global plan as an anchor from the outset, the resulting descriptions often degenerate into a simple list of facts, leading to semantic fragmentation and logical incoherence(Alikhani et al. 2020; Bugliarello and Elliott 2021; Zohourianshahzadi and Kalita 2021; Zhang et al. 2025g). This proves that merely stitching details together cannot effectively compensate for the VLM's lack of planning ability.

We argue that the root of the problem lies in the generation paradigm itself, and the solution is to fundamentally reframe image captioning as a planning problem. To this end, we propose an innovative "top-down" semantic refinement framework (TDSR), which redefines the task from a unidirectional generation process into a coarse-to-fine, goal-oriented, hierarchical planning process(Yarats and Lewis 2018; Zhang et al. 2025h). This idea, illustrated in Figure 1, mimics the human cognitive process(Mefford et al. 2023): first, form a holistic impression of the image to generate a high-level, core description as a "planning blueprint" (e.g., for a picture of people playing cards, an initial description might be "a group of people are sitting in a room doing something").

*These authors contributed equally.
†Corresponding Author: kezewang@gmail.com and tchengp@mail.sysu.edu.cn

Then, using this blueprint as a guide, purposefully and progressively explore and fill in key details (e.g., further specifying it as “a group of men are sitting around a table, engaged in a game of Texas Hold’em poker," and adding that “on the green felt tabletop lie three community cards and a collection of poker chips"). This “global guidance, local refinement" mechanism ensures that all details serve a unified narrative goal, fundamentally guaranteeing high coherence and richness in the description.

Translating this elegant planning concept into an effective computational process hinges on efficient search and planning within the vast language space(Wiher, Meister, and Cotterell 2022). We rigorously formalize this process as a Markov Decision Process (MDP(Puterman 1994)) and employ Monte Carlo Tree Search (MCTS(Kemmerling, Lütticke, and Schmitt 2023)) as the core engine. However, directly applying standard MCTS to a VLM is computationally infeasible due to the model’s massive inference cost(Browne et al. 2012a). Therefore, our core technical contribution lies in deeply optimizing the MCTS algorithm to enable efficient planning within VLMs. By incorporating a Visual-Guided Parallel Expansion mechanism and a lightweight value network, our algorithm reduces the call frequency to the expensive VLM by an order of magnitude without sacrificing planning quality, successfully resolving the efficiency bottleneck. Our framework (TDSR), as a plug-and-play module, significantly enhances the performance of existing VLMs and achieves state-of-the-art or highly competitive results on multiple benchmarks.

Our core contributions can be summarized in three points:

- **A Novel “Planning-based" Generation Paradigm:** We propose a “Top-Down" planning framework (TDSR) that redefines image captioning as a coarse-to-fine hierarchical planning problem. This fundamentally resolves the “myopic" flaw of traditional generative models, ensuring both global narrative coherence and local detail richness.
- **An Efficient MCTS Algorithm Tailored for VLMs:** We design a highly efficient Monte Carlo Tree Search (MCTS) algorithm to address the high inference cost of VLMs. The algorithm broadens search breadth via a “Visual-Guided Parallel Expansion" mechanism and uses a “lightweight value network" for fast value estimation, reducing the call frequency to the expensive VLM by an order of magnitude without sacrificing planning quality.
- **Dynamic, Adaptive Search Control:** We propose a control strategy that improves planning efficiency and quality by steering search with a composite reward (redundancy penalty + depth incentive) and reducing overhead via image-complexity–aware adaptive early stopping.

Related Work

Early Encoder-Decoder Architectures

As a cross-disciplinary field between computer vision and natural language processing(Zhang et al. 2025b; Radford et al. 2021), early research in image captioning primarily adopted the encoder-decoder framework. Methods represented by Show and Tell (Vinyals et al. 2015) used a

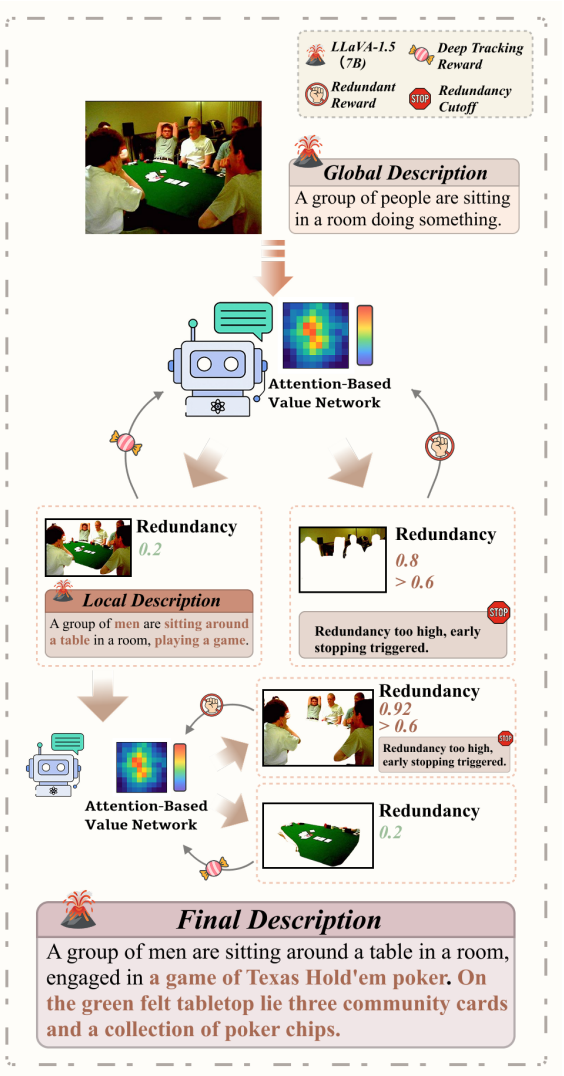

Figure 1: The TDSR framework generates coherent, detailed captions through global-to-local refinement, guided by redundancy-aware stopping and efficient MCTS.

