TOWARDS MULTIMODAL GENERATIVE ACTIVE LEARNING: EFFICIENT LEARNING VIA UNCERTAINTY DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Active learning (AL) aims to improve data efficiency by selectively querying the most informative samples for annotation in deep learning. However, existing AL research focuses predominantly on unimodal and multimodal discriminative learning, leaving multimodal generative learning largely unexplored. We introduce the first framework for *multimodal generative active learning*, where the learner must actively acquire textual ground-truth for multimodal inputs rather than conventional labels. This setting differs fundamentally from existing multimodal AL paradigms, which typically rely on predefined labels obtained through annotation, an assumption that does not hold in multimodal generation. To this end, we propose an innovative AL algorithm that jointly exploits uncertainty and diversity through uncertainty dynamics, facilitating effective and representative sample selection with improved acquisition efficiency. Empirical results on benchmark datasets demonstrate that our approach substantially improves annotation efficiency, achieving comparable performance with up to 33% fewer annotated samples.

1 INTRODUCTION

Modern deep learning models have become increasingly prevalent across diverse domains and applications, but require substantial amounts of human feedback and annotated data. Active learning (AL) (Cohn et al., 1996; Settles, 2009) is a promising approach to alleviate the labeling burden by selecting the most informative samples for annotation. The effectiveness of AL in improving data efficiency has been well established both theoretically (Zhu & Nowak, 2022a) and empirically for deep learning models and large language models (Sener & Savarese, 2017; Ash et al., 2019; Saran et al., 2023; Bhatt et al., 2024; Gupte et al., 2024; Zhang & Zhu, 2025; Xia et al., 2025). Despite its success, extending AL to multimodal generation setting presents additional challenges that are less prominent in conventional scenarios.

In unimodal AL settings, sample acquisition primarily relies on information from a single modality to identify informative samples. Extending this paradigm to multimodal settings further allows the acquisition process to exploit complementary information across modalities. Existing research in multimodal AL has largely focused on classification and cross-modal alignment tasks. For multimodal classification, several approaches have been explored (Zhang et al., 2024; Shen et al., 2023). Among them, Shen et al. (2023) proposes a representative method that assumes pre-aligned vision-language pairs and performs active selection over paired multimodal samples, thereby treating multimodal AL as standard classification-oriented AL over composite inputs. More recent research, such as Zhang & Zhu (2025), formulates the task objective as learning cross-modal correspondence. However, to the best of our knowledge, AL in the multimodal generation setting has not been systematically studied.

In this paper, we introduce the first setting of *multimodal generative active learning*, where the learner begins with multimodal inputs without ground-truth responses and must actively acquire textual annotations for multimodal generation. Unlike traditional multimodal AL, our setting presents two qualitatively new challenges: (i) *hybrid acquisition*, since multimodal generation requires autoregressively producing free-form responses conditioned on multiple modalities, making it non-trivial to jointly leverage uncertainty and diversity for sample acquisition (Perlitz et al., 2023); and (ii) *generation informativeness quantification*, since the output in our setting is not defined over a fixed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

output space, leading to difficulties in translating model generation into a sample-level acquisition signal. To bridge the gap between existing AL and our setting, we propose a novel multimodal AL algorithm that approximates uncertainty dynamics across model layers to identify informative samples, followed by a diversity-aware selection phase that jointly considers uncertainty and diversity. This overall strategy enables the efficient selection of representative and informative multimodal samples under a limited annotation budget.

Our contributions. Our main contributions are as follows:

- We introduce *multimodal generative active learning*, a largely unexplored setting beyond unimodal and existing multimodal AL settings.
- We propose a novel algorithm that leverages uncertainty dynamics to guide sample selection, while jointly incorporating a diversity principle, resulting in an efficient data acquisition process.
- We perform extensive experiments on benchmark datasets. Our method consistently achieves competitive or superior performance while demonstrating substantial annotation savings of up to 33%.

Paper organization. The rest of this paper is organized as follows. Section 2 introduces the problem formulation and discusses the key challenges in multimodal generative active learning. Section 3 presents our algorithm. In Section 4, we conduct extensive experiments to evaluate our method and provide additional analyses. Section 5 concludes the paper. Related work, implementation details, and supplementary experiments are deferred to the Appendix.

2 Preliminaries

2.1 Notations

We study the multimodal generative active learning setting. Given a multimodal dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^N$, we initialize the unlabeled dataset with all data points in $\mathcal{D}$, where I_i and T_i denote the raw vision features and raw language features for the i-th data point, respectively. The ground-truth textual responses $\{(Y_i)\}_{i=1}^N$ are initially unknown. We consider multimodal generation tasks in both open-ended and closed-ended/multiple-choice settings.

The active learning is performed over R iterations. At the r-th iteration, the active learning algorithm selects an informative batch of K data points from the unlabeled dataset $\mathcal{D}_r^l \subseteq \mathcal{D}_r$ and subsequently queries their corresponding ground-truth from the oracle $\mathcal{O}$. Then the labeled and unlabeled datasets are updated as: $\mathcal{S}_r \leftarrow \mathcal{S}_{r-1} \cup \mathcal{D}_r^l$, $\mathcal{D}_{r+1} \leftarrow \mathcal{D}_r \setminus \mathcal{D}_r^l$, where $\mathcal{D}_r^l = \{(I_i, T_i, Y_i)\}_{i=1}^K$ denotes the selected samples together with their ground-truth responses obtained from the oracle. The multimodal large language model (MLLM) f_r is trained on $\mathcal{S}_r$ and used to guide the data selection in the next iteration. The objective of multimodal generative active learning is to maximize the performance of MLLM under a fixed annotation budget. We refer to this setup as *multimodal generative active learning.*

2.2 Challenges of Active Learning in Multimodal Generative Settings

Unlike unimodal active learning (Sener & Savarese, 2017; Ash et al., 2019; Saran et al., 2023; Zhang et al., 2026) and conventional multimodal active learning for classification or cross-modal alignment tasks (Shen et al., 2023; Zhang & Zhu, 2025), our setting—*active learning for multimodal generation*—introduces qualitatively new challenges that are absent in previous active learning paradigms.

