Linking Perception, Confidence and Accuracy in MLLMs

Supplementary Material

A. Prompt Templates

In this section, we provide the specific prompt templates used within our proposed framework. As illustrated in Figure 2, our *Confidence-Aware Test-Time Scaling* (CA-TTS) system employs an Expert Model to fulfill specific roles (specifically as a **Voter** and a **Critic**) to coordinate the self-consistency and self-reflection modules.

A.1. Voter Expert Prompt

The Voter Expert is utilized during the confidence voting process of the *Self-Consistency* phase. It is tasked with analyzing the image and question to assign normalized confidence probabilities to a provided list of candidate options. The specific prompt structure is shown in Figure 5.

A.2. Critic Expert Prompt

The Critic Expert is engaged during the *Self-Reflection* phase. It evaluates the model's initial answer and confidence level to provide a constructive critique. This critique guides the Multi-Modal Large Language Model (MLLM) to rethink and refine its answer. The prompt structure is detailed in Figure 6.

B. Training Datasets Details

Previous research [33] has shown that efficient training driven by a small amount of high-quality data can significantly elicit the scaling ability of models. Inspired by this, we aim to construct a refined and efficient Reinforcement Learning (RL) training set, denoted as D_{RL}.

To establish the source data pool D_{source}, we aggregated data from six public benchmarks, comprising three mathematical reasoning datasets and three general-purpose VQA datasets. These benchmarks include:

- **MathVerse** [64]: A rigorous visual mathematical benchmark designed to decouple visual and textual dependencies, evaluating true multimodal reasoning capabilities across diverse diagrams.
- **DynaMath** [65]: A dynamic visual benchmark that assesses model generalization in mathematical reasoning by programmatically generating diverse problem variations.
- **LogicVista** [51]: A comprehensive benchmark focused on logical reasoning in multimodal contexts, spanning tasks from puzzle solving to complex diagrammatic analysis.
- **RealWorldQA** [50]: A benchmark evaluating real-world spatial understanding and physical reasoning capabilities, primarily derived from challenging real-world environments.
- **MMBench** [29]: An all-around multimodal evaluation pipeline that utilizes a circular evaluation strategy to robustly assess perception and reasoning across diverse tasks.
- **MMMU-Pro** [61]: A robust and refined version of the massive multidisciplinary benchmark MMMU, specifically designed to strictly evaluate expert-level reasoning by filtering out text-shortcut samples.

The aggregated source pool D_{source} contains a total of $N = 11764$ data points. To construct the final D_{RL}, we implemented a two-phase process involving strict data filtering followed by visual augmentation.

Phase 1: Data Filtering. To ensure high training efficiency, we defined a set of strict manual filtering criteria, $\mathcal{C}$, which comprehensively evaluates each data point d based on its **Quality**, **Difficulty**, and **Diversity**. We apply these criteria to filter out a high-quality intermediate set D_{Filtered}:

$$D_{\text{Filtered}} = \{d \in D_{\text{source}} \mid \text{Evaluate}(d, \mathcal{C}) = \text{True}\} \quad (11)$$

Here, $\text{Evaluate}(d, \mathcal{C})$ represents the manual assessment process. Through this procedure, we selected $M = 1936$ high-quality data points from the source pool.

Phase 2: Data Augmentation and Image-Pair Construction. Subsequently, to enhance the model's robustness and perception of critical visual regions, we applied specific visual augmentations to D_{Filtered}. We utilize intermediate-layer attention maps, $\text{Attn}_{\text{CLIP}}$, from the CLIP [37] visual encoder to identify core semantic regions within an image i. We define a noise-injection function $\mathcal{G}_{\text{noise}}$ that applies noise to the original image based on its attention map, yielding a perturbed image $i' = \mathcal{G}_{\text{noise}}(i, \text{Attn}_{\text{CLIP}}(i))$.

The final RL training set D_{RL} expands each filtered data point $d_f = (i, q, a)$ into a tuple containing an *[Original Image, Noised Image]* pair. The formal definition is as follows:

$$\begin{aligned} D_{\text{RL}} = \{((i, i'), q, a) \mid (i, q, a) \in D_{\text{Filtered}}, \\ \text{where } i' = \mathcal{G}_{\text{noise}}(i, \text{Attn}_{\text{CLIP}}(i))\} \end{aligned} \quad (12)$$

This resulting dataset D_{RL}, containing $M = 1936$ data pairs, is employed in our subsequent RL training to teach the model to robustly locate and reason about key information even in the presence of visual disturbances.

