AutoVDC: Automated Vision Data Cleaning Using Vision-Language Models

Aditi Ramadwar*[1], Santosh Vasa*[1], Jnana Rama Krishna Darabattula[1],
Md Zafar Anwar[1], Stanislaw Antol[1], Andrei Vatavu[1], Thomas Monninger[1,2], Sihao Ding[1]

***Abstract*— Training robust autonomous driving systems requires extensive datasets with precise annotations, yet manual data curation is expensive and human labels are often imperfect. In this paper, we conduct a systematic study of Vision-Language Model (VLM) maturity for automated data curation. To derive our findings, we utilize AutoVDC (Automated Vision Data Cleaning), a model-agnostic, two-stage framework that leverages the generalization gap of neural networks to audit large-scale datasets. In the pipeline, a task model first proposes suspicious candidates, which a VLM then acts as a judge to validate errors. Using this framework, we evaluate the readiness of VLMs for annotation cleaning on KITTI and nuImages. Our study reveals that zero-shot VLMs are effective on semantic errors but show a clear spatial reasoning gap for localization noise, making Fine-Tuning with Chain-of-Thought (FT-CoT) essential. We further demonstrate that across variants with intentionally injected erroneous annotations, this VLM-based auditing approach remains robust even when guided by a weak proposer and successfully uncovers 39 previously unknown errors in the original KITTI ground truth.**

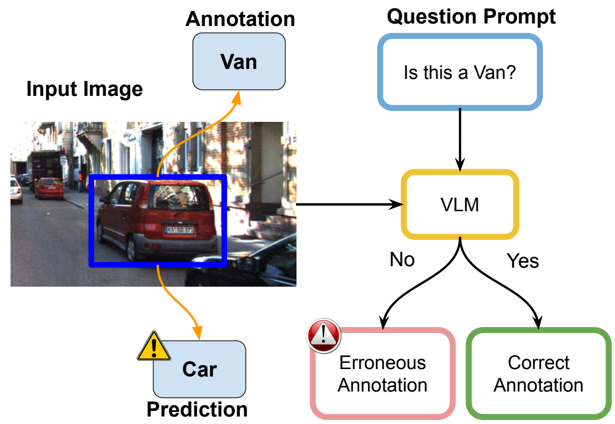

Fig. 1: The two-stage auditing process where the system identifies a label-prediction conflict (Error Proposal) and utilizes a VLM to validate the annotation's correctness (Error Validation).

I. Introduction

In machine learning, high-quality and efficient data annotation is crucial for developing effective models. Traditional manual annotation presents significant challenges, including a large workload, inconsistent quality, and a high cost. These issues limit the scalability and quality of annotations, adversely affecting model performance and evaluation. While auto-labeling algorithms have improved data collection efficiency by automating parts of the annotation process [1], they often introduce systemic biases that are problematic in high-stakes domains such as autonomous driving, where continuous refinement and validation are required.

These challenges are amplified by the rapid growth of dataset scale. Public dataset sizes in autonomous driving have expanded substantially, as seen in the Waymo Open Dataset [2] and Argoverse [3]. Industry and proprietary datasets are growing at an even faster rate, driven by the need for more comprehensive and diverse data to train robust models. This rapid growth makes label-by-label manual inspection increasingly costly and difficult to sustain, motivating the development of automated approaches that can scale.

In response to the above limitations, we conduct a systematic readiness study to evaluate the maturity of Vision-Language Models (VLMs) for automated data curation. To derive our findings, we utilize **AutoVDC** (**Auto**mated **V**ision **D**ata **C**leaning), which we define as a two-stage label cleaning method, Fig. 1, that integrates VLMs to automate the detection of annotation errors. The first stage acts as a triage stage that flags potential error candidates from the entire dataset. A VLM in the second stage assesses these proposals to determine whether or not the annotations are correct, which eliminates or substantially reduces the need for human intervention and effort.

By using AutoVDC as a test framework, we address a broader question: How mature are VLMs for high-precision perception auditing? While VLMs demonstrate strong open-ended understanding [4], perception datasets demand fine-grained spatial inspection in addition to semantic correctness. This motivates our study, focused on both the strengths and limitations of current VLMs when used as automated dataset auditors.

