Detailed Themes Development

Dimension(Themes)	Category(Sub themes)	Sub-Type (Code)	Quote (from ID2)
Completeness	Pre-annotation Issues	Low-quality pre-annotations	"Another example is pre-annotations. If a client provides pre-annotated data, we've developed tools to make the process more efficient. Sometimes pre-annotations are of poor quality, and it's faster to redraw them than fix errors."
	Missing Annotations	Missed objects / Selective omission / Rare case omission	"We had to annotate animals, including moose. Some annotators may not even know what a moose looks like! ... there are two types of moose—the European moose and the North American moose—and they look quite different. If someone doesn't know that, it adds even more confusion."
Accuracy	Boundary Errors	Bounding box errors	"If you ever ran a standard object detection model, you can label all the people and vehicles in a picture, and you'll notice the boxes are drawn quite loosely. Some clients, however, insist on pixel-perfect accuracy—like a 10-pixel margin or 10 cm box drawings."
Consistency	Guideline Issues	Ambiguous instructions	"Clients often provide guidelines, but we frequently revise them to improve efficiency. I advocate for simplification—annotators should not have to read 140 pages to understand the guidelines."
	Drift Over Time	Batch effects / Cross-project drift	"Scaling is a difficult problem. ... There's a difference in quality across annotators, and things can quickly go wrong when you scale to a larger workforce."
	Tool-driven Variability	Automation effects	"Recently, we introduced automatic tracking for sequences. ... The idea is to minimize repetitive tasks so the human annotator can focus on critical decisions rather than manual work. ... Overreliance on automated tools could lead to errors."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID3**
Completeness	Guideline Gaps	Missing feedback loop	"The most striking aspect missing from this process flow is the feedback loop. ... Without feedback loops, they were unable to adjust their approach during the annotation process, leading to inefficiencies and wasted resources."
	Policy Ambiguity	Incorrect annotation skipping	"A common annotation guideline might state that vehicles inside buildings should not be annotated. ... According to the guideline, these vehicles should be ignored, but they might still be relevant for perception systems."
	Unexpected Cases	Edge/unforeseen scenarios	"Insights can be gained by examining the data ... often revealing unexpected cases that were not initially considered. ... Our main challenge is ensuring that annotators flag unexpected cases so that we can consult the client."
Accuracy	Labeling Errors	Wrong class label	"One major challenge is that multiple annotations can be considered correct ... deciding whether an object is visible or not often involves edge cases ... making it difficult to establish a rigid quality assurance process."
	Instruction Errors	Insufficient guidance	"If annotators are given vague or difficult-to-interpret guidelines, they may not annotate data consistently. ... This can lead to discrepancies in annotations and variations in how different annotators interpret the same data."
	Systematic Mislabeling	Bias-driven	"Subsets systematically mislabeled (e.g., red cars) ... The original annotation may not have been wrong, but the supervisor might have simply had a different interpretation."
Consistency	Human Variability	Inter-annotator disagreement	"Human interpretation of guidelines varies, leading to inconsistencies. ... Achieving extremely high quality is possible by increasing the number of review steps, but this significantly raises costs."
	Supervisor Overrides	Reviewer–annotator mismatch	"Another issue arises when a supervisor conducts a spot check and modifies an annotation. The original annotation may not have been wrong, but the supervisor might have simply had a different interpretation. This can lead to unnecessary changes and inefficiencies."
	Guideline Issues	Ambiguous instructions	"What I meant by low-quality guidelines are those that lack clarity and consistency. ... They often contain redundant or unclear instructions, making them difficult to read and follow."
	Standardization Issues	Lack of frameworks	"One of the biggest challenges is the lack of standardization in the annotation market. ... It is difficult to compare different annotation providers because there is no universal benchmark."
	Human Factors	Fatigue / repetition	"Repetitive work does not necessarily produce high-quality annotations. ... A lack of clear understanding due to poorly structured instructions can lead to inconsistencies."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID4**
Completeness	Missing Annotations	Missed objects	"Annotation quality means ensuring every object is annotated correctly. We measure quality based on missed annotations and misclassifications. These errors impact the models directly."
	Policy Ambiguity	Incorrect annotation skipping	"In the older approach, we had an 'unclear' property for objects that were difficult to classify. However, this property was often overused or underused by annotators, making it unreliable. For Z2, we've removed the 'unclear' property altogether."
	Sensor/Data Gaps	Malfunction-driven	"In some projects, we've faced issues where sensors malfunctioned intermittently, leading to gaps in data quality. Ensuring all sensors work properly is critical."