CNN(Krizhevsky, Sutskever, and Hinton 2012) to extract global image features and an RNN(Elman 1990) to generate a fluent description. Subsequently, Show, Attend and Tell (Vinyals et al. 2015) introduced an attention mechanism, allowing the model to focus on local regions of the image. These pioneering works excelled at generating grammatically coherent sentences, but their mechanism of generating a single, holistic description meant they often overlooked fine-grained object details, leading to a lack of richness and specificity. **This is precisely one of the core problems our TDSR framework aims to solve through multi-step refinement.**

The “Bottom-up" Generation Paradigm

To enhance detail capture, subsequent research shifted towards the “bottom-up" paradigm(Zhang et al. 2021). This

approach typically first identifies independent objects or regions in an image, generates local descriptions for them, and finally stitches these segments into a complete sentence. DenseCap (Johnson, Karpathy, and Fei-Fei 2015) is a typical representative of this line of work, which utilizes an object detector to locate and describe regions individually. Follow-up works, such as Patch Matters (Peng et al. 2025b) and FineCaption (Hua et al. 2024b), focused on improving the quality of these local descriptions. Although these methods significantly increased detail richness, their "split-first, stitch-later" process inherently decouples from the global context, often leading to semantic fragmentation and insufficient global coherence. In stark contrast, our "top-down" approach, guided by global context, fundamentally avoids the inconsistency issues inherent in the "split-first, stitch-later" process.

Large Vision-Language Models and Generation Refinement Strategies

In recent years, Large Vision-Language Models (VLMs), such as LLaVA-1.5 (Liu et al. 2023b), Qwen-VL (Bai, Bai, and et al. 2023), and Ferret (You and et al. 2023), have significantly advanced visual narrative capabilities through pre-training on massive image-text data. However, despite their powerful foundational abilities, their standard autoregressive generation still faces the inherent trade-off between detail and coherence. Consequently, several training-free enhancement methods have emerged. For instance, some works employ iterative prompting (e.g., IT (Zhou et al. 2023)) to induce the model to output more details.

A more promising direction involves formalizing the generation process as a search problem and employing planning algorithms like Monte Carlo Tree Search (MCTS) for optimization (Browne et al. 2012b). Against this backdrop, our TDSR framework proposes a more fundamental solution. It also employs MCTS, but its core innovation lies in how to tailor and efficiently execute the search for VLMs. Our method is guided by a sophisticated composite reward function and integrates a suite of efficiency optimization mechanisms, including visual-guided parallel rollouts and dynamic redundancy control. **Consequently, TDSR is not only superior in its paradigm ('top-down') but also innovative in its solution strategy (efficient MCTS), thereby efficiently unifying detail and coherence while significantly mitigating issues like semantic fragmentation and content hallucination.**

Methodology

This section details our Top-Down Semantic Refinement (TDSR) framework. We begin by formalizing the task of progressive image description as a Markov Decision Process (MDP), explicitly framing it as a planning problem. We then present our core contribution: a highly efficient Monte Carlo Tree Search (MCTS)(Kocsis and Szepesvári 2006; Browne et al. 2012b; Yao et al. 2023) algorithm designed to solve this MDP. Our MCTS variant introduces several key optimizations, including a **Visual-Guided Parallel Expansion** strategy, a **lightweight value network** for fast simulations, and dynamic reward shaping, which collectively enable high-quality planning without incurring prohibitive computational costs.

Image Captioning as a Planning Problem

We cast the challenge of generating a detailed and coherent caption $Y = (y_1, y_2, \ldots, y_L)$ for an image I as a sequential decision-making problem. The goal is to find an optimal policy π^* that generates a sequence of tokens maximizing a cumulative reward. This process is formally defined as a Markov Decision Process (MDP), specified by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$:

State $\mathcal{S}$: A state $s_t \in \mathcal{S}$ is the prefix of a caption being generated, represented by the sequence of tokens $(y_1, \ldots, y_t)$. The initial state s_0 can be an empty sequence or a high-level caption generated by a base VLM.

Action $\mathcal{A}$: An action $a_t \in \mathcal{A}$ corresponds to selecting the next token y_{t+1} to append to the current sequence. The set of possible actions is the model's vocabulary.

Transition $\mathcal{P}$: The transition function $\mathcal{P}(s_{t+1}|s_t, a_t)$ is deterministic: taking action a_t in state s_t leads to state $s_{t+1} = s_t \oplus a_t$, where $\oplus$ denotes concatenation.