*Authors denoted with * contributed equally to this paper.
[1]Mercedes-Benz Research & Development North America, San Jose, CA, USA
[2]University of Stuttgart, Institute for Artificial Intelligence, Stuttgart, Germany

To evaluate the effectiveness of VLM-based curation, this paper addresses the following research questions (RQ):

RQ1. VLM Maturity and the Spatial Reasoning Gap. To what extent do current VLMs exhibit the architectural maturity required to resolve fine-grained localization noise compared to semantic classification errors, and is domain alignment through fine-tuning essential to bridge the remaining spatial reasoning gap?

RQ2. Exploiting the Generalization Gap. Does the tendency of task models to learn general semantic features before memorizing noise provide a reliable signal for automated error discovery, even when the task model (model specific to a task) is trained on noisy labels?

RQ3. The Resilience of Triage Logic. Can the VLM act as a robust gatekeeper to maintain high system precision even when utilizing a weak task model with high false-positive rates in the triage stage?

Our contributions are directed towards answering these questions and showcasing the analysis of our findings.

- A VLM readiness study: A systematic evaluation on KITTI [5] and nuImages [6] showing that while zero-shot VLMs work well for semantic errors, they exhibit a spatial reasoning gap; we show that FT-CoT is essential to resolve fine-grained localization errors.
- The AutoVDC framework: A modular, two-stage methodology that combines task-model triage with VLM-based validation for scalable, high fidelity annotation cleaning.
- Robustness and Real-World Impact: We demonstrate that VLM-based auditing is resilient to weak proposers, and successfully uncovers 39 previously unknown errors in the original KITTI ground truth, improving evaluation fidelity.

II. Related Work

To address the challenge of producing accurately labeled datasets, numerous data-driven methodologies have been introduced. For instance, anomaly detection is employed in label cleaning to identify data points that deviate significantly from the norm, thereby flagging samples that are likely to contain labeling inaccuracies [7, 8]. Moreover, several advanced methods integrate multiple strategies to improve the identification of erroneous annotations. Standard error detection utilizes model ensembling, entropy-based measures, and active label cleaning [9-12] to identify prediction-annotation divergences. Although these methods effectively highlight potential label inaccuracies, they still require human review to confirm the flagged samples, making the process labor-intensive. Studies like [13] and [14] have examined the challenges, costs, and efforts associated with these existing approaches, which we aim to obviate.

In recent years, generative models like Large Language Models (LLMs) [15] have increasingly replaced components of machine learning development in both academia and industry. LLMs are now used to evaluate models [16] and generate synthetic data and annotations [17], significantly reducing the need for human labor. More recently, Large Vision Models (LVMs), such as Vision Foundation Models (VFMs) [18-20] and VLMs [21-23], have markedly improved the ability of machines to perceive and understand visual information. These models have demonstrated impressive generalist capabilities [21], enabling them to perform a wide range of tasks across various domains in both vision and multimodal contexts [19, 22]. This versatility is attributed to their training on large, extensive and diverse vision and language datasets [24] that are significantly larger than task-specific model datasets [6]. This allows VLMs to understand and process relationships across modalities, thereby enhancing their performance in tasks that require both textual and visual information, thus ideal for VQA applications. These models can also reason image markups, such as boxes or arrows overlaid onto the image, that can be referenced in the text input [25].

Recent work is starting to leverage these abilities for active learning and label cleaning. [26] utilized VLMs to generate pseudo-labels in an incremental learning approach to train object detection models. [27] employed VLMs to validate pseudo-labels specialized for geo-locating voltage cabins in street view imagery in a semi-supervised setting. ClipGrader [28] fine-tuned CLIP to identify labeling errors in object detection datasets by determining similarity between textual and visual features. However, ClipGrader [28] is specialized for object detection and the use of the CLIP model, requiring substantial modifications to integrate large VLMs. In contrast, AutoVDC is modular, capable of integrating various VLMs, and can be extended to vision tasks other than object detection with minimal modifications. Additionally, with triage stage, AutoVDC framework drastically reduces the amount of annotations processed by a VLM, thereby aiding computational efficiency.

Nevertheless, utilization of VLMs in the ML development cycle, particularly for dataset cleaning, is an emerging area of study. In this work, we present a modular system that can leverage the wide range of VLMs to automatically identify erroneous annotations.