Detailed Themes Development

	Scenario Coverage	Rare case omission	“Improving annotation quality had the single biggest impact on performance. Specifically, eliminating errors in rare traffic sign classes drastically improved the recall and reduced error rates.”
Accuracy	Labeling Errors	Wrong class label	“One of the biggest improvements we saw was in refining traffic sign classification. There were many mistakes in the annotations, and correcting those errors led to substantial improvements.”
	Boundary Errors	Bounding box errors	“Bounding boxes have the most direct impact for our purposes because our end goal is to detect objects in the correct position and orientation.”
	Systematic Mislabeling	Bias-driven	“Biases also occur in driving behavior. For instance, if the data mostly comes from careful drivers, the model won’t learn how to handle erratic driving scenarios. It’s important to account for such biases in both data collection and annotation.”
	Pseudo-label Bias	Rare-class weakness	“One common bias comes from pseudo annotations, which tend to be weaker on rare object classes. We try to mitigate this during data collection by designing processes to capture rare cases.”
Consistency	Guideline Issues	Guideline evolution	“We can make small updates every day. But in terms of publishing a new guideline, that happens every few months. Too many changes too often can be difficult for the annotation companies.”
	Standardization Issues	Lack of frameworks	“We touched on different processes for standards, but there aren’t specific standards that we use. It’s more about addressing our unique needs.”
	Annotator Context Gap	Lack of purpose knowledge	“The annotators sometimes don’t understand why they are annotating certain things or how it ties into the overall purpose. They often say, ‘If we knew what this was supposed to train for, we could do it better.’”
	Team Coordination Gap	Misaligned hand-offs	“We provide guidance and interact with people who are customer-facing, but we don’t always see the full chain of how guidelines are conveyed to annotators or quality assurance teams. Sometimes, not being aware of their internal processes feels like a weakness.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID5**
Completeness	Missing Annotations	Selective omission	“Sometimes the machinery was not working right… the model was maybe not constantly changing, but there were updates… At the end, it’s always a question of time. There are always too few resources for quality assurance, just focusing on getting the rest of the stuff done.”
	Temporal Gaps	Frame skipping	“In our projects, dynamic objects in 3D cuboids are the most challenging aspect… the model flickered frame by frame, unsure if a certain area of the scene was drivable or not.”
	Incomplete Attributes	Attribute omission	“If the scene was dark or contaminated by water droplets, it became very ambiguous. Annotators had different approaches… This noise clearly impacted the perception system.”
	Pre-annotation Issues	Low-quality pre-annotations	“We still have a human in the loop… we can easily modify and feed it with pseudo labels. A larger model provides examples, but these must be fine-tuned by a person.”
	Scenario Coverage	Selection bias	“Some projects focus on specific events rather than statistical sampling… In our case, we had to define the data collection to focus on those events. This inherently introduces bias.”
Accuracy	Instruction Errors	Insufficient guidance	“Initially, it wasn’t clear how to annotate in cases where visibility was poor… Annotators had different approaches to such situations, which introduced a lot of noise.”
Consistency	Guideline Issues	Ambiguous instructions	“There are contradictory requirements—you can’t have very concise guidelines and at the same time cover all details. If guidelines are too descriptive, they become a nightmare to understand.”
	Drift Over Time	Batch effects	“It’s not linear… When you start an annotation project, there is a learning curve… feedback loops and lessons learned may require data to be re-annotated.”
	Edge Case Handling	Inconsistent treatment	“If the model was not constantly changing, it was ambiguous how to annotate in poor visibility or contamination cases. Different annotators treated them differently.”
	Team Coordination Gap	Misaligned hand-offs	“We had to add our requirements for annotation on top of another team’s guidelines… We had to be careful not to create contradictions. Discussions and alignment were critical.”
	QA Capacity Constraints	Limited review & logging	“At the end, it’s always a question of time. There are always too few resources for quality assurance, just focusing on getting the rest of the stuff done.”
	Requirement Contradictions	Dual/re-annotation needed	“First, once you settle on the first version of requirements, new requirements shouldn’t contradict existing ones. If they are contradictory, that’s a problem… It may require dual-annotation or re-annotation.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID6**
Completeness	Missing Annotations	Partial coverage	“Annotating pedestrians in an urban setting introduces ambiguities… If every pedestrian in an image is annotated, including small, distant figures in the background, the model may be overwhelmed by irrelevant details.”