In unimodal active learning, each data point is represented using a single modality, and the active learning algorithm operates without leveraging information from multiple modalities. Existing multimodal active learning methods are primarily designed for classification or cross-modal alignment tasks, where the objective is to predict a discrete class label or identify the correct cross-modal correspondence. *These assumptions no longer hold in multimodal generation setting*, where the model must autoregressively produce free-form textual responses instead of making a single discriminative prediction. This setting gives rise to two distinctive challenges:

- **Hybrid acquisition.** Compared to classification setting, multimodal generation requires autoregressively producing free-form responses conditioned on multiple modalities. Consequently, jointly leveraging uncertainty and diversity for sample acquisition becomes substantially more challenging (Perlitz et al., 2023), making existing hybrid active learning methods (e.g., BADGE (Ash et al., 2019)) non-trivial to directly transfer to our setting.
- **Generation informativeness quantification.** Unlike classification or cross-modal alignment tasks, where model predictions are typically defined over a fixed output space, multimodal generation produces variable-length autoregressive responses with information distributed across multiple decoding steps. As a result, it is difficult to directly summarize the model's generative behavior into a reliable sample-level acquisition signal, making informative sample identification substantially more challenging than in conventional active learning settings.

2.3 Limitations of Existing Paradigms under Multimodal Generation

As illustrated in Figure 1, existing active learning methods applicable to multimodal generation can be broadly grouped into two paradigms. The first is **Uncertainty-Based Selection**, which prioritizes samples with high uncertainty during autoregressive generation. We extend this paradigm from Perlitz et al. (2023), which introduced uncertainty-based active learning for natural language generation. However, uncertainty estimation requires autoregressive decoding after prefilling, making sample acquisition **computationally inefficient** due to sequential token generation. The second category is **Diversity-Based Selection**. In this approach, learner selects representative and diverse samples to maximize coverage of the unlabeled dataset. Bi et al. (2025) showed that this type of method **may yield inferior performance** on multimodal generation tasks and analyzed the underlying reasons.

To address these challenges and limitations, we present our approach in Section 3.

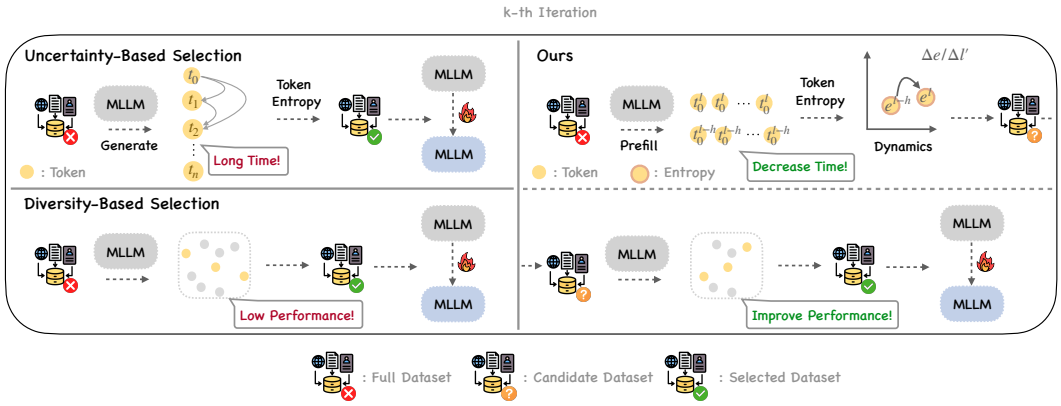


Figure 1: **A Comparison of Active Learning Selection Paradigms.** Existing methods, such as uncertainty-based and diversity-based selection, either incur substantial inference latency due to sequential token generation or suffer from suboptimal performance when relying solely on representation diversity. In contrast, our proposed multimodal generative active learning paradigm first obtains token-level uncertainty in a single prefill pass, avoiding costly sequential generation. We then leverage an approximation of uncertainty dynamics to identify a compact candidate set containing the most uncertain samples. This uncertainty-guided screening narrows the search to highly informative candidates, enabling diversity-aware selection to effectively balance uncertainty and diversity.

3 Methods

In this section, we present our method. The overall framework is illustrated in Figure 1, comprising two main phases: dynamics-guided selection (Section 3.1) and diversity-aware selection (Section 3.2). We provide a detailed illustration of the dynamics-guided selection phase in Figure 2.

3.1 DYNAMICS-GUIDED SELECTION

Dynamics-guided selection aims to identify a subset of samples that exhibit consistently high uncertainty throughout the entire uncertainty dynamics. To improve computational efficiency, this selection process is conducted exclusively during the prefill phase of the multimodal large language model (MLLM), where layer-wise representations are available, rather than during the autoregressive generation phase. To further reduce the computational burden, we approximate the full uncertainty dynamics using uncertainty estimates from selected layers rather than all layers. The selected samples are then passed to the subsequent phase as a candidate set for further diversity-aware selection.

As illustrated in Figure 2, for each data point $d_i = (I_i, T_i)$, we first extract the multimodal feature/representation $\mathbf{H}_i^{(q)}$ from an intermediate layer (i.e., layer q) of pretrained MLLM. In our work, we consider $\forall q \in \{l, l-h\}$, where l denotes the deeper layer and h represents the layer interval. To quantify uncertainty and approximate its dynamics, we measure the uncertainty of each sample using the mean token entropy (Zhao et al., 2020; Perlitz et al., 2023) and estimate its evolution through a finite difference approximation. Formally, let $\phi(\cdot) := \mathrm{softmax}(\mathrm{LMHead}(\mathrm{Norm}(\cdot)))$ denote the token-to-probability mapping of the MLLM. We define the uncertainty of each sample as:

$$U_i^{(q)} := \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left(-\sum_{v=1}^{|\mathcal{V}|} \phi(\mathbf{H}_{i,t}^{(q)})_v \log \phi(\mathbf{H}_{i,t}^{(q)})_v \right), \tag{1}$$

where $\mathcal{V}$ denotes the vocabulary of the MLLM and $\mathcal{T}_i$ represents the set of prefilled tokens for sample d_i. Based on the layer-wise uncertainty signals, we define an approximation of the uncertainty dynamics using a negative first-order backward finite difference approximation:

$$u_i^{'} := -\frac{U_i^{(l)} - U_i^{(l-h)}}{h}, \tag{2}$$

where h is an integer representing the difference in layer indices between the two uncertainty measurements in the language model. A positive u_i', together with high uncertainty levels in deeper layers, suggests that *the sample maintains consistently high uncertainty across layers*. For a negative u_i', uncertainty increases toward deeper layers, reflecting that the model gradually loses confidence during later processing stages. When u_i' is negative but relatively large (i.e., with a relatively small magnitude), *high uncertainty in deeper layers still indicates that the sample remains highly uncertain across layers.* It is worth noting that joint signal $(U_i^{(l)}, u_i')$ only approximates the layer-wise uncertainty evolution across all layers and cannot exactly represent it. We further validate this approximation in Section 4.3.1 by measuring its correlation with the full layer-wise uncertainty evolution.