III. Approach

A. Problem Statement

The goal of automated data cleaning is to identify a subset of erroneous annotations D' within an imperfect dataset of N data points, $D = \{(a_i, x_i)\}_{i=1}^{N}$,

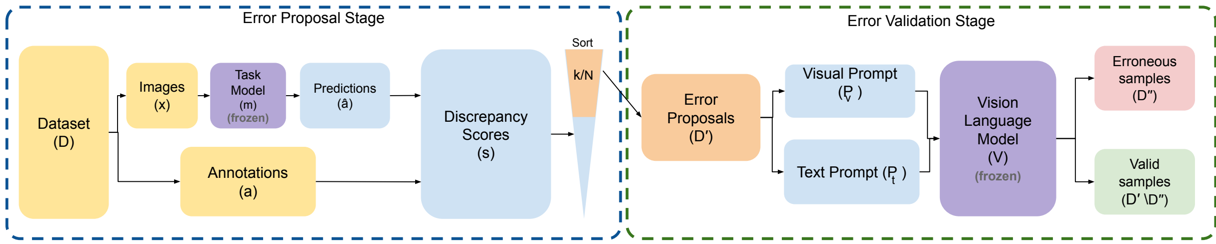

Fig. 2: Illustrates overall system architecture. *Left*: Error Proposal Stage scores samples based on model predictions and annotations. *Right*: Error Validation Stage employs a VLM to examine proposals and detect annotation errors.

where x_i represents an input image and a_i is the corresponding singular annotation (e.g., a bounding box or class label). We define a classification function $f(a_i, x_i) \rightarrow \quad y_i \in \{\text{yes}, \text{no}\}$ that maps each sample to its true correctness.

Our implementation, visualized in Fig. 2, approximates f by first exploiting the generalization gap of task models to propose discrepancies and then utilizing the spatial reasoning of VLMs to audit these proposals. This dual-intelligence approach allows the system to distinguish between inaccurate task model predictions and genuine annotation decay, providing a scalable alternative to exhaustive manual review.

B. Error Proposal (EP) Stage

The EP stage serves as a high-recall filter designed for computational efficiency. Its primary goal is to reduce the volume of data passed to the VLM by identifying a candidate subset that likely contains errors.

We utilize a perception task model m_θ to generate predictions $\hat{a}_i$. Even if m_θ is trained on noisy data, it typically learns consistent semantic features before memorizing random noise. We exploit this generalization gap by calculating a discrepancy score $s_i = \delta(\hat{a}_i, a_i)$, where δ is a task-specific function representing the degree of discrepancy (e.g., IoU, cosine similarity).

We then select the top-k samples or those exceeding a score threshold to form a set of error proposals $D\prime$. This process addresses RQ2 by concentrating the system's attention on suspicious samples, reducing the VLM's total workload significantly.

C. Error Validation (EV) Stage

The EV stage acts as the final judge, using the reasoning power of a VLM, V, to determine the true state of each proposal $(a_i, \hat{a}_i, x_i) \in D'$ by producing $\hat{y}_i$. The VLM is used as an expert auditor. We provide the VLM with two inputs, as seen in Fig. 1:

1) Visual Prompt P^v: A crop of x_i, centered on the discrepancy, overlaid with the bounding box a_i or $\hat{a}_i$ to focus the model's spatial attention.

2) Text Prompt P^t: A specific query requiring the VLM to reason whether the annotation a_i accurately describes the visual content of x_i.

The VLM generates a response, $\hat{y}_i = V(P_i^v, P_i^t)$, that determines the correctness of the annotation. From D', a new subset D'' is created consisting of validated erroneous annotations. This verified set can be used to eliminate (Section IV) or correct the flagged labels.

D. Application to 2D Object Detection

To demonstrate the framework's effectiveness, we instantiate AutoVDC for the task of 2D object detection.

EP Stage: Discrepancies are identified by matching predictions to ground-truth annotations based on an Intersection over Union (IoU) threshold of 0.65. Adjusting this threshold enables finer or coarser selection of samples for processing in the EV stage. Objects that fail to meet the matching criterion are marked as discrepancies with a score of $s_i = 1$. These cases fall into three categories:

1) **Class mismatch**: The model and annotation agree on the location but the object class categories differ.
2) **Missing prediction**: An annotation exists without a matching prediction. A potential false negative in task evaluation.
3) **Missing annotation**: The model predicts a high-confidence object where no annotation exists. A potential false positive in task evaluation.