Reward $\mathcal{R}$: Upon reaching a terminal state s_T (e.g., by generating an end-of-sequence token), the environment returns a terminal reward $R(s_T)$. This reward function is meticulously designed to encourage detailed, coherent, and non-repetitive descriptions:

$$R(s_T) = R_{\text{quality}}(s_T, I) + R_{\text{depth}}(s_T) - P_{\text{redundancy}}(s_T) \quad (1)$$

where R_{quality} assesses fine-grained relevance and compositional correctness (e.g., using CLIP-based scores). The term $R_{\text{depth}} = \alpha \cdot \log(1 + |s_T|)$ provides a **depth incentive** to encourage longer, more detailed descriptions. Finally, $P_{\text{redundancy}}$ penalizes semantic repetition using efficient metrics like n-gram overlap.

The objective is to find a policy $\pi(a_t|s_t)$ that maximizes the expected reward. Given the massive state-action space, we employ MCTS as a powerful online planning algorithm to approximate the optimal policy.

MCTS for Coarse-to-Fine Planning

MCTS is an ideal choice for this problem as it builds a search tree asynchronously(Yao et al. 2023), focusing its computational efforts on more promising regions of the state space. Our key innovation lies in how we integrate the VLM and other components into the four canonical steps of MCTS. The entire TDSR process is outlined in Algorithm 1.

At any node s in the search tree, we store the total action value $W(s, a)$, the visit count $N(s, a)$, and the prior probability $P(s, a)$ for each action a.

1. Selection. Starting from the root node, we recursively select the action that maximizes the Upper Confidence Bound for Trees (UCT) criterion until a leaf node s_L is reached. The summation in the UCT formula is over all valid

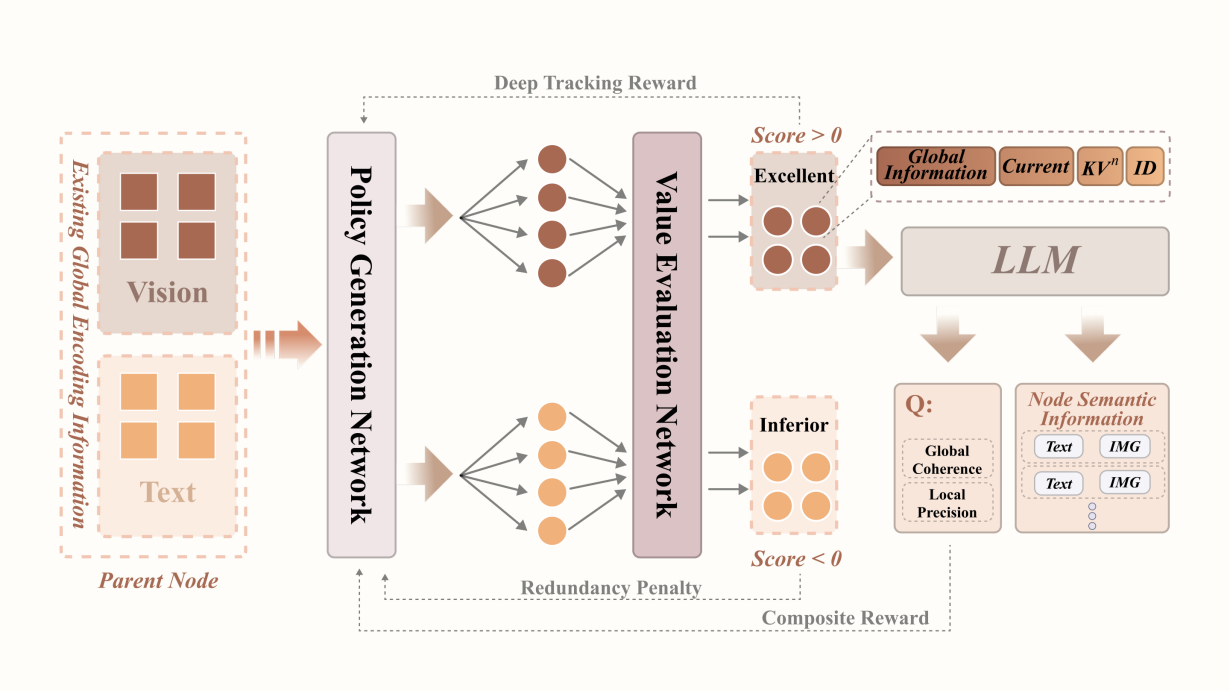

Figure 2: The architecture of TDSR's MCTS planner. The four canonical stages, i.e., selection, visual-guided parallel expansion, lightweight value estimation, and backpropagation, are tailored to efficiently search within VLMs. Composite rewards combine local precision and global coherence.

actions b from state s_t:

$$a_t = \arg\max_a \left(Q(s_t, a) + c_{\text{puct}} \cdot P(s_t, a) \cdot \frac{\sqrt{\sum_b N(s_t, b)}}{1 + N(s_t, a)} \right) \tag{2}$$

Here, $Q(s, a) = W(s, a)/N(s, a)$ is the mean action value (exploitation term), and $P(s, a)$ is the **policy prior** derived from our base VLM to guide the search.