Detailed Themes Development

	Temporal Gaps	Frame skipping	“In practice, dynamic objects in 3D cuboids are challenging—the model flickered frame by frame, unsure if a certain area was drivable or not.”
	Incomplete Attributes	Attribute omission	“If the scene was dark or contaminated by water droplets, it became very ambiguous. Annotators had different approaches… this noise clearly impacted the perception system.”
	Pre-annotation Issues	Low-quality pre-annotations	“Many data annotation companies integrate tool-assisted annotation to streamline the process. Tools suggest annotations, which human annotators review and confirm. However, this raises concerns: are annotators simply approving suggestions without thorough validation?”
	Guideline Gaps	Missing feedback loop	“The iterative process should mature over time, incorporating safety analysis and structured requirements to validate adequacy of datasets. Without feedback loops, there’s no structured improvement in annotation quality.”
	Unexpected Cases	Edge/unforeseen scenarios	“Ambiguities arose—for example, whether a pedestrian on a scooter should be annotated as relevant or not. These unforeseen cases caused inconsistency.”
	Scenario Coverage	Selection bias	“If pedestrian detection data reflects real-world distribution, it may contain fewer samples of wheelchair users. To ensure equal detection probability, data collection must be adjusted—otherwise, bias is introduced.”
Accuracy	Labeling Errors	Wrong class label	“Validation noise in test sets was a big problem—misclassified objects made it difficult to assess true model accuracy.”
	Instruction Errors	Insufficient guidance	“The definition of a ‘relevant pedestrian’ varied among annotators… without well-defined standards, ambiguities in human annotations led to inconsistencies.”
	Systematic Mislabeling	Bias-driven	“Automation tools often rubber-stamped labels—subsets became systematically mislabeled because annotators over-trusted tool outputs.”
Consistency	Human Variability	Inter-annotator disagreement	“One annotator may consider a pedestrian on a scooter relevant… another may prioritize those closer to the vehicle. These differences in interpretation created inconsistencies.”
	Guideline Issues	Ambiguous instructions	“Determining which pedestrians are ‘relevant’ was unclear—guidelines did not define it strictly, leading to divergent interpretations.”
	Automation Effects	Automation validity risk and pseudo-label propagation	“Semi-automatic labels replicated earlier model errors. Human validation is always required; otherwise, errors will persist.”
	Tool-driven Variability	Automation effects	“If an annotation tool mislabels test data, leading to incorrect model validation, it becomes a critical risk factor.”
	Standardization Issues	Lack of frameworks	“The industry lacks common standards for annotation—different providers use different practices, making results inconsistent.”
	Human Factors	Fatigue / repetition	“Human annotators are prone to errors and inconsistencies, especially in long, repetitive tasks. Fatigue reduces accuracy.”
	Annotator Context Gap	Lack of purpose knowledge	“Requirements pass through multiple stakeholders, and annotators often don’t know the true purpose of labeling. This loss of context leads to inconsistent quality.”
	Team Coordination Gap	Misaligned hand-offs	“OEMs define requirements, but these may not be directly shared with suppliers. By the time they reach annotators, the instructions are incomplete or altered—creating inconsistencies.”
	QA Capacity Constraints	Limited review & logging	“Given the vast data, manual review is not economically viable. With too few QA resources, errors remain undetected.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID7**
Completeness	Guideline Gaps	Missing feedback loop	“We need spot checks. We need review. We need feedback. We need to verify our annotation … If we just annotate it and that’s all, the developer will think his model is not performing good, but the problem is with the annotation itself.”
	Scenario Coverage	Rare case omission / Selection bias	“Something missing … people using E-scooters. They are not annotated, sometimes included with pedestrians but mostly skipped … If you don’t include them, the model will not learn and may fail in real-world cases.”
Accuracy	Labeling Errors	Wrong class label	“If you mistakenly put the wrong start or end frame when annotating pedestrian crossing, the model will learn the wrong thing … For example, swapping 200 and 300 in start–end frames makes the prediction error appear as if the model is wrong.”
	Temporal Accuracy	Tracking errors	“I put the numbers like the start crossing at the end crossing and the end crossing I put in the start … 300 before 200 … This is wrong because frames are sequential. Such mistakes directly affect classification.”
Consistency	Human Variability	Inter-annotator disagreement	“Some people will say sunny, some will say foggy … some will annotate high traffic, others low traffic. If there is no common definition, errors appear, and the model evaluation will be wrong.”

Detailed Themes Development

	Human Variability	Intra-annotator drift	“To be consistent and correct in each video for hours and hours is impossible. As humans, we always make mistakes, especially when repeating the same frame-by-frame annotation for days.”
	Guideline Issues	Ambiguous instructions	“We need to have a common definition … For example, defining what is sunny, foggy, or high traffic. Without clear definitions, each annotator applies their own interpretation.”
	Guideline Issues	Guideline evolution	“E-scooters were not in older datasets but today they are common. Guidelines must evolve with new scenarios, otherwise important cases are missed.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID8**
Completeness	Guideline Gaps	Missing feedback loop	“Annotators should have a mechanism to report unclear cases. Continuous updates to annotation requirements based on real-world feedback help ensure guidelines remain relevant.”
	Unexpected Cases	Edge/unforeseen scenarios	“A system for annotators to report edge cases is crucial. Since annotators may lack domain knowledge, expert review of reported cases can help refine annotation requirements.”
	Scenario Coverage	Selection bias / Rare case omission	“Yes, identifying underrepresented data categories can significantly enhance system performance. For example, if a dataset lacks images of bald individuals and the system struggles to detect them, synthetic data can be generated or additional real-world examples can be sourced to improve model accuracy.”
Accuracy	Boundary Errors	Bounding box errors	“Ambiguity arises when guidelines leave room for interpretation. For example, in bounding box annotation, should occluded parts of an object be included?”
	Schema Accuracy	Granularity mismatch	“If the requirement specifies capturing the entire head but annotators only mark facial features, this introduces bias.”
	Systematic Mislabeling	Bias-driven	“Bias can stem from underrepresentation of certain demographics or inconsistencies in labeling. For example, in bounding box annotation, guidelines might instruct annotators to either mark only visible parts of an object or infer the occluded parts. Annotators may interpret these instructions differently, leading to inconsistency in the dataset.”
Consistency	Guideline Issues	Ambiguous instructions	“Ambiguity arises when guidelines are unclear or open to interpretation. Including visual examples and peer review processes can help. Annotators should also have a mechanism to report unclear cases.”
	Guideline Issues	Guideline evolution	“Annotators should be able to report scenarios not covered by existing guidelines so that the requirements can be updated iteratively.”
	Annotator Context Gap	Lack of purpose knowledge	“Annotators sometimes don’t know the true purpose of labeling. Without awareness of why data is annotated, it is difficult to ensure consistent quality.”
	Team Coordination Gap	Misaligned hand-offs	“Companies that outsource annotation to third-party providers often have less control over adherence to guidelines.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID9**
Completeness	Incomplete Attributes	Attribute omission	“Uncertainty in annotations should be logged, particularly for difficult cases. If multiple annotators work on the same data, their agreement level should be considered. Similarly, if annotations are generated by a mix of automated processes and human experts, this should be recorded. Having traceability in annotation decisions would allow issues to be diagnosed and improved later.”
	Guideline Gaps	Missing feedback loop	“Traceability should be maintained throughout the pipeline. Automated and manual quality control measures should be integrated. Metrics should be defined based on the intended use of the data.”
	Policy Ambiguity	Privacy/compliance omissions	“The GDPR requirements remain somewhat unclear, and there is an inherent risk in misinterpreting these guidelines. To ensure compliance, organizations should involve GDPR experts or legal professionals in the process.”
	Scenario Coverage	Selection bias	“If we collect data on varying weather conditions and vehicle speeds but lack sufficient samples, we might unintentionally introduce a correlation between speed and weather. This can lead to biased learning, as the model may assume certain relationships exist when they actually do not.”
Accuracy	Boundary Errors	Bounding box errors	“In quality assurance: bounding box annotations should be evaluated based on their placement, size, and coverage of the intended object. There should also be guidelines to define acceptable levels of error, such as a bounding box margin of no more than 10% beyond the object boundaries.”
	Segmentation Errors	Mask errors	“Annotations should be revisited based on system performance feedback … pixel-wise segmentation, boundary errors between objects can affect results significantly.”
	Schema Accuracy	Granularity mismatch	“Changes in requirements, such as redefining class categories, may necessitate re-annotating data. For example, if gender categories are further subdivided into age groups, previously annotated data may require updates.”