For simplicity, we assign $s_i = 1$ to all such cases, treating them as equally discrepant, and include them in the set D'.

EV Stage: Our visual prompt P^v is created by overlaying the bounding box of interest on the image x_i and then cropping the image around the box of interest with fixed padding. Our textual prompt P^t uses discrepancy information to substitute in for placeholders to produce the final prompts, as shown in the right side of Fig 1.

The VLM then acts as an expert auditor, reasoning whether the box accurately fits the visual evidence. This allows the system to distinguish between a genuine annotation error and a simple model failure, addressing the spatial reasoning gap explored in our research questions.

IV. Experiments

We outline the experimental set up designed to answer the posed research questions for our study. We quantify

Fig. 3: The examples demonstrate different noise types and the performance of fine-tuned with CoT VLM Llama on each noise type and valid annotations. The columns represent annotations, predictions, and Q&A respectively. The green box is the visual prompt P^v.

performance throughout the study using standard metrics: Accuracy (Acc.), Recall (Rec.), Precision (Prec.), and the F1 Score (F1). Our study utilizes recall to measure error discovery and precision to ensure data integrity is maintained.

A. Datasets and Noise Profiles

We evaluate the framework on two standard 2D object detection benchmarks: KITTI (D_{K}) [5] and nuImages (D_{nu}) [6]. Since these datasets are relatively clean and, we assume, lack substantial erroneous annotations, we introduce new known errors by randomly modifying the original annotations. To measure the error detection rate accurately, Fig. 3, we inject four noise types simulating common dataset decays: localization shifts, classification swaps, missing annotations, and extraneous boxes.

TABLE I: (IV-E.1) Evaluation metrics of our system using different VLMs on $D^{\mathrm{noise}}_{\mathrm{K:test}}$. We see that a poor performer (Llama) can be drastically improved with fine-tuning (FT, FT-CoT).

Model	Eval. Type	Rec.	Prec.	Acc.	F1
-	Error Proposals	**0.96**	0.72	0.89	0.82
DeepSeek-VL2 [30]	Error Validation	0.58	**0.99**	0.69	0.73
	AutoVDC System	0.55	**0.99**	0.87	0.71
Gemini Flash 2.0 [31]	Error Validation	0.84	0.83	0.76	0.83
	AutoVDC System	0.80	0.83	0.90	0.81
GPT 4.1 [32]	Error Validation	0.84	0.84	0.77	0.84
	AutoVDC System	0.80	0.84	0.90	0.82
ViP-LLaVA [25]	Error Validation	0.61	0.71	0.53	0.66
	AutoVDC System	0.58	0.71	0.82	0.64
Llama [33]	Error Validation	0.47	0.89	0.58	0.62
	AutoVDC System	0.45	0.89	0.83	0.60
Llama (FT)	Error Validation	0.91	0.91	0.87	0.91
	AutoVDC System	0.88	0.91	0.94	0.89
Llama (FT-CoT)	Error Validation	**0.96**	0.94	**0.92**	**0.95**
	AutoVDC System	**0.92**	0.94	**0.96**	**0.93**

KITTI: This dataset consists of 63,160 bounding box annotations across nine classes, which we divided into 80:10:10 for training, validation, and testing subsets ($D_{\mathrm{K:train}}$, $D_{\mathrm{K:val}}$, $D_{\mathrm{K:test}}$). We filter out non-semantic labels, like *DontCare* and *Misc*, and also heavily occluded or truncated objects to ensure visual clarity for the VLM, The noise-induced version, $D^{\mathrm{noise}}_{\mathrm{K:test}}$, contains around 30% of evenly distributed erroneous annotations.

nuImages: For larger-scale evaluation, we used nuImages, which contains over 800,000 bounding boxes and 23 foreground classes. We followed the official splits for training, validation, and testing ($D_{\mathrm{nu:train}}$, $D_{\mathrm{nu:val}}$, $D_{\mathrm{nu:test}}$). No filtering was applied to test AutoVDC on a more complex and diverse set of real-world data. The noise-injected versions, $D^{\mathrm{noise}}_{\mathrm{nu:train}}$ and $D^{\mathrm{noise}}_{\mathrm{nu:test}}$, contain around 15% of evenly distributed injected noise.