Visual-Guided Parallel Expansion. Upon reaching a leaf node s_L, instead of expanding only one path, we guide the VLM to explore multiple, visually-grounded semantic paths in parallel. This unfolds in two stages:

1. **Salient Region Identification:** We leverage cross-attention maps from the VLM $\mathcal{G}_\theta$ or an external object detector to identify k salient image regions not yet adequately described in the current caption s_L.
2. **Parallel Prompting and Expansion:** For each region, we construct a unique exploratory prompt (e.g., "Describe the person's clothing in more detail."). We then execute the VLM $\mathcal{G}_\theta$ in parallel for these k inputs. This single batch-forward pass yields k policy vectors and k VLM-based value estimates:

$$(p_a^{(i)}, v_{\text{vlm}}^{(i)}) = \mathcal{G}_\theta(\text{prompt}_i, s_L, I) \quad \text{for } i = 1, \ldots, k \tag{3}$$

The node s_L is then expanded with new children corresponding to promising actions from the policy vectors $p_a^{(i)}$. This ensures search breadth is explicitly grounded in diverse visual evidence.

Simulation (and Lightweight Value Estimation). This step is critical for efficiency. Instead of performing a costly "rollout" with the VLM, we estimate the value of the new leaf node s_L using a separate, **lightweight value network** $\mathcal{V}_\phi$. This network is trained to approximate the final reward $R(s_T)$ from an intermediate state:

$$\hat{v} = \mathcal{V}_\phi(s_L, I) \tag{4}$$

This AlphaGo-inspired approach replaces expensive simulations with a single, fast forward pass. The final value estimate V is a weighted combination of the VLM's coarse estimate from the expansion step (v_{vlm}) and the specialized value network's estimate ($\hat{v}$):

$$V = \lambda_v \cdot v_{\text{vlm}} + (1 - \lambda_v) \cdot \hat{v} \tag{5}$$

Value Network Architecture and Training. The lightweight value network $\mathcal{V}_\phi$ is designed for speed. Its architecture consists of a 4-layer Transformer encoder that processes the token sequence s_L, whose output is then concatenated with the global image features from the VLM's vision encoder. This combined representation is passed through a 2-layer MLP head to regress a single scalar value $\hat{v}$. To train $\mathcal{V}_\phi$, we generate a dataset of state-reward pairs by running the full TDSR search on a large corpus of images. For each completed search, we store all intermediate states s_t encountered and the final, true reward $R(s_T)$ of the resulting caption. The network is then trained offline using a Mean Squared Error (MSE) loss between its prediction $\hat{v}$ for a state s_t and the corresponding ground-truth terminal

Algorithm 1: Top-Down Semantic Refinement (TDSR)

function TDSR_GENERATE(I, L)
  $s_{\text{caption}} \leftarrow$ initial prompt or empty sequence
  for $t = 1$ to L **do**
    $s_{\text{root}} \leftarrow s_{\text{caption}}$
    Initialize MCTS tree T with root node s_{root}
    for $i = 1$ to $N_{\text{max\_iterations}}$ **do**
      $s_{\text{leaf}} \leftarrow \text{SelectLeafNode}(s_{\text{root}}, T)$
      $(P, v_{\text{vlm}}) \leftarrow \text{Expand}(s_{\text{leaf}}, I)$ ▷ Via visual-guided parallel expansion
      $\hat{v} \leftarrow \mathcal{V}_\phi(s_{\text{leaf}}, I)$ ▷ Estimate value with lightweight network
      $V \leftarrow \lambda_v \cdot v_{\text{vlm}} + (1 - \lambda_v) \cdot \hat{v}$ ▷ Combine value estimates
      Backpropagate value V from s_{leaf} to s_{root}
      if search has converged at root **then** ▷ Adaptive termination
        break
      end if
    end for
    $y_{t+1} \leftarrow \arg\max_a N(s_{\text{root}}, a)$ ▷ Select best action
    if y_{t+1} is end-of-sequence token **then**
      break
    end if
    $s_{\text{caption}} \leftarrow s_{\text{caption}} \oplus y_{t+1}$
  end for
  return s_{caption}
end function

reward $R(s_T)$. **Backpropagation.** The combined value estimate V is propagated back up the tree to update the visit counts $N(s, a)$ and total action values $W(s, a)$ for all edges on the traversed path from s_L to the root.

Experiment

Experimental Settings

Evaluation Tasks In this study, we evaluate the performance of our method on three distinct and comprehensive benchmark datasets: **DetailCaps**(Dong et al. 2024), **COMPOSITIONCAP**(Hua et al. 2024a), and **POPE**(Li et al. 2023), each of which aims to assess different aspects of image description and reasoning tasks. **DetailCaps**: This benchmark provides a high-quality image captioning dataset to evaluate LVLMs on their ability to generate detailed descriptions at the object, attribute, and relationship levels. The **CAPTURE metric** measures detail coverage across these dimensions, offering a systematic framework for assessing multimodal models' fine-grained image understanding. **COMPOSITIONCAP**: This benchmark evaluates the compositional generalization ability of multimodal models, focusing on their capacity to describe images with novel combinations of objects, attributes, and relationships. It tests models' compositional reasoning by requiring accurate descriptions of unseen combinations. **POPE**: This benchmark is designed to assess the phenomenon of object hallucination in multimodal large models. It focuses on detecting whether models falsely "fabricate" objects or details that do not exist in the image during image description or question-answering tasks.