Detailed Themes Development

	Systematic Mislabeling	Bias-driven	"Relying entirely on automated annotations risks bias, as the model may learn the biases of the annotation system itself. A mixed approach, where human annotations validate and correct machine-generated labels, is preferable."
Consistency	Human Variability	Inter-annotator disagreement	"One approach is inter-annotator agreement, which assesses how consistently different annotators label the same data … some level of disagreement among annotators is expected and even informative, but avoidable annotation errors should be minimized."
	Guideline Issues	Ambiguous instructions	"One of the biggest challenges in defining annotation requirements is reducing ambiguity. Some class definitions are inherently unclear, and different annotators may interpret them differently."
	Guideline Issues	Guideline evolution	"Requirements should not change frequently, but in some cases, they need updates to accommodate new use cases. When changes occur, it is important to track which data needs re-annotation to maintain consistency."
	Tool-driven Variability	Cross-modality alignment	"When multiple sensors record the same scene, they need to be registered together. If we provide separate datasets from different sensors, there must be a method to align them. Some companies have referred to this as sensor fusion data, where different annotation types need to be combined across sensor modalities."
	Standardization Issues	Lack of frameworks	"There is no universal standard for annotation tools because they serve different purposes. However, it is essential to balance human and automated annotations."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID10**
Completeness	Incomplete Attributes	Attribute omission	"Another issue is adding proper metadata. Metadata is crucial because it allows others to find and use the data effectively … older datasets become outdated, and keeping track of changes becomes increasingly difficult. If data is collected one month and then collected again later, there could be discrepancies if someone replaced a sensor or reinstalled it incorrectly. Additionally, software updates can alter sensor outputs, leading to inconsistencies compared to earlier logs."
	Guideline Gaps	Missing feedback loop	"One approach to ensuring data quality is to implement a gateway that checks the quality of incoming data during the preprocessing step … before going out to collect data, a cross-check with well-known reference objects can be performed … after data collection, another check can be conducted … this helps ensure the integrity of the data being included in the dataset."
	Policy/Method Gaps	Modality not annotated	"Typically, camera images are annotated frame by frame, as are LiDAR point clouds. However, radar data is generally not annotated because there is no established method for doing so—unless it involves 4D radar, which is more relevant but still quite noisy."
	Sensor/Data Gaps	Calibration/synchronization issues	"Some sensors capture data momentarily, while others spread data collection over time. This means you need a way to synchronize them so that data points from different sensors align correctly in time and space … otherwise they will show up in different places when you annotate them in the real world."
	Scenario Coverage	Selection bias	"It used to be that you wanted to have as large annotated datasets as possible … but if you have millions of records from Sweden at a certain time, you get commuters but not school children … in that case more data would not help you. Diversity comes into play."
Accuracy	Boundary Errors	Bounding box errors	"When annotating across a sequence of frames or point clouds … the bounding box may grow, shrink, or change dimensions … ideally, annotation quality should ensure that variations across frames are smoother and more consistent."
	Cuboid drift	3D drift errors	"The bounding box may grow, shrink, or change dimensions over a sequence, which can be frustrating because in reality, the object's shape likely remains more stable. The variations might stem from noise, changes in the number of detected points, or vessel movement."
Consistency	Tool-driven Variability	Cross-modality alignment	"If you have multiple sensors you want to be able to annotate in different modalities at the same time … but they often sample at different time instances. You need calibration to make sure they align."
	Standardization Issues	Lack of frameworks	"I suppose … there should be standards for it, I mean for how annotations should be done. Even though there are multiple standards, at least you don't have to start from scratch. But I never used one myself."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID11**
Completeness	Guideline Gaps	Missing feedback loop	"It would be best to test the guidelines—maybe do some AB testing before rolling them out. For example, you could try crowd-sourcing or small-scale testing to ensure they're clear and effective. Putting many people to work on annotation without testing the guidelines can have a big impact on perception models and their performance."
	Scenario Coverage	Selection bias	"Let's say you have equal amounts of negative and positive cases, but in reality, it's 10% negative to 90% positive—you would need to do sampling to address that difference."
Accuracy	Instruction Errors	Insufficient guidance	"If the initial manual annotation isn't handled with care, the system will fail to train accurately … One approach is to define crisp, clear criteria for the annotation process to reduce ambiguity."