B. EP Stage: High-Recall Filtering

The Error Proposal (EP) stage utilizes DETR [29] detector to generate candidates.

KITTI Setup: The task model trained on clean set, $D_{\mathrm{K:train}}$, to establish a performance ceiling using a strong proposer.

nuImages Setup: Intentionally trained on a corrupted set ($D^{\mathrm{noise}}_{\mathrm{nu:train}}$) to mimic a realistic weak proposer where existing labels are already imperfect.

C. EV Stage: Fine-Tuning and CoT parameters

We utilize the Llama-3.2-Vision-11B model [33] in both zero-shot mode, fine-tuned(FT) and fine-tuned with Chain of Thought (FT-CoT) modes. We evaluate a diverse selection of VLMs to provide a representative snapshot of the multi-modal reasoning maturity. To achieve dataset alignment:

Fine-Tuning Split: We re-purposed the validation sets from both of our datasets, $D_{\mathrm{K:val}}$ and $D_{\mathrm{nu:val}}$,

TABLE II: (IV-E.2) Evaluation on $D^{\text{noise}}_{\text{nu:test}}$ of AutoVDC with DETR trained with $D^{\text{noise}}_{\text{nu:train}}$ and Llama. Even with a weak task model, AutoVDC can still reliably find erroneous annotations.

Model	Eval. Type	Rec.	Prec.	Acc.	F1
-	Error Proposals	0.86	0.19	0.43	0.31
Llama [33]	Error Validation	0.51	0.30	0.68	0.38
	AutoVDC System	0.44	0.30	0.76	0.36
Llama (FT-CoT)	Error Validation	0.89	0.78	0.93	0.83
	AutoVDC System	**0.77**	**0.78**	**0.93**	**0.77**

by injecting 50% noise to create noise-induced subsets, $D^{\text{noise}}_{\text{K:FT}}$ and $D^{\text{noise}}_{\text{nu:FT}}$ for fine-tuning.

Training Setup: We employed LoRA [34] for fine-tuning, updating the query and value projection modules. The input consisted of image with either noise-injected or clean annotations, and the target Y indicated whether they are erroneous. In the FT-CoT variant, the target output Y was augmented with a text-based justification.

For the text prompt, we use *"Does this box contain the* $\{\text{class\_label}\}$. *If yes, does it fit the object well enough?"*.

D. Evaluation Criteria

The **EP evaluation** addresses the triage logic of the framework. By comparing the error proposals D' against the ground truth D, we quantify the Error Proposal stage's ability to act as a high-recall filter.

The **EV evaluation** focuses on the reasoning capabilities of the VLM. By comparing the validated output D'' against the candidates provided by the first stage D', we can isolate the VLM's expert judgment to measure its ability to resolve the spatial reasoning gap and the effectiveness of domain-alignment, such as FT-CoT.

The **AutoVDC System Evaluation** assesses the end-to-end effectiveness. We compare the final validated errors D'' against the original dataset D to evaluate performance across varying noise scales.

The **Task Model Evaluation** demonstrates the practical impact on task development. By re-evaluating the perception model across clean D, noisy D^{noise}, and cleaned D^{cleaned} datasets, we show how using VLMs can have a real impact on data curation and, in turn, restores evaluation fidelity.

E. Experimental Results

In order to quantify the effectiveness of the proposed framework, we conduct the following experiments.

*1) **VLM Sensitivity and the Spatial Reasoning Gap**:* Our first study investigates whether general-purpose foundation models can identify annotation errors out-of-the-box or if they require task-specific alignment. We use the DETR task model trained on $D_{\text{K:train}}$ as our task model for the EP stage for this part of the study.

In Tab. III, zero-shot models such as Gemini Flash 2.0 and GPT-4.1 demonstrate decent semantic capabilities,

TABLE III: (IV-E.1) Error Validation Recall metrics of different VLMs for each noise type on $D^{\text{noise}}_{\text{K:test}}$. The VLMs are particularly weak on handling localization errors, which benefits most from fine-tuning.