Based on these observations, we incorporate both the uncertainty signals and an approximation of the layer-wise uncertainty evolution into a unified uncertainty-dynamics score to guide data selection. Before that, we first define a raw uncertainty-dynamics vector as: $\mathbf{J}_i := [u_i, u_i'], \forall d_i \in \mathcal{D}_r$, where $u_i := U_i^{(l)}$. Since the two components capture different aspects of uncertainty dynamics and may have different scales, we normalize each component before aggregation to prevent either term from dominating the resulting score:

$$\bar{\mathbf{J}}_i[j] \leftarrow \frac{\mathbf{J}_i[j] - \min_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j]}{\max_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j] - \min_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j] + \varepsilon}, \quad \forall d_i \in \mathcal{D}_r,\ j = 1, 2. \tag{3}$$

We then construct a unified uncertainty-dynamics score for each data point by aggregating the normalized components:

$$s_i := \bar{\mathbf{J}}_{i,1} + \lambda \cdot \bar{\mathbf{J}}_{i,2}, \tag{4}$$

where λ controls the relative contribution of the uncertainty evolution component. We select the top-B data points with the highest uncertainty-dynamics scores to form the candidate set. It is worth noting that a larger s_i indicates that the corresponding sample exhibits higher selection priority, as it captures samples that consistently exhibit high uncertainty throughout the uncertainty dynamics, in agreement with our previous observations.

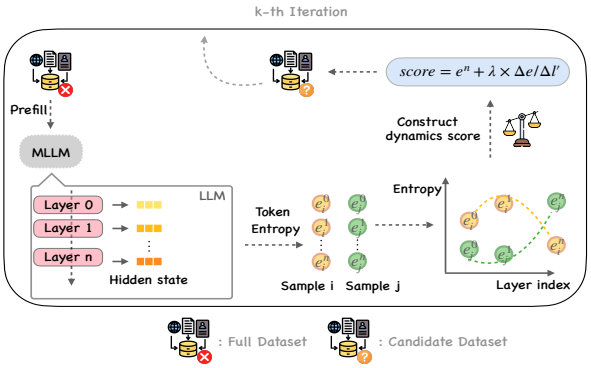


Figure 2: Detailed illustration of the dynamics-guided selection phase of our method. During the prefill stage, we estimate sample uncertainty across layers using mean token entropy, obtained by averaging token-level entropies. In practice, we use only two representative layers from the language model: one shallow layer and the last layer. The resulting uncertainty dynamics estimate how model uncertainty evolves across layers, from which we derive a score to rank unlabeled samples and construct a compact candidate set for subsequent diversity-aware selection.

3.2 DIVERSITY-AWARE SELECTION

To account for both uncertainty and diversity, we perform diversity-aware selection over the candidate set obtained from the preceding phase. The final selected subset is expected to improve performance by balancing diversity and uncertainty, and is then added to the labeled dataset to update the model.

Inspired by a prior diversity-based active learning method in unimodal settings (Sener & Savarese, 2017), we perform diversity-aware selection by constructing a coreset $\mathcal{C}_r$ of size K that, together with the already labeled data, maximizes coverage of the candidate set. We formulate this objective as:

$$\mathcal{C}_r := \underset{\mathcal{C}:|\mathcal{C}|=K,\mathcal{C}\subseteq\mathcal{P}_r}{\arg\min} \max_{d_i\in(\mathcal{P}_r\setminus\mathcal{C})} \min_{d_j\in(\mathcal{S}_{r-1}\cup\mathcal{C})} \text{dist}\left(\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\mathbf{H}_{i,t}^{(l)}, \frac{1}{|\mathcal{T}_j|}\sum_{t\in\mathcal{T}_j}\mathbf{H}_{j,t}^{(l)}\right), \tag{5}$$

where dist is the distance metric and $\mathbf{H}_{i,t}^{(l)}$ denotes the multimodal feature of sample i from a deeper layer l. Based on the analysis in Sener & Savarese (2017), exactly solving the optimization problem in Eq. 5 is NP-Hard. Accordingly, we adopt a greedy approximation to Eq. 5 following the procedure in Zhang & Zhu (2025). The selected coreset is then submitted to the oracle for annotation. It is worth noting that the hyperparameter B (i.e., the size of candidate set) plays an important role in balancing diversity and uncertainty in our method. As B increases, the candidate set becomes larger, and the selection increasingly relies on diversity, eventually approaching pure diversity-based selection. Conversely, a smaller B places greater emphasis on uncertainty, making the selection closer to uncertainty-based selection. We present a complete algorithm in Alg. 1 of the Appendix.

4 EXPERIMENTS

4.1 SETUP

Datasets. We conduct experiments on three multimodal datasets for generative tasks: VQAv2-small, a subset of VQAv2 (Goyal et al., 2017), ChartQA (Masry et al., 2022), and MMBench (Liu et al., 2024). VQAv2-small is a compact subset of the classic VQAv2 benchmark, designed for visual question answering based on given images and natural languages. ChartQA is a benchmark with human-written questions that primarily evaluates models' visual and logical reasoning abilities. MMBench is another multimodal benchmark that objectively evaluates models using multiple-choice questions across a broad range of capabilities. To evaluate the effectiveness of our method across different dataset scales, we conduct experiments on the full MMBench and ChartQA datasets, as well

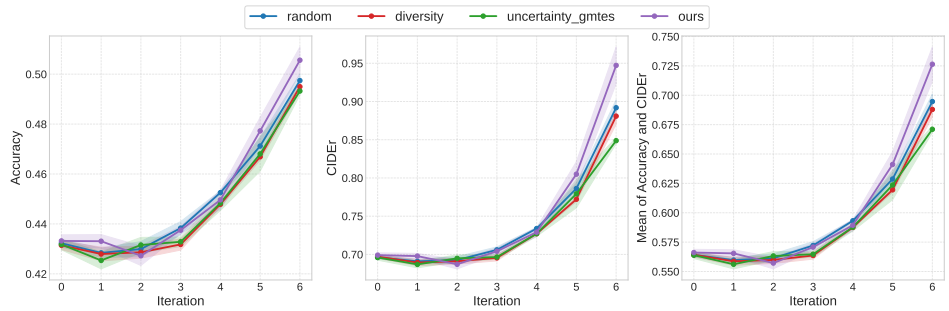


Figure 3: Performance comparison of our proposed method with baseline methods using the SmolVLM-256M model. The x-axis denotes the active learning iteration, while the y-axis reports VQA accuracy (*left*), CIDEr (*middle*), and their mean (*right*) as learning progresses. The experiments are conducted on the VQAv2-6K dataset.

as subsets of ChartQA and VQAv2-small, denoted as ChartQA-8K and VQAv2-6K, respectively. MMBench is evaluated in a closed-ended/multiple-choice setting, whereas VQAv2-6K, ChartQA-8K, and ChartQA are evaluated in open-ended settings. Details on dataset construction and split configurations are provided in Appendix C.