C. Additional Results

C.1. Tables

a) Generalizability on Emerging Reasoning-Capable MLLMs. To further scrutinize the universality and robust-

VOTER EXPERT PROMPT:

Image: {image}
Question: {question}

Look at the image carefully, and you will be given a list of candidate options: {options_list}.
Generate a normalized confidence (probability) score for each option in this list.
The order of the output probabilities must strictly correspond to the order of the options in options_list.
The sum of all probabilities **must equal 1.**

Your output must strictly follow the format below, and must **only** be this array. Do not include any other text, labels, or explanations:
[p_1, p_2, ..., p_n]

Figure 5. The prompt template used for the Voter Expert. The model acts as a discriminator to assign probability scores to candidate choices, facilitating confidence-weighted voting.

CRITIC EXPERT PROMPT:

Given the following information:

Image: {image}
Question: {question}
Model Answer: {model_answer}
Model Confidence: {confidence}

Please generate a self-reflection critique according to the given information above. Your output must strictly follow the format below, without any other text or explanation:

Critique: Based on this question, your answer is "{model_answer}", <Please fill in your concise, objective critique of this answer here, for example, questioning its accuracy, relevance, or completeness>

Figure 6. The prompt template used for the Critic Expert. This prompt induces the expert model to generate a critique based on the original input and the model's initial low-confidence response, aiding in the self-reflection process.

ness of our approach, we extended our evaluation to *Qwen3-VL-2B-Thinking*, a representative of the latest MLLMs equipped with intrinsic chain-of-thought (CoT) capabilities. This experiment investigates a pivotal question: does CA-TTS provide additive value to models that already possess optimized "thinking" processes?

As detailed in Table 5, the answer is affirmative. CA-TTS consistently outperforms both the standard Majority Voting and the DeepConf baseline across all evaluated benchmarks. Notably, our framework achieves a commanding **Overall Avg of 74.41**, securing a clear lead over the Majority Voting baseline (71.09).

The performance gains are particularly pronounced in reasoning-intensive domains. On *Math-Vista*, CA-TTS elevates the score to **83.81**, significantly surpassing the Pass@1 baseline of 73.60. Moreover, on the comprehensive *MMMU* benchmark, our method achieves a remarkable **79.63** (vs. 75.46 for Majority Voting). These results suggest that even for models with native CoT designs, CA-TTS effectively modulates the reasoning trajectory to correct errors and refine outputs, thereby unlocking a higher ceiling of performance in both mathematical reasoning and general visual understanding.

b) Additional Model Scaling Results. While the main text visualizes the scaling trends primarily on the *Math-Vision* dataset due to space constraints, here we present the comprehensive numerical data across all four benchmarks in Table 6. This detailed breakdown allows for a granular analy-

Table 5. **Performance on Qwen3-VL-2B-Thinking.** Comparison of CA-TTS against other TTS baselines. Abbreviations follow previous tables. Best results are **bold**, and second-best results are underlined. The "Overall Avg" represents the mean of the "ALL" scores across the four datasets.

Methods	Math Reasoning						General VQA						Overall
	Math-Vista$_{testmini}$			Math-Vision$_{testmini}$			MMStar			MMMU			
(Qwen3-VL-2B-Think)	OE	MC	ALL	OE	MC	ALL	PE	RE	ALL	STEM	HASS	ALL	**Avg.**
Pass@1	64.95	82.51	73.60	47.69	50.00	48.82	62.79	56.00	61.24	58.84	66.95	61.40	61.27
Majority Voting	72.30	89.86	80.95	**60.87**	62.50	61.70	67.10	66.49	66.24	72.90	81.01	75.46	71.09
DeepConf	72.77	87.92	80.24	52.17	62.50	57.45	67.36	68.09	66.24	73.83	82.28	75.93	69.97
Ours (CA-TTS)	**77.46**	**90.34**	**83.81**	**60.87**	**66.67**	**63.83**	**69.71**	**75.53**	**70.36**	**78.50**	**83.54**	**79.63**	**74.41**

Table 6. **Detailed Scaling Results.** Comparison of performance with varying number of samples (N). Best results are **bold**.