Model	Localization Noise	Classification Noise	Missing Ann. Noise	Extraneous Ann. Noise
DeepSeek-VL2 [30]	0.01	0.57	**0.99**	0.80
Gemini Flash 2.0 [31]	0.65	0.96	0.74	0.99
GPT 4.1 [32]	0.59	0.95	0.80	1.00
ViP-LLaVA [25]	0.58	0.66	0.41	0.74
Llama [33]	0.14	0.38	0.89	0.52
Llama (FT)	0.85	0.9	0.92	1.00
Llama (FT-CoT)	**0.90**	**0.97**	0.95	**1.00**

achieving over 0.95 recall on classification noise. However, they exhibit a significant spatial reasoning gap. In Tab. I, DeepSeek-VL2 has high precision of 0.99, it has low recall of 0.55, indicating that it is unable to detect almost half the noise in the dataset. Tab. III clarifies this inability by showing that DeepSeek-VL2 is an expert at identifying missing annotations (0.99 recall) but it fails almost entirely on localization noise (0.01 recall).

As shown in Tab. I, FT-CoT transforms the base Llama model from a low zero-shot performer (F1: 0.60) to our strongest validator (F1: 0.93). Most importantly, Tab. III highlights that FT-CoT recovers the missing spatial performance, raising localization recall from the baseline 0.14 to a robust 0.90. This confirms that for high-fidelity curation, domain-specific alignment through CoT reasoning is a fundamental requirement to resolve intricate spatial errors.

*2) **Resilience to Weak Candidate Proposals**:* We tested this by intentionally training a DETR model on a dataset with a high noise rate of 15%, $D^{\text{noise}}_{\text{nu:train}}$. This created a weak proposer for the nuImages dataset. We evaluate using $D^{\text{noise}}_{\text{nu:test}}$. As shown in Tab. II, the weak task model was very noisy, producing error proposals with a precision of only 0.19. However, our Llama (FT-CoT) was able to filter through this noise, maintaining a high validation precision of 0.78. This proves the EP stage's role as Economic Triage by flagging suspicious samples, it reduced the VLM's total workload substantially. This allowed the VLM to focus its expensive reasoning power only on the most likely errors.

*3) **Real-World Discovery and Noise Scaling**:* We tested the performance on the KITTI test set with injected noise rates ranging from 5% to 30%, in $D^{\text{noise}}_{\text{K:test}}$. Llama (FT-CoT) model, Fig. 4, maintains consistent accuracy and recall metrics across the ranges. While precision naturally decreases at lower noise rates as there are fewer errors to find giving more weightage to the false positives, the overall performance remains stable.

Beyond synthetic noise, we also evaluated the system's ability to find actual errors in the original, un-

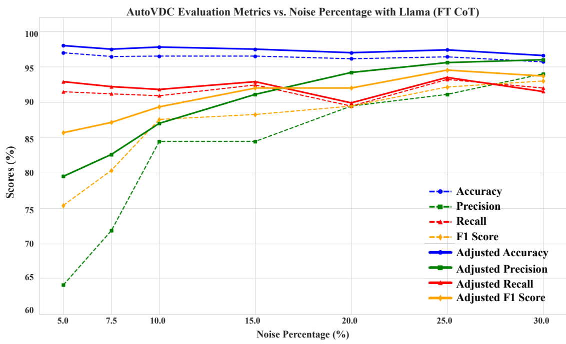

Fig. 4: (IV-E.3) Real, non-injected erroneous annotations can lead to biased results at lower noise rates, so removing the identified 39 real errors is critical for accurate assessment (Adjusted), but is mitigated by using larger noise rates.

altered KITTI dataset, $D_{\text{K:test}}$. Even at a 0% injected noise rate, AutoVDC identified 71 suspicious annotations, 39 of which were confirmed as genuine mistakes in the original ground truth, mainly included completely occluded objects, even though the dataset was filtered based on the occlusion annotation attribute. This proves that the study isn't just on artificial errors we added but it actually finds real mistakes that humans missed in prominent datasets.

For datasets with varying noise rates, Fig. 4, we report the performance metrics both before and after mitigating the effect of 39 erroneous annotations ("Adjusted"). Accounting for these 39 real errors significantly improves our performance metrics across all noise rates. Overall, we see that at all noise rates, AutoVDC keeps consistent accuracy and recall metrics while dropping precision at lower noise rates.