Performance evaluation. In this work, we adopt the same evaluation metrics as those used by Marafioti et al. (2025), namely VQA accuracy (denoted as **Accuracy** hereafter) and **CIDEr**. To measure the first metric, we follow an evaluation protocol similar to that of Lu et al. (2024): we first generate model responses, extract the predicted answers, and then compute accuracy by comparing them against the ground-truth answers. For CIDEr (Vedantam et al., 2015), we use it to assess the similarity between predicted answers and ground-truth answers, following the original evaluation procedure. All main and analysis results are averaged over 4 random runs, with shaded regions in plots indicating 2/3 of a standard deviation.

Baselines. We compare our method with three baselines adapted to our setting: random, diversity, and uncertainty_gmtes. Random is included as a passive learning baseline that acquires samples for annotation without using any selection criterion. As shown in Figure 1, we adapt two representative active learning paradigms—diversity-based and uncertainty-based selection—to our setting for comparison. For the diversity baseline, we apply the k-Center-Greedy algorithm (Sener & Savarese, 2017; Guo et al., 2022), using the model's final-layer hidden states as embeddings for greedy selection. Inspired by Perlitz et al. (2023), we extend the max-entropy acquisition rule to our setting using Mean Token Entropy (MTE), where token-level entropies are averaged across the generated sequence to quantify instance-level uncertainty. We adopt this MTE-based acquisition strategy as our uncertainty baseline, denoted as *uncertainty_gmtes*.

Implementation details. Our algorithm and all baselines are implemented with SmolVLM-256M (Marafioti et al., 2025) and its larger variant, SmolVLM-2.2B, to evaluate performance across different model scales. Further details on the experimental setup are provided in Appendix C.

4.2 Main Results

Open-ended setting. We first consider the open-ended setting, where the model generates free-form answers without being restricted to a predefined set of candidates. We first evaluate our method against three baselines on the VQAv2-6K dataset. As shown in Figure 3, our method generally outperforms the competing baselines across both evaluation metrics. Notably, at the final iteration, our algorithm achieves an accuracy of 50.56% and a CIDEr score of 0.95. Compared with the best-performing baseline, random, which achieves an accuracy of 49.74% and a CIDEr score of 0.89, the improvement is particularly notable in CIDEr, where our method achieves a relative gain of 6.74%.

We next compare our algorithm with the baselines on ChartQA-8K, another open-ended VQA benchmark that requires free-form answer generation based on chart understanding and reasoning.

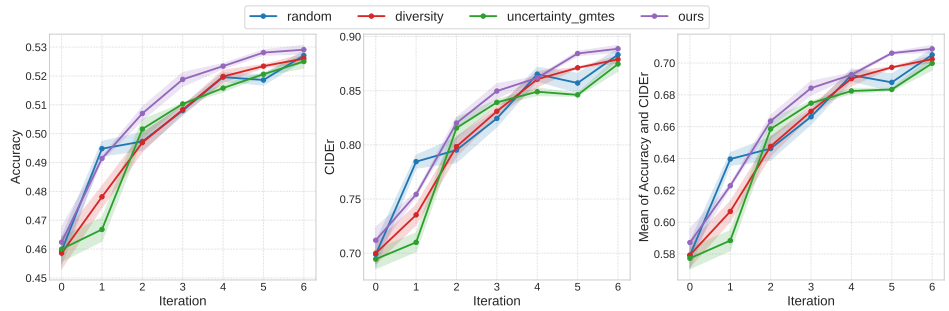


Figure 4: Performance comparison of our proposed method with other baselines on the ChartQA-8K dataset using the SmolVLM-256M model.

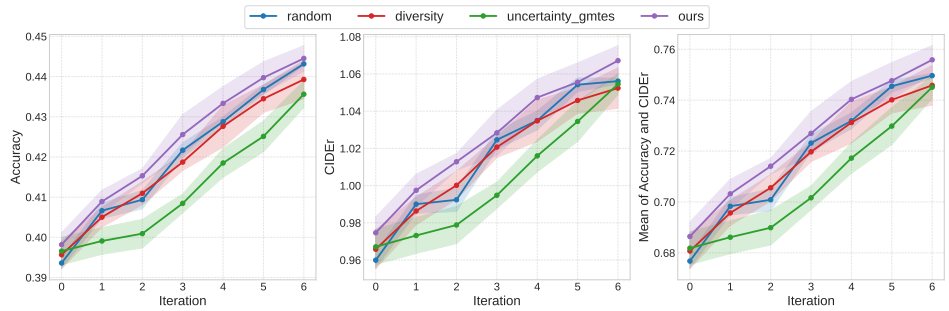


Figure 5: Performance comparison across different active learning methods on the MMBench dataset using the SmolVLM-256M model.

Results on the same evaluation metrics are reported in Figure 4. The results show that our method generally achieves better performance than all baselines on both evaluation metrics. For example, our method achieves an accuracy of 51.88% using only 12% of labeled samples (at iteration 3 in Figure 4), which is comparable to the performance of random and uncertainty_gmtes using 18% of annotated data (iteration 5). This corresponds to a 33.33% reduction in annotation cost while achieving similar performance. In addition, for CIDEr, our method achieves a score of 0.88 with 18% labeled data, surpassing the performance of all baselines at a 21% labeling budget.

Closed-ended/multiple-choice setting. To assess performance in the closed-ended setting, we conduct experiments on MMBench, a multimodal benchmark that requires models to select answers from predefined multiple-choice options. As presented in Figure 5, our proposed method generally surpasses all baselines in both accuracy and CIDEr. It should be noted that our algorithm achieves a mean score of 0.703 across accuracy and CIDEr using only 5% of annotated samples (at iteration 1 in Figure 5), closely matching the diversity baseline (0.706) while outperforming the uncertainty_gmtes and random baselines at iteration 2, corresponding to a 33.33% reduction in annotation cost.