Dataset	Method	Number of Samples (N)					
		1	**2**	**4**	**8**	**16**	**32**
Math-Vision$_{testmini}$	Majority Voting	26.38	25.00	29.93	30.26	30.92	34.41
	DeepConf	26.38	26.64	28.95	28.95	30.26	32.15
	Ours (CDRL+CA-TTS)	**30.60**	**34.21**	**41.78**	**43.75**	**46.38**	**48.44**
Math-Vista$_{testmini}$	Majority Voting	**64.70**	67.78	68.56	68.39	73.11	75.60
	DeepConf	**64.70**	67.89	67.69	69.28	71.32	75.50
	Ours (CDRL+CA-TTS)	64.15	**72.15**	**73.80**	**74.19**	**77.85**	**80.60**
MMStar	Majority Voting	60.20	60.55	63.68	69.00	64.56	69.96
	DeepConf	60.20	59.63	62.88	61.87	64.08	69.77
	Ours (CDRL+CA-TTS)	**61.21**	**63.30**	**71.16**	**71.27**	**70.87**	**74.03**
MMMU	Majority Voting	48.77	54.14	55.84	57.18	58.87	58.62
	DeepConf	48.77	54.61	56.31	56.24	58.25	58.00
	Ours (CDRL+CA-TTS)	**52.63**	**61.96**	**65.65**	**66.28**	**68.72**	**69.58**

sis of how performance evolves with the number of samples (N) ranging from 1 to 32.

The results clearly demonstrate the superior scaling efficiency of **Our Method (CDRL+CA-TTS)**. Compared to standard Majority Voting and DeepConf, our approach consistently achieves higher accuracy gains as the sample size increases. Notably, on the *MMMU* and *Math-Vision* datasets, our method maintains a substantial lead at every scaling step. For instance, at $N = 32$, CDRL+CA-TTS outperforms the best baseline on Math-Vision by a margin of over 14% (48.44% vs 34.41%), validating that our confidence-aware test-time scaling strategy effectively leverages increased test-time computation to resolve complex visual reasoning tasks.

c) Additional Results of Performance by Using Different Expert Models. Quantitative results in Table 7 further substantiate these observations. The framework exhibits a positive correlation between the capability of the expert model and the final performance. Specifically, Gemini-2.5-Pro achieves the state-of-the-art performance with an **Overall Avg of 64.85**, significantly outperforming the Majority Voting baseline (55.35). It dominates across most benchmarks, particularly in *Math-Vista* (79.50) and *MM-Star* (71.27). GPT-5 follows closely as the second-best performer with an Overall Avg of 64.14, while demonstrating superior capability in the *MMMU* benchmark (66.51 compared to Gemini's 66.28), specifically in STEM tasks. Notably, even smaller models like Qwen2.5-VL-7B provide a clear boost over the baseline (57.16 vs. 55.35), validating the effectiveness of our framework regardless of the expert model's size. This trend confirms that our approach effectively leverages the distinct strengths of various foundation models, from reasoning-heavy tasks in Math Reasoning datasets to general visual question answering.

d) Compatibility with Existing Frameworks. To validate whether our proposed CA-TTS can serve as a universal plug-in to enhance existing models, we integrated it with several state-of-the-art baselines, utilizing Gemini-2.5-Pro as the expert model. As shown in Table 8, under identical settings, CDRL combined with CA-TTS achieves the

Table 7. **Performance of Different Expert Models.** Abbreviations: OE (Open-Ended), MC (Multi-Choice), PE (Perception), RE (Reasoning), STEM (Science, Technology, Engineering, and Mathematics), HASS (Humanities, Arts, and Social Sciences), and ALL (Overall). Best results are **bold**, and second-best results are underlined. The "Overall Avg" column is calculated as the average of the "ALL" scores from the four datasets.

