UNDER REVIEW

Research Study: CS 421

ImplicitAVE Reproducibility Study

University of Illinois Chicago

Reproducibility Summary

This study focuses on reproducing the primary claims of the original paper *ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction*. Specifically, it aims to validate:

- The **ImplicitAVE** dataset's comprehensiveness in addressing implicit attribute value extraction (AVE) tasks across diverse domains and attributes.
- The inherent challenges of implicit AVE tasks for state-of-the-art multimodal large language models (MLLMs) in zero-shot settings, particularly for nuanced attributes and domains.

Methodology

The code and dataset provided in the **ImplicitAVE** GitHub repository were utilized to evaluate the Qwen-VL model. The approach involved reproducing both domain-level and attribute-level evaluations, leveraging NVIDIA A100 GPUs for computations. Despite challenges, such as incomplete dependencies and the need for multiple environment adjustments, the results for Qwen-VL were successfully reproduced and extended.

Results

The reproduced results are consistent with the original paper's claims, demonstrating comparable performance across domains and attributes. At the domain level, Qwen-VL achieved an average micro-F1 of 70.36%, closely matching the original 70.86%. Attribute-level evaluations revealed robust performance for reported attributes and provided new insights for attributes not previously analyzed, such as *Shape in Home* (*85.00%*) and *Occasion in Food* (*38.09%*).

What was Easy

The objective of the paper and the structure of the **ImplicitAVE** dataset were clearly articulated, facilitating straightforward understanding and usage. The instructions for running Qwen-VL were well-defined, enabling a smooth setup and evaluation process.

What was Difficult

Reproducing the experiments presented challenges in setting up the computational environment to match the original GPU specifications and addressing missing or outdated requirements for several models. The process required extensive adjustments and troubleshooting to establish a functional environment, which increased the time required. Several models referenced in the paper could not be reproduced due to incomplete documentation and dependencies.

Communication with Original Authors

An issue was raised on the GitHub repository regarding the missing and outdated requirements file. As of this report, no response has been received from the original authors.

1 Introduction

The paper *ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction* [1] presented in the ACL 2024 (finding papers) [2], introduces **ImplicitAVE**, a dataset addressing limitations in existing attribute value extraction (AVE) benchmarks by focusing on implicit attribute values and incorporating multimodality. This dataset combines product text and images, enabling the evaluation of multimodal large language models (MLLMs) in extracting implicit attributes. The study benchmarks six state-of-the-art MLLMs across experimental settings, including zero-shot and few-shot learning, highlighting the challenges of implicit value extraction. The objective is to reproduce the experimental results, validate the claims, and assess the feasibility of the proposed methods using the provided dataset and benchmarks.

2 Scope of reproducibility

The work addresses the problem of implicit attribute value extraction (AVE), a critical task in e-commerce for identifying attribute values that are not explicitly mentioned in product descriptions but can be inferred from contextual clues or product images. The study introduces **ImplicitAVE**, a multimodal dataset designed specifically for implicit AVE. It benchmarks state-of-the-art multimodal large language models (MLLMs) to assess their performance in this challenging task.
Claims to be tested:

- **Claim 1:** The **ImplicitAVE** dataset provides a comprehensive resource for implicit AVE tasks, incorporating multimodal information and high-quality annotations across diverse domains and attributes, addressing the limitations of existing datasets.
- **Claim 2:** Implicit attribute value extraction is a challenging task for current MLLMs. State-of-the-art models struggle to achieve high performance in zero-shot and few-shot settings, particularly on nuanced attributes and domains.

3 Methodology

The methodology primarily involves authors providing code and data from the **ImplicitAVE** GitHub repository to reproduce their experiments. Necessary modifications were applied to adapt the setup for evaluating the Qwen_VL model.

3.1 Model descriptions

The experiments utilized the QwenV2, a state-of-the-art multimodal large language model (MLLM). Qwen_VL's unique architecture integrates visual and textual inputs through advanced cross-modal training. It is parameterized with:

- **Base Model:** Qwen-7B
- **Parameters:** Approximately 7 billion
- **Special Features:** Includes a visual receptor and position-aware adapters optimized for multimodal tasks.

3.2 Datasets

The **ImplicitAVE** dataset, as described in the paper, was used. Details include:

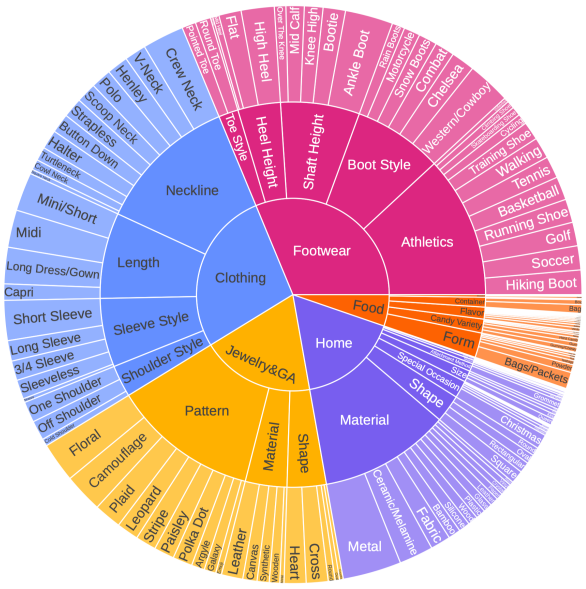

Figure 1. Overview of the train set composition.