*4) **Impact on Task Model Evaluation**:* In Tab. IV, we established a baseline using the clean KITTI test set $D_{\text{K:test}}$, where a DETR achieves 0.38 precision and 0.54 recall. The model's perceived performance collapses to 0.22 precision and 0.36 recall on $D^{\text{noise}}_{\text{K:test}}$. We created clean versions of the noisy dataset using three VLM variants: zero-shot Llama, fine-tuned Llama (FT), and fine-tuned Llama with Chain-of-Thought (FT-CoT). $D^{\text{cleaned}_{\text{FT-CoT}}}_{\text{K:test}}$ had 696 erroneous annotations removed and regained 11% to reach a precision of 0.33 and match the baseline's recall at 0.54.

V. Discussion

VLM Readiness and the Spatial Gap: Our findings in IV-E.1 provide a readiness report for current VLMs in automated data curation. We observe a clear hierarchy in task difficulty. While zero-shot VLMs are highly mature for semantic classification, a significant spatial reasoning gap persists. As shown in Tab. II, foundation models cannot yet reliably audit fine-grained localization

TABLE IV: (IV-E.4) Evaluation of DETR on cleaned up datasets along with a breakdown of which annotations were removed ("Ann. Rm.") after cleaning KITTI test dataset with 758 noise-injected erroneous annotations. AutoVDC gets close to the "true" evaluation numbers.

Dataset	DETR Evaluation		# Erroneous	# Non-Erroneous
	Avg. Rec.	Avg. Prec.	Ann. Rm.	Ann. Rm.
$D_{\text{K:test}}$	0.54	0.38	-	-
$D^{\text{noise}}_{\text{K:test}}$	0.36	0.22	-	-
$D^{\text{cleaned}}_{\text{K:test}}$	0.4	0.24	344	41
$D^{\text{cleaned}_{\text{FT}}}_{\text{K:test}}$	0.52	0.32	666	67
$D^{\text{cleaned}_{\text{FT-CoT}}}_{\text{K:test}}$	**0.54**	0.33	696	46

without specific alignment. The need for extensive fine-tuning may decrease as the native spatial reasoning of industry VLMs improves.

The Efficiency of Triage Logic: A core contribution of the AutoVDC framework is the two-stage methodology that exploits the generalization gap of neural networks. As demonstrated in IV-E.2, the system is resilient to a weak proposer. Even with an EP Stage precision of only 0.19, the VLM acts as a robust gatekeeper, maintaining high system precision. This economic triage logic is vital for scalability, as it reduces the human auditing workload exponentially and making high-fidelity curation feasible for massive production datasets.

Human-in-the-Loop and the Long-Tail: As seen in Tab. IV, while FT-CoT is the most precise validator, removing only 46 valid samples compared to the 67 removed by standard fine-tuning, we recognize that automated cleaning cannot yet totally replace human oversight. These removed samples often represent the long-tail of the dataset, cases that are difficult or ambiguous that they appear erroneous even to advanced models. The discovery of 39 genuine, non-synthetic errors in the original KITTI ground truth empirically validates that the FT-CoT reasoning generalizes beyond injected patterns to capture real-world dataset decay. Rather than eliminating manual review, AutoVDC transforms it from an exhaustive search into a targeted verification of high-confidence candidates.

Systematic Bias: A limitation of the current approach is systemic label bias. If a dataset contains consistent, repetitive error patterns, the task model may learn them as features, preventing the EP stage from flagging them as discrepancies. Addressing systematic label biases represents a promising area for future research.

Noise Modeling: We view noise modeling as an adjustable parameter. This helps the finetuned VLM understand the specific characteristics of the dataset during testing. We can then better align the VLM's error detection capabilities with the nuances of the dataset, thereby improving its performance.

Methodological Extension: To evaluate cross-task modularity, we conceptually extended AutoVDC to 3D

Bird's Eye View (BEV) occupancy task. By projecting 3D voxel discrepancies into the 2D perspective domain, we enabled the VLM to act as a cross-view auditor, successfully identifying ghost objects and missed obstacles in a preliminary proof-of-concept. This provides a robust road-map for future research, suggesting that general-purpose VLMs can act as universal auditors for complex perception tasks, provided the errors are correctly mapped into the model's native visual space.

VI. Conclusion

This work introduced AutoVDC, a modular framework designed to evaluate the maturity of Vision-Language Models (VLMs) for automated data curation. By employing a two-stage architecture for our study, leveraging the latent generalization of task models to propose error candidates and the reasoning power of VLMs for final validation, we successfully bridge the gap between model intuition and ground-truth integrity.