Detailed Themes Development

	Systematic Mislabeling	Bias-driven	"You can sample your data to mitigate bias, but that can also cause unintended issues … blindly removing bias is possible but may not yield the results you want."
Consistency	Guideline Issues	Ambiguous instructions	"A good annotation tool should have a pleasant interface, easy-to-follow guidelines, and be hard to misinterpret. This would speed up the process and reduce frustration … If the guidelines are flaky and lead to too many discrepancies, you won't get a good model in the end."
	Guideline Issues	Guideline evolution	"Annotation processes should be iterative—like software development. Take feedback, improve the guidelines, add edge cases, and ensure annotators follow the latest versions of the guidelines."
	Annotator Context Gap	Lack of purpose knowledge	"For edge cases, you'd need someone with domain expertise to explain why a car behaves a certain way. So, lack of communication and lack of domain expertise are two issues I've seen."
	Team Coordination Gap	Misaligned hand-offs	"If a team does annotation internally instead of outsourcing, they might not document it properly. Instead, they'll convey details in meetings, and that information can be very ephemeral—easily forgotten."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID12**
Completeness	Scenario Coverage	Selection bias	"The data needs to be collected in diversified situations to ensure good coverage of all conditions, especially within the operational design domain (ODD) for autonomous driving."
	Scenario Coverage	Requirement-driven slices	"Through my analysis, I might identify gaps in the data—certain motorcycle behaviors or interactions that I've observed or imagined but don't yet have in the dataset. In that case, I would specify requirements for the data collection team: collect more data specifically on motorcycle behavior in certain driving situations."
	Unexpected Cases	Edge/unforeseen scenarios	"The guidelines should cover all the known scenarios that have been analyzed—those that are reasonably foreseeable. At the same time, we should acknowledge the existence of unknown scenarios, particularly edge cases. The challenge with unknown scenarios is that we cannot write explicit guidelines for them because they are unknown."
Accuracy	Boundary Errors	Bounding box errors	"When annotating objects like vehicles in an image using bounding boxes, it's important to avoid optimistic annotations. You can annotate slightly larger bounding boxes to be cautious, but you must avoid annotating smaller boxes. Underestimating the size could cause issues with decision-making and planning in autonomous systems—like collision avoidance or distance estimation."
	Instruction Errors	Insufficient guidance	"For annotation, if the guidelines are not written clearly, the quality can drop significantly. Annotation guidelines are critical … depending on how skilled annotators are, if they are new, they might just follow what the guidelines say without fully understanding them."
	Human Factors	Psychological strain / Fatigue	"The shortcomings lie with the people. I read an article before that said when people are doing annotations continuously, they tend to develop psychological issues. This can dramatically decrease the quality of annotated data."
Consistency	Human Variability	Intra-annotator drift / Mood variability	"Annotation consistency is tricky because it depends on factors like the annotator's mood, stress level, or the amount of data they need to annotate."
	Guideline Issues	Ambiguous instructions	"For annotation, if the guidelines are not written clearly, the quality can drop significantly. Annotation guidelines are critical."
	Guideline Issues	Guideline evolution	"Handling evolving annotation requirements is a challenge, especially when the ODD expands or changes … Even within the same ODD, if you discover new edge cases, those requirements would also change. The main challenge is making this process structured and traceable."
	Annotator Context Gap	Lack of purpose knowledge	"Because depending on how skilled annotators are, if they are new, they might just follow what the guidelines say without fully understanding them. They don't have experience, so they don't know how to make good use of the guidelines or apply their own expertise."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID13**
Completeness	Policy Ambiguity	Privacy/compliance omissions	"One thing that might be under quality assurance is the need for policies regarding masking sensitive data. For example, policies to blur pedestrians' faces or license plates could be considered part of quality assurance."
	Scenario Coverage	Requirement-driven slices	"The data must contribute to high-level requirements, such as safety goals. For example, if there's a requirement for data involving pedestrians wearing red shirts, the annotated dataset must include such cases."
	Unexpected Cases	Edge/unforeseen scenarios	"If harsh braking occurs, it can trigger data collection, capturing instances before and after that event. I call this 'golden data' because it can lead to detecting unknown scenarios or edge cases, which are crucial for improving machine learning software or for verification and validation."
Accuracy	Boundary Errors	Bounding box errors	"For image annotation, I see bounding boxes—that's clear … when annotating across frames, the bounding box may grow, shrink, or change dimensions … ideally, annotation quality should ensure smoother, consistent variations."