Models/Datasets	Math Reasoning						General VQA						Overall
	Math-Vista$_{testmini}$			Math-Vision$_{testmini}$			MMStar			MMMU			
	OE	MC	ALL	OE	MC	ALL	PE	RE	ALL	STEM	HASS	ALL	Avg
Majority Voting	68.39	73.73	69.80	26.20	33.24	30.08	60.20	71.20	64.00	53.20	62.50	57.53	55.35
Qwen2.5-VL-7B	69.49	77.99	73.00	30.91	34.03	32.57	65.33	71.47	64.59	53.35	65.99	58.46	57.16
Qwen2.5-VL-72B	67.90	78.62	74.20	28.52	37.70	34.21	61.00	71.60	64.33	55.08	61.56	58.11	57.71
Qwen-VL-Max	69.35	79.14	74.47	35.71	36.13	35.97	61.20	73.40	64.80	56.25	67.05	61.45	59.17
GPT-5	73.74	82.71	78.20	**38.94**	43.93	41.45	64.00	77.00	70.40	**61.85**	72.62	**66.51**	64.14
Gemini-2.5-Pro	**74.19**	**84.67**	**79.50**	38.44	**45.60**	**42.35**	**67.80**	**77.20**	**71.27**	59.90	**73.50**	66.28	**64.85**

best overall average performance (64.9%), outperforming the second-best framework (We-Think) by a clear margin of 3.9%. This demonstrates that calibrated confidence provides a fundamentally stronger foundation for effective test-time scaling.

e) Ablation Study and Inference Cost. We conducted a comprehensive ablation study to isolate the contribution of each module within the CA-TTS framework and analyzed their corresponding inference time costs. As detailed in Table 9, the removal of any module leads to a noticeable performance drop, validating our carefully coordinated design. Regarding inference efficiency, compared to the standard Majority Voting baseline, CA-TTS consumes only 0.91× more time while delivering a substantial 8.4% increase in average accuracy, proving it is highly efficient within the paradigm of test-time scaling.

C.2. Visualization

As shown in Figure 7, 8 and 9, we illustrates additional visualized scaling results. Detailed analysis can be found in figure captions, Section 4.4 and Appendix C.1. Additionally, Figure 10 illustrates the generality of confidence miscalibration under visual degradation.

C.3. More Case Studies

Figures 11 and 12 illustrate additional case studies of our framework. Consistent with the analysis in the main text 4.5, our approach exhibits a distinct superiority over tree-based paradigms (e.g., ToT). While ToT is often characterized by a cumbersome and non-decouplable architecture with complex branching logic, it is further limited by single-round verification, leading to an excessive reliance on the inherent capability of the Expert model.

D. Discussions

Justification for Confidence Calculation. In our implementation, we adopt the Normalized Mean Log-Probability (NMLP) as the core metric, consistent with DeepConf [10].

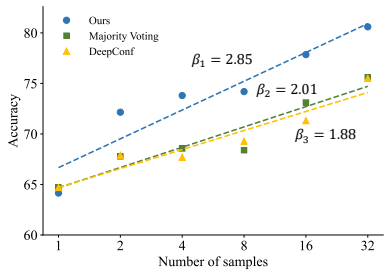

Figure 7. **Scaling Superiority on Math-Vista.** This plot compares the accuracy of Our method (blue dots) against the Majority Voting (green squares) and DeepConf (yellow triangles) baselines as the number of samples increases from 1 to 32. The results show that the slope of our trendline ($\beta_1 = 2.85$) is substantially steeper than those of the baseline methods ($\beta_2 = 2.01$ and $\beta_3 = 1.88$), indicating that the performance advantage and potential of our method widens as more samples are provided.

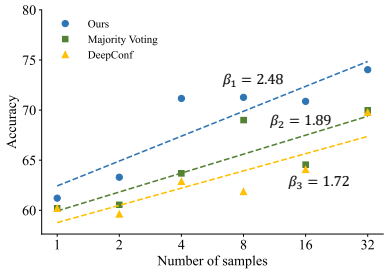

Figure 8. **Scaling Superiority on MMStar.** This plot compares the accuracy of Our method (blue dots) against the Majority Voting (green squares) and DeepConf (yellow triangles) baselines as the number of samples increases from 1 to 32. The results show that the slope of our trendline ($\beta_1 = 2.48$) is substantially steeper than those of the baseline methods ($\beta_2 = 1.89$ and $\beta_3 = 1.72$), indicating that the performance advantage and potential of our method widens as more samples are provided.

Although we experimented with alternative aggregation

Table 8. **Baseline Models Results with Gemini 2.5 Pro.** Comparison of various frameworks when equipped with our CA-TTS plug-in.