Method	CAPTURE	F1_obj	F1_attr	F1_rel
Shikra-13B	60.5	61.9	55.4	56.4
MiniGPT-v2-7B	61.2	62.7	55.9	56.8
Ferret-13B	62.8	63.2	56.4	57.3
VisionLLM-H-7B	57.9	59.3	53.4	53.9
KOSMOS-2	58.5	60.8	53.1	54.2
Alpha-CLIP-13B	59.2	61.9	57.2	56.5
FINECAPTION-8B	63.4	63.7	58.1	58.3
VistaLLM-13B	63.2	63.5	60.3	59.2
LLaVA-1.5-7B+IT	51.98	56.3	48.2	50.4
LLaVA-1.5-7B+Patch Matters	58.05	62.2	56.1	52.5
LLaVA-1.5-7B	49.99	55.7	44.4	49.4
LLaVA-1.5-7B+ours	66.7	66.2	62.4	63.4
Qwen2.5-VL-7B	64.7	66.7	62.5	62.3
Qwen2.5-VL-7B+ours	**72.2**	**72.3**	**65.2**	**64.7**

Table 1: Performance comparison on the DetailCaps benchmark. TDSR-enhanced models achieve consistent improvements across all fine-grained metrics.

Baselines To systematically evaluate the generalization and practical effectiveness of **TDSR** across different model architectures, we deploy it on two widely adopted multimodal large language models: **Qwen2.5-VL**(Bai et al. 2025) and **LLaVA-1.5 (7B)**(Liu et al. 2023a). We compare it against a diverse set of representative baselines, which fall into two major paradigms: **Training-free image captioning enhancement methods**: including *IT*(Pi et al. 2024), *Patch Matters*(Peng et al. 2025a), and *FINECAPTION*, which improve visual description quality without additional training. All baselines in this category are implemented on top of LLaVA-1.5 (7B). **Foundation vision-language models**: including *Shikra-13B*(Chen et al. 2023c), *MiniGPT-v2*(Chen et al. 2023b), *Ferret-13B*(You et al. 2023), *VisionLLM-H*(Wang et al. 2023), *KOSMOS-2*(Peng et al. 2023), *Alpha-CLIP-13B*(Sun et al. 2023), and *VistaLLM*(Pramanick et al. 2023), representing the dominant modeling paradigms in the open-source multimodal field.

Implementation Details Our TDSR framework is implemented in PyTorch. For the core MCTS algorithm, we set the UCT exploration constant c_{puct} to 1.5. The reward depth incentive weight α is set to 0.1. During inference, we applied TDSR to refine the outputs of both Qwen2.5-VL and LLaVA-1.5 (7B). **A comprehensive list of all hyperparameters, including the value network architecture and training specifics, is provided in Appendix A for reproducibility.**

Benchmark Model Comparison Experiment

DetailCaps Result The experimental results presented in Table 1 on the DETAILCAPS benchmark demonstrate that **TDSR** significantly enhances fine-grained semantic understanding in multimodal models, particularly in object, attribute, and relation-level comprehension. Under the LLAVA architecture, LLAVA-1.5+OURS exhibits notable improvements across all three fine-grained metrics ($F1_{\text{obj}}$, $F1_{\text{attr}}$, $F1_{\text{rel}}$) compared to the base model. In particular, $F1_{\text{attr}}$ rises markedly from 44.4 to 62.4, validating the effectiveness of TDSR in capturing detailed semantic signals

Method	ROUGE-L↑	BLEU-4↑	METEOR↑	CIDEr↑	BERT Score↑
Shikra-13B	32.4	11.9	19.5	108.4	78.4
MiniGPT-v2-7B	31.9	11.5	18.7	106.2	78.2
Ferret-13B	33.6	12.8	19.6	114.6	79.1
VisionLLM-H-7B	31.2	10.7	15.4	90.2	76.5
KOSMOS-2	30.8	10.1	14.9	88.9	76.7
Alpha-CLIP-13B	35.6	10.9	19.3	93.9	77.7
FINECAPTION-8B	40.6	13.9	20.9	118.6	79.5
VistaLLM-13B	40.9	14.1	21.4	117.5	80.2
LLaVA-1.5-7B +IT	32.9	10.6	15.7	95.2	78.2
LLaVA-1.5-7B +Patch Matters	34.6	12.5	21.2	118.8	79.1
LLaVA-1.5-7B	30.3	8.6	11.4	86.5	73.2
LLaVA-1.5-7B +ours	44.3	16.6	23.5	124.2	82.5
Qwen2.5-VL-7B	41.2	14.5	21.9	120.3	81.3
Qwen2.5-VL-7B +ours	**47.5**	**19.7**	**27.5**	**129.4**	**88.9**

Table 2: Benchmark comparison on the COMPOSITIONCAP dataset. Our method significantly outperforms all baselines across all metrics.

in image descriptions. Within the stronger QWEN2.5-VL architecture, TDSR further advances overall performance, achieving a CAPTURE score of 72.2, with $F1_{obj}$ and $F1_{rel}$ reaching 72.3 and 64.7 respectively, both significantly outperforming all other baselines. These results highlight the robust semantic modeling and visual-linguistic alignment capacity brought by TDSR. The proposed semantics-driven exploration mechanism exhibits consistent and effective improvements across both architectures, markedly enhancing the model's ability to capture key semantic units from images.

COMPOSITIONCAP Result The experimental results presented in Table 2 on the COMPOSITIONCAP benchmark demonstrate the effectiveness of the proposed **TDSR** method across different vision-language model architectures. Within the **LLaVA** framework, the incorporation of TDSR (*LLaVA-1.5+ours*) consistently improves performance over the base model, with ROUGE-L increasing to **44.3** and CIDEr reaching **124.2**, indicating enhanced descriptive completeness and detail sensitivity. In the stronger **Qwen2.5-VL** framework, TDSR yields further performance gains, achieving a CIDEr of **129.4** and a BERTScore of **88.9** (the best results to date), highlighting its superior modeling of visual-semantic consistency and linguistic precision.