Beyond the main experimental results, we further evaluate our method across different data scales and model variants. These additional experiments are presented in Appendix D.

4.3 Analyses and Ablations

4.3.1 Correlation Analysis for Uncertainty Dynamics Approximation

In this section, we validate the effectiveness of the uncertainty-dynamics score as an approximation of the layer-wise uncertainty evolution in Alg. 1. The approximation is designed to estimate the layer-wise uncertainty across all layers of the language model component in the MLLM, thereby serving as a proxy for the actual evolution of uncertainty across layers. To verify the reliability of this approximation, we conduct experiments following the same setup as the main experiments on the VQAv2-6K and ChartQA-8K datasets and compute the Pearson product-moment correlation

coefficient (Schober et al., 2018) over the unlabeled pool. Specifically, we compute the correlation between the average uncertainty across all language model layers and the uncertainty-dynamics score. As shown in Table 1, the correlation coefficients across all iterations are generally close to or above 0.7, which is classified as a strong correlation according to the criterion in Schober et al. (2018). These results support the effectiveness of the proposed uncertainty-dynamics score in approximating layer-wise uncertainty dynamics within the language model.

Table 1: Pearson product-moment correlation results across active learning iterations using the SmolVLM-256M model.

Datasets	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6
VQAv2-6K	0.70± 0.02	0.67± 0.01	0.68± 0.02	0.70± 0.02	0.70± 0.02	0.70± 0.01
ChartQA-8K	0.72± 0.06	0.73± 0.05	0.78± 0.05	0.73± 0.03	0.73± 0.06	0.67± 0.07

4.3.2 Selection Overlap Analysis

To further evaluate the effectiveness of the uncertainty-dynamics approximation (Eq. 4) in Alg. 1, we measure the overlap between the samples selected by our method using the approximate dynamics and those selected by the same method with Eq. 4 replaced by full layer-wise uncertainty dynamics, on ChartQA-8K using the SmolVLM-256M model. The results are shown in Table 2, where the selection overlap at each iteration is computed based on the cumulative selected samples up to that iteration. By the final iteration, our method and its full-dynamics variant exhibit notable overlap in their selected samples, suggesting that the proposed approximation closely captures the selection behavior of full layer-wise uncertainty dynamics.

Table 2: Selection overlap between our method and its full-dynamics variant on the ChartQA-8K dataset using the SmolVLM-256M model.

Datasets	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6
ChartQA-8K	0.66± 0.02	0.55± 0.02	0.56± 0.02	0.60± 0.01	0.63± 0.01	0.70± 0.01

4.3.3 Effectiveness of Full Layer-wise Uncertainty Dynamics

To assess the effectiveness of full layer-wise uncertainty dynamics, we conduct additional experiments on the ChartQA-8K dataset with the SmolVLM-256M model. Specifically, we replace the approximated uncertainty dynamics (Eq. 4) with the average of the actual uncertainty computed across all layers. We report the mean of accuracy and CIDEr in Table 3. The results show that using the full dynamics leads to improved performance, demonstrating that it provides a reliable criterion for data acquisition. Combined with the analysis in Table 1 and 2, these results suggest that our method effectively approximates a valid sample selection signal. Additional results, including ablation studies examining the insufficiency of final-layer uncertainty and other performance comparisons across the remaining main datasets, are deferred to Appendix D.

Table 3: Performance comparison between our method and its full-dynamics variant on the ChartQA-8K dataset using the SmolVLM-256M model.

Methods	Iter. 0	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6
ours	0.59± 0.03	0.62± 0.00	0.66± 0.01	0.68± 0.01	0.69± 0.00	0.71± 0.00	0.71 ± 0.01
ours with full dynamics	0.59± 0.01	0.62± 0.01	0.67± 0.01	0.68± 0.02	0.70± 0.00	0.71± 0.02	0.71 ± 0.01

4.3.4 Effect of B

We study the effect of varying hyperparameter B on model performance in this section. As mentioned earlier, this hyperparameter controls the relative contributions of uncertainty- and diversity-based selection in our method. Therefore, we conduct experiments with $B \in \{3, 4, 5\}$ on the VQAv2-6K

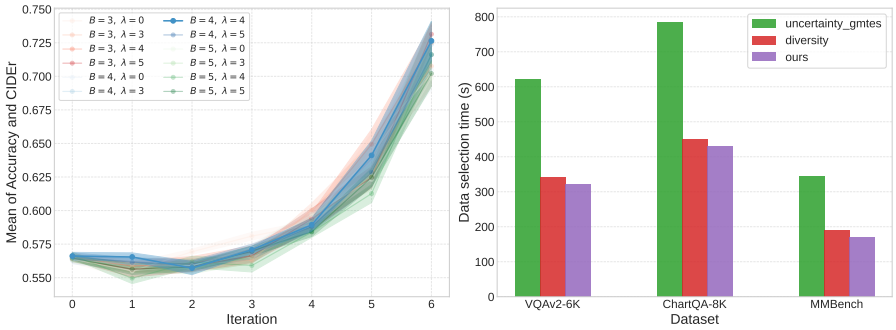


Figure 6: *Left:* Parameter study of Alg. 1 with different values of B and λ on the VQAv2-6K dataset using the SmolVLM-256M model. We report the average of accuracy and CIDEr as learning progresses. *Right:* Per-round data acquisition/selection time (seconds) across various datasets for different data acquisition strategies using the SmolVLM-256M model.

dataset and report the average of accuracy and CIDEr in Figure 6. Based on the results at the final iteration, model performance degrades under an inappropriate setting of B. For example, averaging across different values of λ, $B = 4$ achieves a performance of 0.718, compared with 0.716 for $B = 3$ and 0.711 for $B = 5$. This indicates that a well-chosen B can balance the uncertainty- and diversity-based selection criteria. Accordingly, we set $B = 4$ for this task.

4.3.5 Impact of λ

To evaluate the impact of λ in our Alg. 1, which determines the weight assigned to the uncertainty evolution term, we conduct experiments with $\lambda \in \{0, 3, 4, 5\}$ while keeping all other settings the same as those in Section 4.3.4. As reported in Figure 6, setting λ to 4 leads to better performance when $B = 4$. Specifically, the model achieves an average score of 0.73 with $\lambda = 4$, compared with 0.70 when $\lambda = 0$. The latter corresponds to a new baseline that considers only the uncertainty from the final layer. These results demonstrate the benefit of incorporating uncertainty evolution/dynamics into our method.