- **Statistics:** The dataset comprises 68,000 training examples as shown in Figure 1 and 1,600 evaluation examples visible in Figure 2 across five domains: Clothing, Footwear, Jewelry & GA, Food, and Home.
- **Preprocessing:** Explicit attribute values were removed to create implicit tasks and redundant attributes were filtered. Dataset preprocessing scripts from the repository were used without modifications.

Below, we include two images from the paper illustrating the train set and evaluation set compositions, providing a clear overview of the structure and examples.

3.3 Experimental setup and code

The experimental setup was adapted from the authors' repository:

- **Evaluation Metric:** Micro-F1 score at domain and attribute levels was the primary performance measure.
- **Code Availability:** The code, along with adaptations made to reproduce the results, is publicly available [3].

3.4 Computational requirements

The experiments required significant computational resources:

- **Hardware:** NVIDIA A100 GPU with 40GB RAM.
- **Average Runtime:** 50 minutes/run approximately.
- **Total Colab Computational Units:** Almost 100 computational units.

4 Results

All the numbers reported in our results are the averages of two runs to ensure greater consistency and reliability. Our results provide evidence supporting the main claims

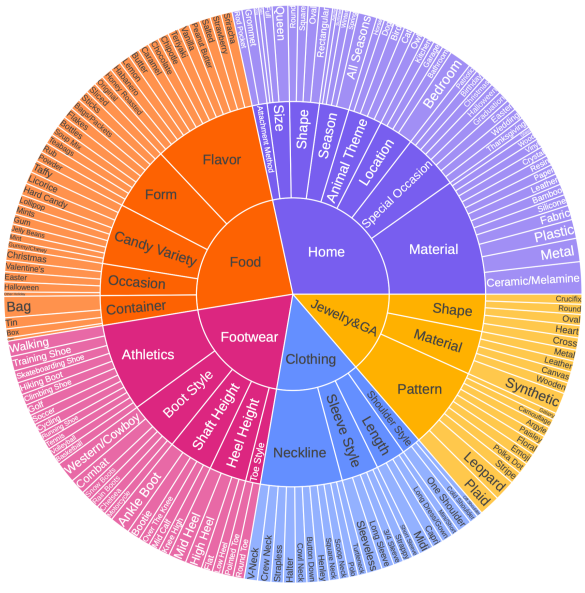

Figure 2. Overview of the evaluation set composition.

of the original paper. A summary of the reproduced results and comparison with the original paper's results is shown in Table 1, which details the domain-level performance. Additionally, Table 2 provides a fine-grained attribute-level comparison for Qwen-VL.

Table 1. Comparison of Qwen-VL original results and reproducibility results across categories (Domain-level comparison)

Method	Clothing	Footwear	Jewelry & GA	Food	Home Product	All
Qwen-VL (Original)	59.73	57.72	84.09	76.92	73.96	70.86
Qwen-VL (Reproducibility)	58.67	58.54	85.39	78.41	72.81	70.36

According to the original paper [1], only the most influential prompt is used for model interaction. As shown in Table 3, *prompt 8* is identified as the most influential in the original study. This finding is consistent with our reproducibility study, where *prompt 8* also proved to be the most influential, demonstrating coherence between the original results and our replication efforts.

4.1 Results reproducing original paper

The reproduced results confirm the main claims of the original paper. Each experiment is analyzed with respect to the associated claims:

Claim 1 — *The **ImplicitAVE** dataset provides a comprehensive resource for implicit AVE tasks, incorporating multimodal information and high-quality annotations across diverse domains and attributes, addressing the limitations of existing datasets.*
The results affirm this claim. Using the **ImplicitAVE** dataset, the evaluation showed consistent performance across a wide range of domains and attributes. Domain-level results shown in Table 1 reflect strong performance in categories such as **Jewelry & GA** (*85.39%*) and **Food** (*78.41%*), while attribute-level performance visible in Table 2 demonstrates robust results in attributes like **Flavor** (*89.93%*) and **Container in Food** (*87.50%*). These findings emphasize the dataset's capability to enable thorough evaluations of implicit attribute value extraction tasks.