Overall, compared to traditional non-trained augmentation methods (e.g., *IT*, *Patch Matters*, *FINECAPTION*) and mainstream multimodal models (e.g., *Ferret*, *VistaLLM*, *KOSMOS-2*), **TDSR** consistently demonstrates strong capabilities in cross-modal reasoning, compositional understanding, and expressive generation under both **Qwen** and **LLaVA** backbones.

Hallucination Evaluation

As shown in Figure 3, we conduct a systematic evaluation of several state-of-the-art multimodal models on the POPE benchmark, which is designed to assess hallucination robustness under three types of semantic perturbations:

Method	Accuracy	Precision	Recall	F1 Score
MSCOCO Dataset				
Ferret-13B	87.6	95.8	76.4	89.8
Shikra-13B	85.3	95.1	75.3	86.4
VistaLLM	88.2	96.7	78.6	90.5
LLaVA-1.5	87.6	96.9	78.2	89.2
LLaVA-1.5 + ours	**90.9**	**98.5**	**80.8**	**91.3**
Popular Dataset				
Ferret-13B	86.4	94.1	76.5	84.2
Shikra-13B	84.7	93.2	75.3	83.8
VistaLLM	87.3	95.2	77.9	86.5
LLaVA-1.5	86.7	95.7	78.4	86.9
LLaVA-1.5 + ours	**89.9**	**97.2**	**81.5**	**89.7**
Rare Dataset				
Ferret-13B	82.3	91.7	70.8	80.2
Shikra-13B	80.1	90.4	68.3	78.5
VistaLLM	83.4	92.8	72.5	82.1
LLaVA-1.5	82.7	93.1	71.9	81.2
LLaVA-1.5 + ours	**85.8**	**94.5**	**74.2**	**83.9**

Table 3: Benchmark comparison on the POPE dataset across MSCOCO, Popular, and Rare settings[cite: 213, 215, 216]. Our method (LLaVA-1.5 + ours) consistently achieves the best performance and demonstrates superior robustness against hallucinations.

Random, **Popular**, and **Adversarial**. POPE explicitly tests whether a model hallucinates non-existent entities or attributes in response to misleading prompts. Results indicate that **LLaVA-1.5+TDSR** consistently achieves the best performance across all settings, demonstrating superior robustness. Notably, under the most challenging **Adversarial** condition, it maintains an Accuracy of **86.3** and an F1 Score of **84.3**, significantly outperforming other models. In the relatively simpler **Random** setting, it achieves the highest F1 Score of **91.3**, slightly ahead of **VistaLLM (90.5)**. In the more ambiguous **Popular** setting, where semantically frequent entities such as "person" or "cat" may induce biased responses, most models experience a notable performance drop, i.e., **Ferret-13B** and **Shikra-13B** fall to 84.2 and 83.8 respectively, while **LLaVA-1.5+TDSR** remains stable at **87.1**, highlighting its robustness and generalization to biased prompts. We attribute TDSR's resistance to hallucination to its **top-down semantic reasoning**: a strong global context steers attention to the truly relevant regions when parsing fine-grained objects, thereby minimizing misalignment and fabricated details.

Efficiency Analysis

To comprehensively assess the efficiency-performance tradeoff of the TDSR framework, we conduct a series of controlled experiments comparing the full TDSR architecture, its variants with individual efficiency components disabled, and several representative vision-language baselines.

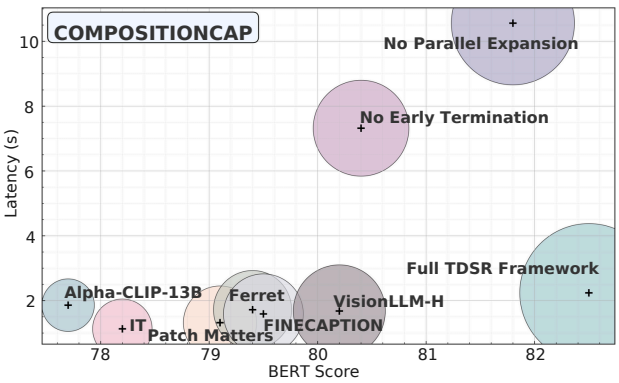

Figure 3: Efficiency-performance tradeoff of TDSR. The full framework achieves the best generation quality (BERTScore) with only a marginal latency increase, clearly outperforming prior methods.

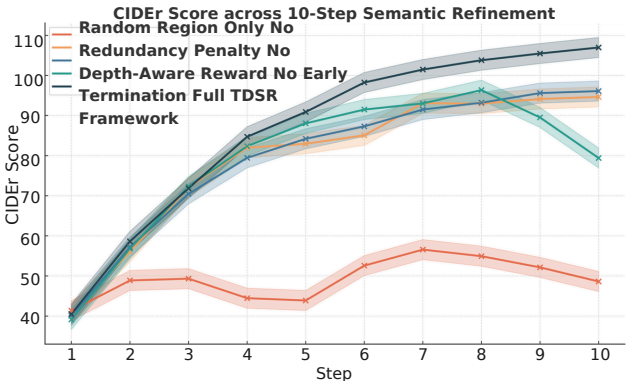

Figure 4: Step-wise CIDEr score under ablation settings. Removing any core component from TDSR results in significant performance drops, especially in early stopping and value-guided region selection.