4.3.6 Data Acquisition Runtime Comparison

To assess whether our method overcomes the computational inefficiency limitation discussed in Section 2, we report the per-round data acquisition/selection time of each active learning strategy in Figure 6. As expected, our algorithm is substantially more efficient than the uncertainty_gmtes baseline, as it avoids the costly sequential generation process required by the latter. Compared with the diversity baseline, our method performs diversity-aware selection over a smaller candidate set, thereby reducing the associated computational cost. The dynamics-guided selection introduces extra computational overhead, but the computational savings from the reduced candidate set outweigh this overhead, resulting in lower overall acquisition times across datasets.

5 Conclusion

In this paper, we present the first study of active learning in *multimodal generation settings*. By leveraging uncertainty dynamics to quantify uncertainty and incorporating diversity into the selection process, we develop an efficient active learning algorithm applicable to both open- and closed-ended scenarios. Experiments across benchmark datasets show that our method achieves comparable performance with up to 33.33% fewer annotated samples, demonstrating the effectiveness of active learning in reducing annotation costs for multimodal generative learning.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results. We used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work, which consisted of polishing the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Gantavya Bhatt, Yifang Chen, Arnav Das, Jifan Zhang, Sang Truong, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Du, Kevin Jamieson, et al. An experimental design framework for label-efficient supervised finetuning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6549–6560, 2024.

Jinhe Bi, Zengjie Jin, Yifan Wang, Danqi Yan, Wenke Huang, Xiaowen Ma, Sikuan Yan, Artur Hecker, Mang Ye, Xun Xiao, et al. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024b.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *International conference on computational learning theory*, pp. 249–263. Springer, 2005.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Juan Elenter, Navid NaderiAlizadeh, and Alejandro Ribeiro. A lagrangian duality approach to active learning. *Advances in Neural Information Processing Systems*, 35:37575–37589, 2022.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pp. 181–195. Springer, 2022.

Sanket Rajan Gupte, Josiah Aklilu, Jeffrey J Nirschl, and Serena Yeung-Levy. Revisiting active learning in the era of vision foundation models. *arXiv preprint arXiv:2401.14555*, 2024.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein Dor. Active learning for natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9862–9877, 2023.

Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In *Conference on learning theory*, pp. 3806–3832. PMLR, 2021.

Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, and Jordan T Ash. Streaming active learning with deep neural networks. In *International Conference on Machine Learning*, pp. 30005–30021. PMLR, 2023.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles. Active learning literature survey. 2009.

Meng Shen, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See. Towards balanced active learning for multimodal classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3434–3445, 2023.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, 35:25171–25184, 2022.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, et al. From selection to generation: A survey of llm-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14552–14569, 2025.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. Hide and seek in noise labels: Noise-robust collaborative active learning with llms-powered assistance. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 10977–11011, 2024.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Jiancheng Zhang and Yinglun Zhu. Towards multimodal active learning: Efficient learning with limited paired data. *arXiv preprint arXiv:2510.03247*, 2025.

Jiancheng Zhang, Meiqing Li, Qi Zhang, and Yinglun Zhu. Active learning with foundation model priors: Efficient learning under class imbalance. *arXiv preprint arXiv:2606.07630*, 2026.

Jifan Zhang, Yifang Chen, Gregory Canal, Stephen Mussmann, Arnav M Das, Gantavya Bhatt, Yinglun Zhu, Jeffrey Bilmes, Simon Shaolei Du, Kevin Jamieson, et al. Labelbench: A comprehensive framework for benchmarking adaptive label-efficient learning. *arXiv preprint arXiv:2306.09910*, 2023.

Yiqian Zhang, Buyu Liu, Jun Bao, Qiang Huang, Min Zhang, and Jun Yu. Learnability matters: active learning for video captioning. *Advances in Neural Information Processing Systems*, 37: 37928–37954, 2024.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active learning approaches to enhancing neural machine translation. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 1796–1806, 2020.

Yinglun Zhu and Robert Nowak. Active learning with neural networks: Insights from nonparametric statistics. *Advances in Neural Information Processing Systems*, 35:142–155, 2022a.

Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. *Advances in Neural Information Processing Systems*, 35:35379–35391, 2022b.

A ALGORITHM

We present our complete algorithm in Alg. 1.

Algorithm 1 Multimodal Generative Active Learning

Input: Multimodal dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$, number of active learning rounds R, per-round selection size K, initial number of data points M, interval h, candidate size B, weight λ, multimodal large language model (MLLM) components (VE, Proj, Emb, $\text{LLM}^{(l)}$).

Initialize the labeled dataset $\mathcal{S}_0$ with M data points drawn uniformly at random from $\mathcal{D}$, together with their queried ground-truth responses.

for $r = 1, \cdots, R$ **do**

Consider unlabeled dataset $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{S}_{r-1}$.

Multimodal Feature Extraction

for all $d_i = (I_i, T_i) \in \mathcal{D}_r$ **do**

$\mathbf{v}_i \leftarrow \text{VE}(I_i)$,

$\mathbf{z}_i^v \leftarrow \text{Proj}(\mathbf{v}_i)$,

$\mathbf{z}_i^t \leftarrow \text{Emb}(T_i)$,

$\mathbf{X}_i \leftarrow [\mathbf{z}_i^v; \mathbf{z}_i^t]$,

$\mathbf{H}_i^{(q)} \leftarrow \text{LLM}^{(q)}(\mathbf{X}_i), \quad \forall q \in \{l, l-h\}$. `// Extract multimodal hidden states from multiple layers.`

Dynamics-Guided Selection

Let $\phi(\cdot) := \text{softmax}(\text{LMHead}(\text{Norm}(\cdot)))$ denote the token-to-probability mapping of the MLLM. For each data point d_i, compute the mean token entropy as its uncertainty score:

$$U_i^{(q)} := \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left(-\sum_{v=1}^{|\mathcal{V}|} \phi(\mathbf{H}_{i,t}^{(q)})_v \log \phi(\mathbf{H}_{i,t}^{(q)})_v \right).$$

Approximate the corresponding uncertainty dynamics based on two selected layers l and $l-h$:

$$u_i := U_i^{(l)}, \qquad u_i^{'} := -\frac{U_i^{(l)} - U_i^{(l-h)}}{h},$$

Construct the raw uncertainty-dynamics vector: $\mathbf{J}_i := [u_i, u_i'], \forall d_i \in \mathcal{D}_r$.