Table 2. Comparison of Qwen_VL Results: Paper vs. Reproducibility Study

Domains	Attributes	# Values	Qwen_VL (Paper %)	Qwen_VL (Reproducibility %)
Food	Flavor	14	89.21	89.93
	Container	4	80.00	87.50
	Form	9	75.58	75.58
	Occasion	5	-	38.09
Home	Material	13	67.09	65.61
	Shape	4	-	85.00
	Special Occasion	8	88.15	89.47
	Attachment Method	2	100.00	85.00
Jewelry & GA	Pattern	10	89.19	88.28
	Material	5	88.14	86.21
Footwear	Boot Style	6	72.05	73.13
	Heel Height	4	54.00	52.00
	Shaft Height	5	-	35.00
	Toe Style	2	-	75.00
Clothing	Sleeve Style	5	66.00	60.00
	Shoulder Style	3	80.77	84.62
	Length	4	-	60.00
	Neckline	11	52.73	50.46

Table 3. Prompt's influence comparison on Qwen_VL

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8
Paper	0.45	0.60	0.59	0.50	0.46	0.61	0.56	0.62
Reproducibility	0.48	0.57	0.59	0.50	0.47	0.59	0.54	0.61

Claim 2 — *Implicit attribute value extraction is a challenging task for current MLLMs. State-of-the-art models struggle to achieve high performance in zero-shot and few-shot settings, particularly on nuanced attributes and domains.*
The results support this claim by exposing the limitations of MLLMs in nuanced scenarios. For example, attribute-level evaluations reveal difficulties with attributes like **Heel Height** (*52.00%*) and **Occasion in Food** (*38.09%*), indicating that even state-of-the-art models such as Qwen_VL encounter challenges in zero-shot settings. This variability highlights the inherent complexity of implicit AVE tasks.

4.2 Results beyond original paper

In addition to reproducing the original results, we evaluated attributes not reported in the original paper, providing further insights into Qwen-VL's performance:

- **Unreported Attributes:** We analyzed attributes such as **Occasion in Food** (*38.09%*), **Shape in Home** (*85.00%*), **Shaft Height in Footwear** (*35.00%*), **Toe Style in Footwear** (*75.00%*), and **Length in Clothing** (*60.00%*). These evaluations highlight the model's strengths and weaknesses in previously unexplored areas.
- **Challenges in Nuanced Attributes:** Consistent with the original study, attributes like **Heel Height in Footwear** (*52.00%*) and **Material in Home** (*65.61%*) remain challenging, reflecting the limitations of current MLLMs.
- **Domain Variability:** Within domains, the model demonstrated significant variability. For example, in the **Home** domain, strong performance was observed for **Attachment Method** (*85.00%*), while **Material** (*65.61%*) posed more difficulty.

This extended analysis provides a more complete picture of Qwen-VL's strengths and limitations in implicit attribute value extraction.

5 Discussion

Reproducing the results of the original paper presented a mixed experience. While the paper clearly articulated its objectives and provided a comprehensive dataset, the lack of adequate documentation and missing requirements for many models introduced significant challenges. Specifically, the incomplete dependencies and setup instructions hindered the reproducibility of several models referenced in the study, limiting the broader scope of our evaluation. Despite these difficulties, the Qwen-VL model was successfully evaluated, with results aligning with the original claims. This experience underscores the importance of robust documentation and detailed requirements in ensuring the reproducibility of machine learning research across diverse computational environments.

5.1 What was easy

The paper's objective and the gap it aimed to fill are clearly defined and well-articulated. The **ImplicitAVE** dataset, along with its accompanying descriptions, is structured to facilitate understanding and usage. The instructions for running Qwen-VL are straightforward, enabling the setup of the initial environment for this model with relative ease.

5.2 What was difficult

Reproducing the experiments, despite the clarity of the paper, presented several challenges:

- **Environment Setup:** Configuring the appropriate environment to run the models required significant time. Local execution without GPU support was prohibitively slow. Transitioning to cloud-based environments introduced additional challenges, such as matching GPU specifications with those used in the original paper and ensuring flexibility for managing repository files.
- **Requirements Issues:** The repository did not include an updated and complete requirements file for many models. Numerous dependencies were outdated or missing, necessitating manual adjustments. Consequently, several models referenced in the original paper could not be reproduced due to nonfunctional or poorly documented setups.
- **Iterative testing of configurations:** Testing various platforms and configurations was required to identify a functional setup for running the Qwen-VL model. This process was iterative and demanded substantial time and computational resources.

5.3 Communication with original authors

An issue was raised on the repository's GitHub page regarding the missing and outdated requirements file. The original authors responded with detailed guidance on proceeding with the setup of other modules. Future work will focus on incorporating these additional modules to validate and extend the findings of this study.

References

1. H. P. Zou, V. Samuel, Y. Zhou, W. Zhang, L. Fang, Z. Song, P. S. Yu, and C. Caragea. "ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction." In: **arXiv preprint arXiv:2404.15592** (2024).
2. L.-W. Ku, A. Martins, and V. Srikumar, eds. **Findings of the Association for Computational Linguistics: ACL 2024**. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024.
3. A. A. Marina and M. Laurenzi. **ImplicitAVE Reproducibility Study**. https://github.com/marcolaurenzi/ImplicitAVE-Reproducibility-Study. 2024.