The evaluation focuses on two primary metrics: inference latency and generation quality (measured by **BERTScore**).

As shown in Figure 3, the full TDSR framework achieves strong generation performance while maintaining a reasonable inference delay. Specifically, although its average latency slightly increases to **2.24s/frame**, this overhead remains marginal compared to mainstream baselines such as *VisionLLM-H* (1.68s), *FINECAPTION* (1.59s), and *Ferret* (1.72s). In contrast, TDSR yields a substantial improvement in output quality, achieving the highest **BERTScore of 82.5** on the *COMPOSITIONCAP* benchmark, significantly outperforming the aforementioned models (e.g., Ferret: 79.4, FINECAPTION: 79.5, VisionLLM-H: 80.2).

Ablation studies

To assess the contribution of each TDSR component, we randomly sample **100 COCO images** and track step-wise **CIDEr** and **BLEU-4** scores across 10 exploration steps. The ablation variants are: **w/o value estimation**: Disable value network; select regions randomly without semantic lookahead; **w/o redundancy penalty**: Remove penalties on repeated or overlapping descriptions; **w/o depth-aware reward**: Drop the reward term for fine-grained region tracking; **w/o early termination**: Always run 10 steps regardless of confidence; **Full TDSR**: All modules enabled as the default configuration. The ablation results (Fig. 5–6) show that the five modules of TDSR are complementary and non-redundant. **Value-guided region selection**. Disabling it (*Random Region Only*) causes the steepest decline, as the planner no longer attends to salient areas, dropping scores to CIDEr 48.62, and BLEU-4 7.45. **Redundancy penalty & depth-aware reward.** Removing either one slows convergence and yields repetitive or shallow sentences, with final scores stalled around CIDEr 94.64/96.10 and BLEU-4 11.8/11.6. **Early termination.** When always running the full 10 steps, the model initially rivals the complete framework but then over-generates, causing CIDEr to fall from 96.34 to 79.41 and BLEU-4 to 9.34. These drops underline that each component is essential for maintaining both

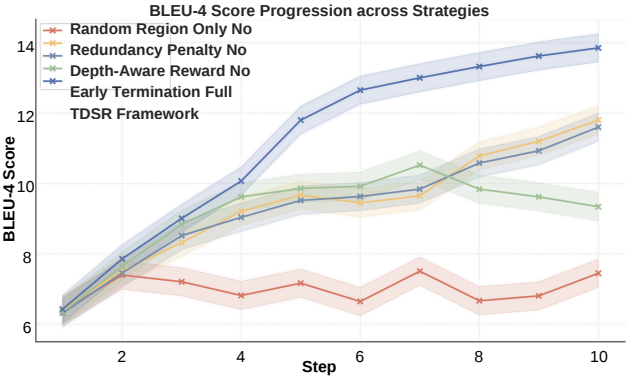

Figure 5: Step-wise BLEU-4 score under ablation settings. Full TDSR achieves the highest and most stable performance; removing value estimation or early stopping severely degrades output fluency.

descriptiveness and coherence.

Conclusion

We propose **TDSR**, a top-down semantic-refinement framework that reformulates image captioning as a coarse-to-fine planning task. Driven by an MCTS planner, TDSR first drafts a global caption and then incrementally enriches it with visually grounded details. Three key techniques, i.e., (i) a lightweight value network, (ii) redundancy-aware early stopping, and (iii) adaptive rollout depth, jointly deliver high caption quality at modest computational cost. Across detail, compositionality, and hallucination benchmarks, TDSR consistently raises factual accuracy, descriptive richness, and robustness to visual perturbations. Ablation experiments show that removing any one component leads to sizable drops in CIDEr and BLEU-4, underscoring their complementarity.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMA-JBGS202517, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012985, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006, and in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11.

References

Alikhani, M.; Sharma, P.; Li, S.; Soricut, R.; and Stone, M. 2020. Cross-modal Coherence Modeling for Caption Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6525–6535. Association for Computational Linguistics.

Bai, J.; Bai, S.; and et al. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012a. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1): 1–43.

Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012b. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1): 1–43.

Bugliarello, E.; and Elliott, D. 2021. The Role of Syntactic Planning in Compositional Image Captioning. arXiv:2101.11911.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023a. VLP: A Survey on Vision-language Pre-training. *Machine Intelligence Research*, 20(1): 38–56.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023b. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. *arXiv:2310.09478*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023c. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Dong, H.; Li, J.; Wu, B.; Wang, J.; Zhang, Y.; and Guo, H. 2024. Benchmarking and Improving Detail Image Caption. *arXiv preprint arXiv:2405.19092*.

Elman, J. L. 1990. Finding Structure in Time. *Cognitive Science*, 14(2): 179–211.

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2020. Image Captioning: Transforming Objects into Words. arXiv:1906.05963.

Hua, H.; Liu, Q.; Zhang, L.; Shi, J.; Sooye, K.; Zhang, Z.; Wang, Y.; Zhang, J.; and Luo, J. 2024a. FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity. *arXiv preprint arXiv:2411.15411*.