Normalize the uncertainty dynamics representation of each data point:

$$\bar{\mathbf{J}}_i[j] \leftarrow \frac{\mathbf{J}_i[j] - \min_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j]}{\max_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j] - \min_{d_k \in \mathcal{D}_r} \mathbf{J}_k[j] + \varepsilon}, \quad \forall d_i \in \mathcal{D}_r,\ j = 1, 2.$$

Compute an uncertainty-dynamics score for each data point: $s_i := \bar{\mathbf{J}}_{i,1} + \lambda \cdot \bar{\mathbf{J}}_{i,2}$.

Select the top-B data points with the largest uncertainty-dynamics scores to form the candidate subset $\mathcal{P}_r$.

Diversity-Aware Selection

Construct a coreset $\mathcal{C}_r \subseteq \mathcal{P}_r$ of size K:

$$\mathcal{C}_r := \underset{\mathcal{C}:|\mathcal{C}|=K, \mathcal{C} \subseteq \mathcal{P}_r}{\arg\min} \max_{d_i \in (\mathcal{P}_r \setminus \mathcal{C})} \min_{d_j \in (\mathcal{S}_{r-1} \cup \mathcal{C})} \text{dist} \left(\frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{H}_{i,t}^{(l)}, \frac{1}{|\mathcal{T}_j|} \sum_{t \in \mathcal{T}_j} \mathbf{H}_{j,t}^{(l)} \right).$$

Annotate the selected coreset, and update $\mathcal{S}_r \leftarrow \mathcal{S}_{r-1} \cup \{(I_i, T_i, Y_i)\}_{i=1}^{K}$.

Model update. Train multimodal large language model f_r on the updated annotation set $\mathcal{S}_r$.

Output: Actively trained model.

B RELATED WORK

Active learning. Active learning aims to achieve strong model performance with fewer annotations by selectively acquiring informative samples for labeling (Settles, 2009). Prior theoretical work has demonstrated provable benefits of active learning over passive learning across a variety of learning

settings (Castro & Nowak, 2008; Dasgupta et al., 2005; Hanneke, 2014; Zhang & Chaudhuri, 2014; Puchkin & Zhivotovskiy, 2021; Zhu & Nowak, 2022a,b). Empirical studies have also demonstrated the effectiveness of active learning across diverse deep learning applications (Tong & Koller, 2001; Beluch et al., 2018; Citovsky et al., 2021; Wang et al., 2022; Zhang et al., 2023), and recent work has further extended active learning to foundation models (Bhatt et al., 2024; Yuan et al., 2024; Zhang et al., 2026). Existing active learning algorithms generally fall into three classes: uncertainty-based approaches that target samples with high predictive uncertainty (Gal et al., 2017; Ducoffe & Precioso, 2018), diversity-based methods that emphasize representative coverage of the data space (Sener & Savarese, 2017; Citovsky et al., 2021), and hybrid strategies that incorporate both uncertainty and diversity into the acquisition process (Ash et al., 2019; Elenter et al., 2022; Zhang & Zhu, 2025).

Multimodal generative learning. Multimodal generative learning seeks to integrate information from diverse modalities, such as text, images, and audio, to enhance generation performance and produce outputs in a target modality conditioned on multimodal inputs. Early methods combined pretrained vision and language models for conditional text generation. For instance, Alayrac et al. (2022) introduced Flamingo with 80B parameters, which generates text from interleaved visual and textual inputs, achieving strong few-shot performance. Laurençon et al. (2023) adopted Flamingo's architecture, exemplifying large-scale multimodal learning. BLIP-2 (Li et al., 2023) achieves better performance than prior methods while using substantially fewer trainable parameters. Similarly, LLaVA (Liu et al., 2023) connected a pretrained image encoder to a LLaMA backbone (Llama, 2023), achieving strong visual conversational performance while being smaller and faster than Flamingo. More recent work has pursued greater efficiency along two main directions. Models such as Qwen-VL (Wang et al., 2024; Bai et al., 2025) and InternVL (Chen et al., 2024b,a) have focused on improving computational efficiency through architectural innovations without sacrificing multimodal capability. In parallel, compact multimodal models such as MiniCPM-V (Hu et al., 2024) and SmolVLM (Marafioti et al., 2025) reduce model size while maintaining competitive multimodal performance.

Multimodal active learning. Most existing multimodal active learning methods primarily focus on classification and cross-modal alignment tasks. Shen et al. (2023) studies multimodal active learning for classification in a pre-aligned setting, where the correspondence between modalities is assumed to be given. They therefore avoid the need to infer cross-modal correspondences during active learning. Zhang et al. (2024) takes a similar perspective on the multimodal active learning setting. More recently, Zhang & Zhu (2025) has proposed an efficient algorithm for multimodal active learning in the unaligned setting, explicitly targeting cross-modal alignment. Existing multimodal active learning methods assume that the learner either predicts discrete labels or identifies cross-modal correspondences. These assumptions fail to naturally extend to the multimodal generation setting. To address this gap, we propose the first active learning algorithm for the multimodal generation setting.

C Additional Implementation Details

Dataset construction and splits. For the full MMBench dataset, we randomly partition the official validation split into 75% for training and 25% for testing, since ground-truth annotations are only available for the validation split. For the full ChartQA dataset, we use the official training and test splits. To construct ChartQA-8K, we randomly sample 30% of the official training split, resulting in a training set of approximately 8K examples, while retaining the official test split for evaluation. For VQAv2-6K, since ground-truth annotations are only available for the publicly released validation split, we randomly sample 40% of this split to construct the subset and further partition it into 75% for training and 25% for testing. We list detailed descriptions of these datasets in Table 4.

Additional evaluation details. On the VQAv2-6K dataset, we report standard VQA accuracy between predicted and ground-truth answers. For CIDEr, we compute the score between the ground-truth answers and both the extracted answers and the full generated responses, and report the average of the two scores. For the other datasets, the generated responses and extracted answers are very similar in both content and length, making separate evaluation redundant. In contrast, the generated responses on VQAv2-6K are often longer than the extracted answers. We therefore evaluate both to more comprehensively assess the performance of active learning methods.

Table 4: Datasets used in our experiments.