Framework + CA-TTS	Math-Vista	Math-Vision	MMStar	MMMU	Average
Qwen2.5-VL-7B	75.8	39.1	65.6	59.2	59.9
R1-OneVision	70.9	38.2	60.9	59.8	57.5
VL-Rethinker	74.3	38.1	64.3	57.1	58.5
We-Think	77.1	39.8	66.8	60.2	61.0
Ours (CDRL)	**79.5**	**42.4**	**71.3**	**66.3**	**64.9**

Table 9. **Ablation Study and Inference Cost.** Evaluation of individual modules and their corresponding time consumption.

Settings	Math-Vista	Math-Vision	MMStar	MMMU	Avg.	Time Cost (s)
Ours (CA-TTS)	**79.5**	**42.4**	**71.3**	**66.3**	**64.9**	11.03
w/o Self-Consistency	70.5	35.0	67.4	58.4	57.8	4.66
w/o Self-Reflection	74.6	37.9	69.1	65.5	61.8	8.55
w/o Self-Check	74.2	39.1	70.5	65.9	62.4	8.85
Majority Voting	69.8	30.1	69.0	57.5	56.6	5.76

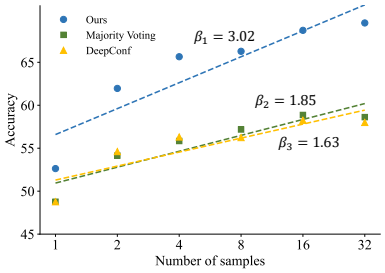

Figure 9. **Scaling Superiority on MMMU.** This plot compares the accuracy of Our method (blue dots) against the Majority Voting (green squares) and DeepConf (yellow triangles) baselines as the number of samples increases from 1 to 32. The results show that the slope of our trendline ($\beta_1 = 3.02$) is substantially steeper than those of the baseline methods ($\beta_2 = 1.85$ and $\beta_3 = 1.63$), indicating that the performance advantage and potential of our method widens as more samples are provided.

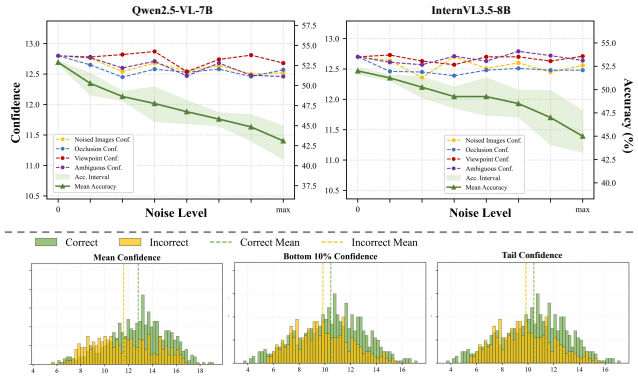

Figure 10. **Generality of Confidence Miscalibration under Visual Degradation.** We tested the impact of various visual degradation types on two representative MLLMs. The overlapping confidence distributions indicate that MLLMs struggle to calibrate their confidence correctly when handling degraded visual inputs, highlighting the necessity of our Confidence-Aware framework.

strategies, such as *Tail Confidence* (prioritizing final tokens) and *Bottom-Group Confidence* (averaging the lowest probability tokens based on the "weakest link" theory), our empirical results consistently favor the global average confidence. We attribute this to the holistic nature of chain-of-thought reasoning, where the validity of the final answer is contingent upon the semantic coherence of the entire reasoning path. Furthermore, local metrics proved overly sensitive to noise, often penalizing high-entropy tokens associated with benign stylistic choices (e.g., synonyms or formatting) rather than factual errors. Consequently, the global average acts as a robust smoothing filter, effectively representing the model's overall certainty regarding the semantic integrity of the output.

Future Prospects. The vast potential of Test-Time Scaling (TTS) remains largely untapped. In this work, we have empirically demonstrated that confidence awareness acts as a pivotal catalyst in optimizing test-time scaling laws. Looking ahead, we envision extending the principles of our CA-TTS framework to broader paradigms. Specifically, the high-quality reasoning trajectories filtered by our system can serve as gold-standard supervision for *Data Synthesis Frameworks* and *Reinforcement Fine-Tuning*, effectively closing the loop between inference-time scaling and post-training improvements. Furthermore, integrating this dynamic evaluation capability into *Agent-Evolving* systems and *Self-Play* mechanisms holds the promise of enabling models to autonomously refine their strategies. We remain