Our systematic study reveals that while zero-shot VLMs are highly effective for semantic classification, specialized Chain-of-Thought (CoT) fine-tuning is essential for resolving fine-grained localization noise. Furthermore, we proved that the generalization gap of task models provides a reliable signal for noise discovery even when trained on corrupted labels. Our experiments confirm that the framework remains robust when guided by weak task models, with the VLM acting as a high-precision gatekeeper that reduces manual review workloads exponentially.

Instead of auditing the entire dataset or even the full list of error proposals, manual verifiers can focus exclusively on the high-confidence candidates flagged by the VLM. This significantly cuts down on human effort and expense while ensuring that the long-tail samples are preserved. Ultimately, this methodology offers a robust roadmap for shifting toward automated, high-fidelity data engines across diverse vision-based applications.

References

[1] X. Chen, B. Mersch, L. Nunes *et al.*, "Automatic labeling to generate training data for online lidar-based moving object segmentation," *RA-L*, vol. 7, no. 3, 2022. 1

[2] P. Sun, H. Kretzschmar, X. Dotiwalla *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, June 2020. 1

[3] B. Wilson, W. Qi, T. Agarwal *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *NeurIPS Datasets and Benchmarks*, 2021. 1

[4] S. Antol, A. Agrawal, J. Lu *et al.*, "VQA: Visual Question Answering," in *ICCV*, 2015. 1

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012. 2, 4

[6] H. Caesar, V. Bankiti, A. H. Lang *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020. 2, 4

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. 2

[8] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *SIGMOD/PODS*, 2000. 2

[9] M. Bernhardt, D. C. Castro, R. Tanno *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature Communications*, vol. 13, no. 1, 2022. 2

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017. 2

[11] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR*. Springer London, 1994. 2

[12] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep. TR 1648, 2009. 2

[13] H.-M. Heyn, K. M. Habibullah, E. Knauss *et al.*, "Automotive perception software development: An empirical investigation into data, annotation, and ecosystem challenges," 2023. 2

[14] M. Liu, E. Yurtsever, J. Fossaert *et al.*, "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook," 2024. 2

[15] L. Zheng, W.-L. Chiang, Y. Sheng *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. 2

[16] H. Li, Q. Dong, J. Chen *et al.*, "Llms-as-judges: A comprehensive survey on llm-based evaluation methods," 2024. 2

[17] X. Guo and Y. Chen, "Generative AI for synthetic data generation: Methods, challenges and the future," 2024. 2

[18] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," 2021. 2

[19] B. Xiao, H. Wu, W. Xu *et al.*, "Florence-2: Advancing a unified representation for a variety of vision tasks," 2023. 2

[20] N. Ravi, V. Gabeur, Y.-T. Hu *et al.*, "SAM 2: Segment anything in images and videos," 2024. 2

[21] P. Agrawal, S. Antoniak, E. B. Hanna *et al.*, "Pixtral 12b," 2024. 2

[22] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. 2

[23] M. Abdin, J. Aneja, H. Awadalla *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," 2024. 2

[24] C. Schuhmann, R. Beaumont, R. Vencu *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," 2022. 2

[25] M. Cai, H. Liu, D. Park *et al.*, "ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts," 2024. 2, 4, 5

[26] J. Kim, Y. Ku, J. Kim, J. Cha, and S. Baek, "Vlm-pl: Advanced pseudo labeling approach for class incremental object detection via vision-language model," in *CVPR Workshops*, 2024. 2

[27] M. Diab, G. Barchi, and D. Moser, "Vision-language models as pseudo-label validators in semi-supervised learning: Geolocating medium voltage cabins from google street view imagery," 01 2025. 2

[28] H. Lu, Y. Bian, and R. C. Shah, "ClipGrader: Leveraging vision-language models for robust label quality assessment in object detection," 2025. 2

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020. 4

[30] Z. Wu, X. Chen, Z. Pan *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," 2024. 4

[31] Google, "Gemini flash 2.0," accessed on April 2025. 4, 5

[32] OpenAI, "Chatgpt (gpt-4.1)," accessed on April 2025. 4, 5

[33] A. G. et al., "The llama 3 herd of models," 2024. 4, 5

[34] E. J. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022. 5