Datasets	# Samples in Splits	
	Train	Test
VQAv2-6K	6,430	2,144
ChartQA-8K	8,489	2,500
MMBench	3,282	1,095
ChartQA	28,299	2,500

Hyperparameter settings. Table 5 lists the hyperparameters used in our experiments across datasets and model variants.

Table 5: Hyperparameter settings for different datasets and model variants.

	VQAv2-6K	ChartQA-8K	MMBench	ChartQA	VQAv2-6K
Epochs	1	1	1	1	1
Batch Size	8	8	8	8	8
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning Rate	1×10^{-5}	4×10^{-5}	1×10^{-5}	4×10^{-5}	5×10^{-6}
Per-round Budget K	200	250	110	808	200
Model	SmolVLM-256M	SmolVLM-256M	SmolVLM-256M	SmolVLM-256M	SmolVLM-2.2B
B	4	26	16	26	2
λ	4	3.5	0.3	3	2
h	28	28	28	23	23

D Other Analyses and Ablations

D.1 Insufficiency of Final-Layer Uncertainty

To evaluate the contribution of Eq. 2 to our algorithm, we conduct ablation studies by removing this component and using only the final-layer uncertainty to select the candidate set. We first measure the Pearson correlation between the final-layer uncertainty and the average uncertainty across all layers, with the results reported in Table 6. The results indicate moderate and weak correlations across the two datasets, respectively, suggesting that the final-layer uncertainty provides a less accurate approximation than the proposed uncertainty-dynamics score. These findings are also reflected in Figure 6, where using only the final-layer uncertainty yields a mean accuracy and CIDEr score of 0.70.

Table 6: Pearson product-moment correlation between the average uncertainty across all layers and the uncertainty at the final layer across active learning iterations using the SmolVLM-256M model.

Datasets	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6
VQAv2-6K	0.53± 0.01	0.49± 0.01	0.49± 0.01	0.51± 0.01	0.54± 0.01	0.57± 0.00
ChartQA-8K	0.04± 0.06	0.14± 0.12	0.25± 0.04	0.31± 0.11	0.14± 0.13	0.34± 0.06

D.2 Additional Evaluation of Full Layer-Wise Uncertainty Dynamics

We conduct the experiments on the VQAv2-6K and MMBench datasets following the same setup as in Section 4.3.3. As shown in Table 7, using the full layer-wise uncertainty dynamics achieves slightly better performance, further supporting the conclusion in the main text that our method approximates the good sample selection signal provided by the full layer-wise uncertainty dynamics.

D.3 Robustness on Larger Dataset

To evaluate the robustness of Alg. 1 across different data scales, we conduct additional experiments on the full ChartQA dataset, which contains nearly 30,000 samples, using the SmolVLM-256M

Table 7: Performance comparison between our method and its full-dynamics variant on the VQAv2-6K and MMBench datasets using the SmolVLM-256M model.

Dataset	Methods	Iter. 0	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6
VQAv2-6K	Ours	0.57± 0.01	0.57± 0.01	0.56± 0.01	0.57± 0.01	0.59± 0.01	0.64± 0.03	0.73± 0.04
	Ours with full dynamics	0.57± 0.01	0.56± 0.01	0.56± 0.01	0.57± 0.01	0.59± 0.02	0.65± 0.01	0.73± 0.02
MMBench	Ours	0.69± 0.02	0.70± 0.02	0.71± 0.01	0.73± 0.03	0.74± 0.02	0.75± 0.02	0.76± 0.02
	Ours with full dynamics	0.69± 0.02	0.70± 0.02	0.72± 0.02	0.73± 0.02	0.74± 0.02	0.75± 0.02	0.76± 0.01

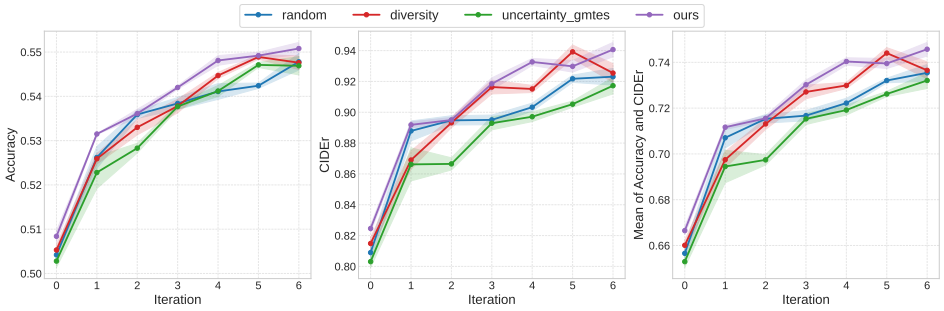


Figure 7: Performance comparison of our proposed method and baseline methods on the full ChartQA dataset using the SmolVLM-256M model.

model. As shown in Figure 7, our method generally outperforms the baselines in terms of the average of accuracy and CIDEr scores. As expected, performance improves for all methods as the per-round labeling budget increases, while the advantage of our method is maintained, further validating its robustness at a larger data scale.

D.4 Generalization across Model Variants

To assess the sensitivity of our Alg. 1 to the choice of model, we conduct additional experiments on the VQAv2-6K dataset using the SmolVLM-2.2B model. As shown in Figure 8, our method generally achieves higher accuracy while maintaining a comparable CIDEr score to the baselines. Similarly, all methods benefit from scaling to the larger model, resulting in improved performance. Our method also maintains its relative advantage, demonstrating consistent performance across different model variants.

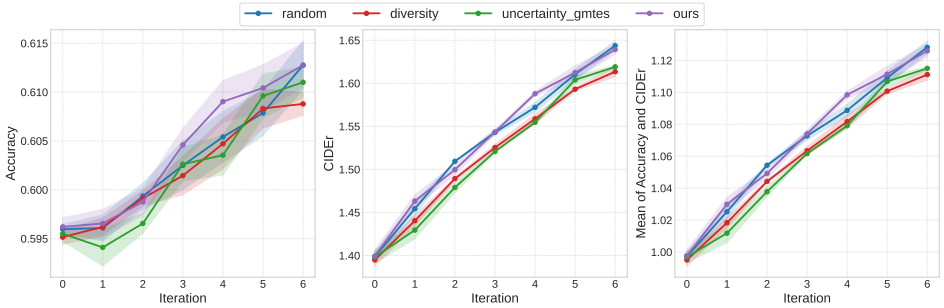


Figure 8: Performance comparison of our proposed method and baseline methods on the VQAv2-6K dataset using the SmolVLM-2.2B model.