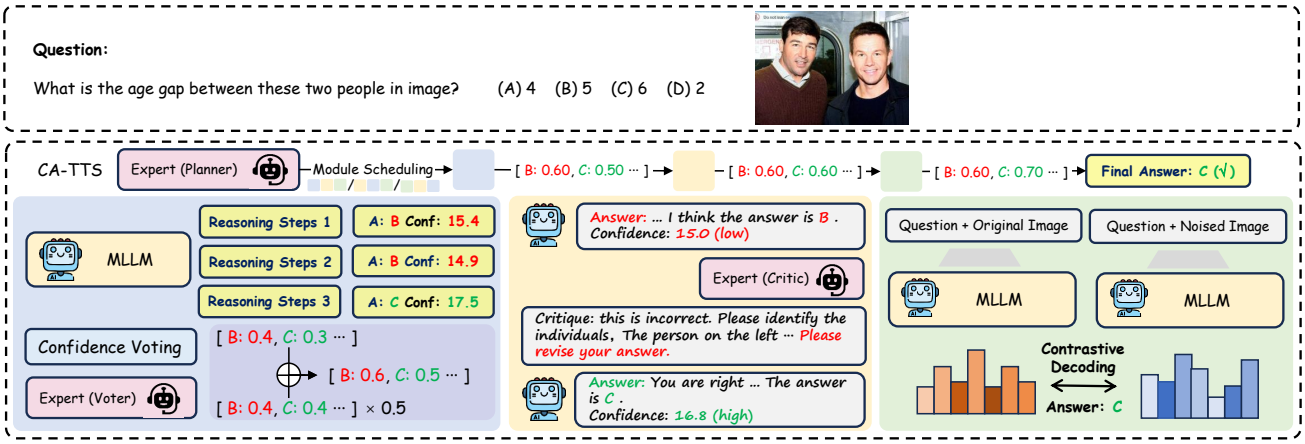

Figure 11. **A case study of CA-TTS on MMStar.** Our method (bottom) demonstrates a multi-stage, resilient process: an initial error from **Self-Consistency** (Answer: B) is corrected by **Self-Reflection** (Answer: C) and confirmed by **Self-Check**.

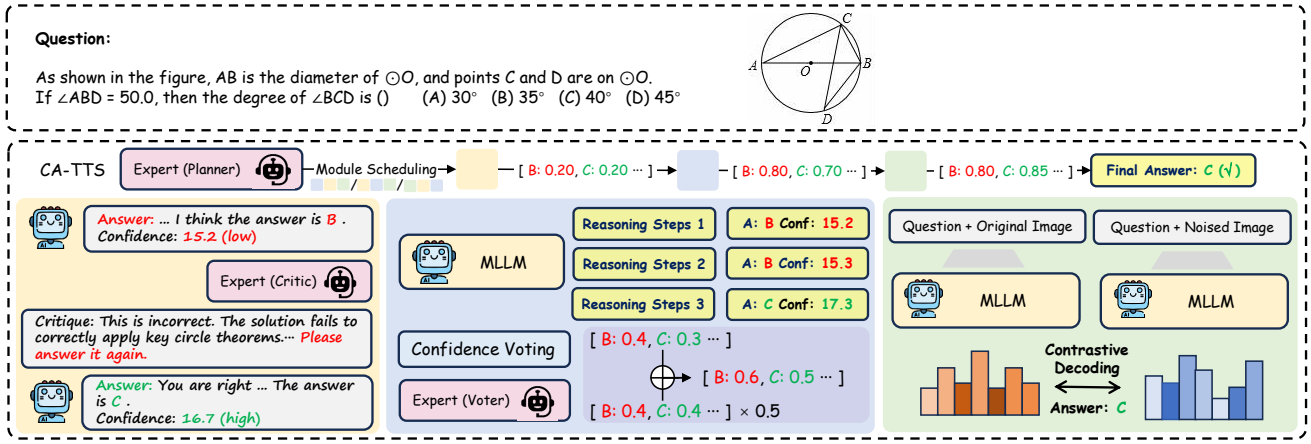

Figure 12. **A case study of CA-TTS on Math-Vista.** Our method (bottom) demonstrates a multi-stage, resilient process: a error from **Self-Reflection** and **Self-Consistency**(Answer: B) is corrected by **Self-Check** (Answer: C).

committed to exploring these avenues to further unlock the versatility and impact of TTS in multimodal intelligence.