Detailed Themes Development

	Instruction Errors	Insufficient guidance	"Annotation guidelines: Ensure they are clear, detailed, and leave no room for ambiguity … communication is key—having an effective channel for annotators to ask questions and clarify guidelines."
Consistency	Guideline Issues	Ambiguous instructions	"Annotation guidelines: Ensure they are clear, detailed, and leave no room for ambiguity … Accurate and consistent data points depend on following a clear roadmap and standard."
	Guideline Issues	Guideline evolution	"Handling evolving data requirements is a challenge, especially when the ODD expands or changes … Even within the same ODD, if you discover new edge cases, those requirements would also change. The main challenge is making this process structured and traceable."
	Tool-driven Variability	Cross-modality alignment	"Some sensors capture data momentarily, while others spread data collection over time … you need calibration to make sure they align."
	Standardization Issues	Lack of frameworks	"I suppose … there should be standards for it, I mean for how annotations should be done. Even though there are multiple standards, at least you don't have to start from scratch."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID14**
Completeness	Scenario Coverage	Selection bias	"For example, if we look at automotive perception systems, you need data from different driving scenarios, different road conditions, different weather conditions (snow, rain, fog, sunlight), and different lighting conditions (daylight, dawn, dusk, nighttime). These are critical 'checkpoints' to ensure the data covers all possible scenarios."
	Missing Annotations	Missed objects	"Sometimes you might see corrupted annotations. For instance, there can be negative values in bounding boxes that can't translate properly to an image. Imagine you have 3,000–4,000 annotations, and 200–300 of them are corrupted. The model's performance after training for 100 epochs would be lower."
Accuracy	Boundary Errors	Bounding box errors	"If you have inconsistencies or errors … Sometimes you might see corrupted annotations. For instance, there can be negative values in bounding boxes that can't translate properly to an image."
	Instruction Errors	Insufficient guidance	"One key aspect is ensuring there are no erroneous or corrupted annotations in the dataset used for training. However, I don't have experience assessing annotation quality directly, so I'm not certain how it's ensured."
Consistency	Guideline Issues	Guideline evolution	"Since ODDs evolve continuously, we deal with constantly changing requirements. You cannot stop developing the systems. You keep retraining the models with more data so they understand additional scenarios and achieve higher accuracy over time. It's challenging but expected."
	Team Coordination Gap	Misaligned hand-offs	"What I believe happens is that sometimes requirements might get missed. When that occurs, the team must go back to the annotation team and say, 'We are not fulfilling this requirement; we need these specific features or annotations to address it.' This back-and-forth ensures that everything aligns, even if some aspects are initially overlooked."
	QA Capacity Constraints	Limited review & logging	"If 5–10% of your dataset annotations are corrupted or incorrect, it impacts the performance … To handle this, you need to either remove the images with corrupted annotations, or fix the annotation files to ensure consistency."
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID15**
Completeness	Scenario Coverage	Selection bias	"The completeness of this process is an important question. One thing that stands out to me is that the process begins with data collection, but I believe it should start with identifying data needs before collection begins … Data collection can be costly, so having a clear goal beforehand can be beneficial."
	Guideline Gaps	Missing feedback loop	"Yes, iterations should always be present in such a process … Quality assurance, which is crucial, should not be left until the end but rather incorporated throughout the process. If you only check data quality at the final stage, you might have already lost a lot of time."
	Policy Ambiguity	Privacy/compliance omissions	"Ensuring data privacy and compliance with ethical standards is particularly critical when dealing with sensitive data … The best way to handle this is to involve legal experts to ensure compliance with privacy regulations and ethical standards."
Accuracy	Instruction Errors	Insufficient guidance	"Annotation quality depends on multiple factors, such as … the cost implications of false positives and false negatives. These considerations are highly context-dependent."
	Systematic Mislabeling	Bias-driven	"Fairness is more related to data collection than annotation itself … If adjustments are made to data collection to mitigate bias, this should be reflected in the annotation process as well."
Consistency	Human Variability	Inter-annotator disagreement	"This is one of the biggest challenges. Even with well-defined guidelines, annotators may interpret cases differently based on context … Consistency can be improved by providing clear and detailed annotation guidelines, conducting regular training sessions, and using iterative refinement."