Hua, H.; Liu, Q.; Zhang, L.; Shi, J.; Zhang, Z.; Wang, Y.; Zhang, J.; and Luo, J. 2024b. FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity. arXiv:2411.15411.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2015. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. arXiv:1511.07571.

Kemmerling, M.; Lütticke, D.; and Schmitt, R. H. 2023. Beyond games: a systematic review of neural Monte Carlo tree search applications. *Applied Intelligence*, 54(1): 1020–1046.

Kocsis, L.; and Szepesvári, C. 2006. Bandit Based Monte-Carlo Planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML)*, 282–293. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, 1097–1105.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. arXiv:2304.08485.

Mefford, J. A.; Zhao, Z.; Heilier, L.; Xu, M.; Zhou, G.; Mace, R.; Sloane, K. L.; Sheppard, S. M.; and Glenn, S. 2023. Varied performance of picture description task as a screening tool across MCI subtypes. *PLOS Digital Health*, 2(3): e0000197.

OpenAI. 2023. GPT-4V(ision) System Card. https://openai.com/research/gpt-4v-system-card.

Peng; Ruotian; He; Haiying; Wei; Yake; Wen; Yandong; and Hu, D. 2025a. Patch Matters: Training-free Fine-grained Image Caption Enhancement via Local Perception. *arXiv preprint arXiv:2504.06666*.

Peng, R.; He, H.; Wei, Y.; Wen, Y.; and Hu, D. 2025b. Patch Matters: Training-free Fine-grained Image Caption Enhancement via Local Perception. arXiv:2504.06666.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *ArXiv*, abs/2306.

Pi, R.; Zhang, J.; Zhang, J.; Pan, R.; Chen, Z.; and Zhang, T. 2024. Image Textualization: An Automatic Framework for Creating Accurate and Detailed Image Descriptions. arXiv:2406.07502.

Pramanick, S.; Han, G.; Hou, R.; Nag, S.; Lim, S.-N.; Ballas, N.; Wang, Q.; Chellappa, R.; and Almahairi, A. 2023. Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model. *arXiv preprint arXiv:2312.12423*.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: Wiley-Interscience.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Technical report, OpenAI.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1179–1195.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2023. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. arXiv:2312.03818.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; and Dai, J. 2023. VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv:2305.11175.

Wiher, G.; Meister, C.; and Cotterell, R. 2022. On Decoding Strategies for Neural Text Generators. arXiv:2203.15721.

Yao, J.; Zhang, J.; Pan, X.; Wu, T.; and Xiao, C. 2024. DepthSSC: Monocular 3D Semantic Scene Completion via Depth-Spatial Alignment and Voxel Adaptation. arXiv:2311.17084.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Yarats, D.; and Lewis, M. 2018. Hierarchical Text Generation and Planning for Strategic Dialogue. arXiv:1712.05846.

You, H.; and et al., H. Z. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. arXiv:2310.07704.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704*.

Zhang, J.; Cai, K.; Fan, Y.; Wang, J.; and Wang, K. 2025a. CF-VLM:CounterFactual Vision-Language Fine-tuning. arXiv:2506.17267.

Zhang, J.; Cai, K.; Fan, Y.; Wang, J.; and Wang, K. 2025b. CF-VLM:CounterFactual Vision-Language Fine-tuning. arXiv:2506.17267.

Zhang, J.; Cai, K.; Yang, J.; and Wang, K. 2025c. Learning Dynamics of VLM Finetuning. arXiv:2510.11978.

Zhang, J.; Cai, K.; Zeng, Q.; Liu, N.; Fan, S.; Chen, Z.; and Wang, K. 2025d. Failure-Driven Workflow Refinement. arXiv:2510.10035.

Zhang, J.; Fan, Y.; Cai, K.; Huang, Z.; Sun, X.; Wang, J.; Tang, C.; and Wang, K. 2025e. DrDiff: Dynamic Routing Diffusion with Hierarchical Attention for Breaking the Efficiency-Quality Trade-off. arXiv:2509.02785.

Zhang, J.; Fan, Y.; Cai, K.; Sun, X.; and Wang, K. 2025f. OSC: Cognitive Orchestration through Dynamic Knowledge Alignment in Multi-Agent LLM Collaboration. arXiv:2509.04876.

Zhang, J.; Fan, Y.; Cai, K.; Yang, J.; Yao, J.; Wang, J.; Qu, G.; Chen, Z.; and Wang, K. 2026. Why Keep Your Doubts to Yourself? Trading Visual Uncertainties in Multi-Agent Bandit Systems. arXiv:2601.18735.

Zhang, J.; Fan, Y.; Lin, W.; Chen, R.; Jiang, H.; Chai, W.; Wang, J.; and Wang, K. 2025g. GAM-Agent: Game-Theoretic and Uncertainty-Aware Collaboration for Complex Visual Reasoning. arXiv:2505.23399.

Zhang, J.; Huang, Z.; Fan, Y.; Liu, N.; Li, M.; Yang, Z.; Yao, J.; Wang, J.; and Wang, K. 2025h. KABB: Knowledge-Aware Bayesian Bandits for Dynamic Expert Coordination in Multi-Agent Systems. In *Forty-second International Conference on Machine Learning*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5575–5584.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *ICLR*.

Zohourianshahzadi, Z.; and Kalita, J. K. 2021. Neural attention for image captioning: review of outstanding methods. *Artificial Intelligence Review*, 55(5): 3833–3862.