Detailed Themes Development

	Guideline Issues	Ambiguous instructions	“How do we ensure that annotation guidelines are well-defined and useful? ... Annotation quality depends on multiple factors ... Annotation guidelines must be fit for purpose and application-specific.”
	Guideline Issues	Guideline evolution	“When discussing evolving data annotation requirements ... This means that annotation efforts must be continuously revisited. The process is never truly complete. Continuous monitoring is essential.”
	Tool-driven Variability	Automation effects	“One challenge is that annotation work can feel isolating ... It is crucial for annotation tools to foster a sense of collaboration and purpose ... Some existing tools incorporate gamification elements, such as progress bars and small feedback indicators, but this could be expanded further.”
	Standardization Issues	Lack of frameworks	“Yes, I strongly support standardization ... A standardized process that captures best practices from across the industry would add significant value.”
	Human Factors	Attention & motivation gaps	“One challenge is that annotation work can feel isolating. Annotators often work alone on assigned batches, which can lead to a lack of motivation ... Gamification should encourage teamwork and shared progress.”
	Annotator Context Gap	Training/onboarding gaps	“Annotation requirements should be integrated into various aspects, including training materials for annotators ... Explicit documentation and traceability are essential.”
	QA Capacity Constraints	Limited review & logging	“Some key limitations include ... the lack of standardized metrics for annotation quality ... High costs and resource demands associated with independent verification.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID16**
Completeness	Incomplete Attributes	Attribute omission	“I think the most specific requirement for data annotation is to measure the ambiguity. Know the amount of ambiguity as opposed to what you’re supposed to answer. So you always should know if it’s possible to even answer the question.”
	Pre-annotation Issues	Automation-first omissions from foundation models	“In the future, I don’t think geometry annotations will be done manually anymore. Instead, the focus will shift to correcting those annotations, which will represent about 10% of the data ... Over-reliance on foundation models without targeted human correction can create omissions.”
	Guideline Gaps	Missing feedback loop	“Normally, you have this loop. You train the model, realize the model cannot learn from that data, and then you update your label guide. This loop has to be fast—happening maybe once a month, not once a year or once a project.”
	Unexpected Cases	Edge/unforeseen scenarios	“For traffic signs, you might say, ‘Draw a box around anything that looks like a sign.’ ... This approach avoids false negatives, which is a super important topic. It also allows you to start mining edge cases.”
	Sensor/Data Gaps	Calibration/synchronization issues	“For cleaning, you might focus on removing incorrect frames, repairing calibration issues ... Balancing helps ensure a dataset is representative, complete, and consistent.”
Accuracy	Labeling Errors	Hallucinated/ambiguous objects	“A typical error is deciding if an object is still identifiable or too occluded. Annotators often label ‘hallucinated’ objects—things that look like a pedestrian but are not. This confuses the model because the training process amplifies these errors.”
	Boundary Errors	Occlusion policy bias	“A typical error is deciding if an object is too occluded ... Annotators guess the extents, which leads to systematic bias.”
	Instruction Errors	Insufficient guidance	“The challenge is that it’s impossible to define unambiguous annotation requirements ... The real challenge is that you somehow have to embrace the ambiguity and document it.”
	Systematic Mislabeling	Incentive-driven vendor shortcuts/noise	“Later phases: many label providers have an incentive to reduce costs ... To save money, they may encourage annotators to be inconsistent but fast, or even have some annotators click randomly.”
	Pseudo-label Bias	Automation over-trust	“Automation-dominant. But the problem is, you can’t fully trust an AI to build another AI. If the process is fully automated, it will be biased. You still need human oversight.”
Consistency	Human Variability	Mental models diverge	“The label guide creates confusion, leading to random, biased answers from annotators. Each annotator ends up building their own mental model of what they’re supposed to do.”
	Guideline Issues	Ambiguous instructions	“The challenge is that it’s impossible to define unambiguous annotation requirements ... For example, in a demonstration with thousands of people, you can’t draw thousands of boxes ... You come up with a lot of weird rules, which are all ambiguous.”
	Guideline Issues	Guideline evolution	“Everything is evolving all the time ... For example, is a group two people or three? This is how things evolve all the time, and you always have that in every annotation project I’ve seen.”
	Tool-driven Variability	Process scales better when UI encodes step-by-step guide	“So the only way to do that, in my opinion ... is to enforce it in software somehow. Not train and retrain people ... You need software-defined processes so the annotator doesn’t need training, and then you can scale it.”
	Standardization Issues	Immature/partial standards (ISO 5259-2)	“Standardization: For example, the European AI community now has ISO standard 5259-2, which defines which metrics should be used. However, it doesn’t define how these metrics should be implemented.”

Detailed Themes Development

	Team Coordination Gap	Intent/competence mismatch	“There could also be a mismatch of competence—you just don’t understand what it’s about. This isn’t limited to the annotator and the machine learning engineer. Even the ML engineer might not understand the annotator's needs … This creates a lot of friction.”
	QA Capacity Constraints	Vendor ‘100% quality’ claims statistically impossible	“Some companies even say, ‘Our quality is 100% when we deliver it,’ but obviously, this is a statistical impossibility … Currently, QA processes in vendors are fundamentally flawed.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID17**
Completeness	Guideline Gaps	Missing feedback loop	“Once the first data has been annotated … you realize there’s something missing or the annotations are not precise enough … then you go back. Exploitation step seems to be missing to me.”
	Sensor/Data Gaps	Calibration/synchronization issues	“One of the problems that we encountered … was that the time stamps don’t fit. It’s all things with sampling frequencies and synchronized clocks … this came up after the fact.”
	Sensor/Data Gaps	Sampling limits / throughput	“There was something with the sampling rate and the bandwidth … we had planned to do 30 frames per second and then it was less than that because the hardware couldn’t handle the three streams.”
	Scenario Coverage	Selection bias	“We didn’t collect in the rain … we couldn’t pick the day. That definitely has an impact.”
	Scenario Coverage	Legal/logistical constraints	“You can’t just fly a drone over an autobahn … you’re not allowed to, and with good reason. You can also not block the autobahn because you want to fly a drone.”
	Scenario Coverage	Requirement-driven slices	“One of the requirements was … we wanted to do it at nightfall … include scenarios that you don’t find in published data sets as well.”
Accuracy	Schema Accuracy	Granularity mismatch	“We used a taxonomy … but it didn’t include trams, which we see in our European street scene. The taxonomy was from the US … so this was a problem.”
	Instruction Errors	Insufficient guidance	“Annotation quality depends on linking the system purpose to the guideline rules … the quality depends on how well these rules reflect what the system is supposed to achieve.”
	Pseudo-label Bias	Automation validity risk	“As soon as you introduce automation, you have a problem of validity … if the automated system annotates your data and it’s good enough, why not just use it in practice? You still need validation and verification.”
Consistency	Guideline Issues	Guideline evolution	“Once you start talking to the authorities … you get new requirements that you didn’t think of … sampling rate and bandwidth also led to new requirements as you go.”
	Tool-driven Variability	Cross-modality alignment	“We wanted three different modalities synchronized … but the synchronization didn’t work properly. The modalities are individually good, but not aligned.”
	QA Capacity Constraints	Limited review & logging	“Essentially we do spot checks … that’s all you can do. Doing inter-annotator agreement is prohibitively expensive.”
Dimension	**Category**	**Sub-Type (Code)**	**Quote from ID18**
Completeness	Guideline Gaps	Missing feedback loop	“We had a different labeling company where we didn’t have these feedback loops in that intensive way, which definitely led to some misunderstandings.”
	Policy Ambiguity	Privacy/compliance omissions	“You have to provide a data privacy concept already … In preprocessing you have to make sure that no sensitive data go to the data annotation company … GDPR is very difficult when dealing with countries outside of Europe.”
	Policy/Method Gaps	Annotation type coverage gaps	“What I would be missing here is panoptic segmentation … And I think the map probably is missing … sometimes you also need to annotate the map.”
	Unexpected Cases	Edge/unforeseen scenarios	“Even though if you have examples then there will always be something you are still unsure about … you will probably never cover all edge cases in your guidelines.”
	Sensor/Data Gaps	Calibration/synchronization issues	“If you screw up something in the data collection, everything that happens after that is affected … slight miscalibrations can have a very huge effect on the quality.”
	Scenario Coverage	Selection bias	“Usually most of the data sets only capture daytime or sunny conditions … models later have problems with identifying pedestrians in rainy conditions or at night.”
Accuracy	Labeling Errors	Wrong class label	“For example, in autonomous driving … you have different classes of vehicles. And you say, OK, this is a car, this is a truck, this is a van … different people might annotate the same object differently.”
	Boundary Errors	Overlap/geometry conflicts	“If the annotated bounding box occludes with each other … or dimension/velocity conflicts … then you could sense the annotation quality is low.”

Detailed Themes Development

	Schema Accuracy	Granularity mismatch	“When talking about dimensions of objects … in 2D images you don’t have 3D information … two annotators would annotate the same object differently even with a clear definition.”
	Instruction Errors	Insufficient guidance	“Sometimes it’s very hard to make a guideline that can constantly measure it … annotators will interpret it in different ways.”
	Systematic Mislabeling	Bias-driven / Benchmark contamination	“We tested RGB vs infrared vs fusion … the poor model predictions fit better to the poor annotation quality … This is a dataset that is heavily used by everyone but actually has not good labels.”
	Pseudo-label Bias	Automation validity risk	“As soon as you introduce automation … using model predictions as annotation benchmarks risks validity.”
	Temporal Accuracy	Tracking errors	“If you are processing sequence data … you want tracking ID and accurate positions … Otherwise, IDs across frames may be inconsistent.”
Consistency	Human Variability	Inter-annotator disagreement	“Different annotators … they will subjectively annotate the object depending on distance or shape … then we need a second person to correct this.”
	Guideline Issues	Ambiguous instructions	“It’s very hard to include all edge cases … annotators still behave differently … sometimes misunderstanding the guideline.”
	Guideline Issues	Guideline evolution	“New requirements emerge when you get data back in batches … then you realize this is not what we wanted and it requires re-annotation.”
	Tool-driven Variability	Cross-modality alignment	“We projected annotations from point cloud onto the image plane and checked alignment … sometimes the projection didn’t capture the object correctly.”
	Standardization Issues	Lack of frameworks	“There is no common standard … every company follows their own guideline … sometimes datasets lack information about how annotations were done.”
	QA Capacity Constraints	Limited review & logging	“Annotation companies rely on manual reviewing … they said they want 95% accuracy, but it was all done manually … automated QA was lacking.”