Reasoning-guided Ego-path Segmentation for Autonomous Trains using Vision-language Models

Mohammadjavad Ghorbanalivakili[1], Ashley Varghese[1], Gunho Sohn[1]

[1] Dept. of Earth and Space Science and Engineering, York University, Toronto, Ontario, Canada - (mvakili, ashleyva, gsohn)@yorku.ca

Keywords: Vision–language Models, Reasoning Segmentation, Autonomous Trains, Ego-path Detection, Rail Switch Understanding.

Abstract

Autonomous train perception must identify the train's valid path under complex railway geometry, particularly at merging and diverging switches where multiple candidate tracks coexist. Existing approaches are primarily trained as purely visual predictors and typically do not provide justification for route selection, despite the fact that valid paths depend on structured cues such as blade–stock contact, rail gaps, and track continuity. In this work, we adapt the Large Language Instructed Segmentation Assistant (LISA) to railway ego-path perception and formulate the task as reasoning-guided segmentation: given a forward-facing railway image and a natural-language query, the model predicts the valid ego-path mask and, when prompted, generates a textual explanation grounded in visible switch geometry. Our approach integrates railway-specific prompting, a tailored annotation scheme, and efficient finetuning, along with semantic segmentation supervision to support general scene understanding. Experiments on a RailSem19-based evaluation set show improved ego-path segmentation performance over the original LISA checkpoint and increased robustness to prompt variation, while qualitative results indicate that the model can produce plausible, though not always consistent, reasoning. Notably, these capabilities emerge despite the reasoning-specific dataset consisting of only 54 samples, highlighting the data efficiency of the approach. These results highlight the potential of vision-language models for more interpretable railway perception, while also underscoring the need for stronger supervision and evaluation in safety-critical settings. Code and reasoning segmentation data are available at `https://github.com/mvakili96/Railway_Perception_FoundationModel`.

1. Introduction

Reliable ego-path perception is a prerequisite for next-generation autonomous trains. While highly automated train systems can already operate in constrained settings, extending autonomy to more open and infrastructure-light railway environments requires onboard perception to determine which track the train can safely follow (Ghorbanalivakili and Sohn, 2025). This becomes particularly difficult in switch regions, where several rails may be visible simultaneously but only one route is physically continuous for the train. In these scenes, perception is not only a segmentation problem. It is also a reasoning problem over track geometry.

In this study, we define a *track* as an instance composed of a left rail and a right rail. *Track bed* is the region bounded by a track's left and right rails. Railway *path* or *route* is defined as a track with its corresponding track bed. A *switch region* is a railway section enabling transitions between paths. A *turnout* branches one path into another, while a *merge* combines two paths into one.

Most existing railway vision pipelines treat track extraction or ego-path detection as a fully supervised image understanding task. They typically predict rails, track centers, or ego-track regions from monocular or multi-camera images, and then rely on geometric constraints, post-processing, or task-specific decoders to recover the valid route (Jaß and Thomas, 2025, Guo et al., 2026). These methods have shown encouraging results on public rail benchmarks, especially after the introduction of RailSem19 as a rail domain scene understanding dataset (Zendel et al., 2019), dedicated route extraction pipelines such as TPE-Net (Ghorbanalivakili et al., 2023), and more recent end-to-end ego-path detectors such as TEP-Net (Laurent, 2024). However, they usually optimize only for visual output quality, not for explainability. As a result, they provide limited support for debugging, operator trust, or downstream safety assurance when the scene becomes ambiguous.

Recent advances in multimodal learning point toward a fundamentally different paradigm for visual understanding. Instead of treating vision tasks as fixed, narrowly defined prediction problems, emerging approaches leverage instruction-following models that interpret images through flexible, language-guided interactions. Therefore, such systems demonstrate strong generalization to previously unseen tasks and inputs (Liu et al., 2023). In parallel, prompt-driven segmentation frameworks have shown that a single model can generate meaningful object masks across diverse image distributions without task-specific retraining (Kirillov et al., 2023). Building on these ideas, recent approaches integrate high-level reasoning with dense prediction, enabling models to infer implicit intent from natural language (Lai et al., 2024).

The railway switch domain is a natural fit for this paradigm. Determining the valid route at a switch depends on interpretable structural cues: which switch blade (moving rail in a switch) is closed against its stock rail (fixed rail in a switch), which branch retains rail continuity, and which alternative path is blocked by an open gap between the rails. These cues are already described in railway operating logic and can be verbalized by a model. This creates an opportunity to go beyond silent mask prediction and instead train a system that can explain why the segmented route is correct.

In this work, we adapt LISA (Lai et al., 2024) to the task of

railway ego-path segmentation in switch scenes and extend the supervision with explicit railway explanations. The core idea is simple: instead of asking only for the track or other rail infrustructure intances, we ask for the train's valid path and supervise the model with natural-language reasoning tied to blade and stock positions, switch topology, and route continuity. This allows the model to behave as a reasoning-guided segmentation assistant for autonomous trains.

1.1 Contributions

- We formulate railway switch understanding as a reasoning-guided ego-path segmentation task, where the model predicts the valid route and can explain the underlying geometric evidence.
- We construct a railway-specific adaptation for finetuning LISA, including switch-aware prompt templates, polygon ego-path masks, and explanatory supervision aligned with turnout and merge logic.
- We establish an initial experimental protocol on RailSem19 benchmark, showing improved segmentation performance and prompt robustness over the base model, while providing qualitative insights into the strengths and limitations of the current version.

2. Related Work

2.1 Scene Understanding and Ego-path Perception

Railway computer vision has historically lagged behind road-scene perception because of the limited availability of public datasets and the domain-specific structure of rail environments (Olivier et al., 2025). RailSem19 partially addressed this gap by introducing the first widely used public benchmark for semantic rail scene understanding, with ego-perspective train and tram imagery and rail-relevant labels (Zendel et al., 2019). More recent work has pushed from semantic understanding toward route-level perception. TPE-Net extracts left-right rail pixels and associates them into route instances for rail path proposal generation, explicitly targeting topological interpretation in switch and multi-track scenes (Ghorbanalivakili et al., 2023). Laurent (Laurent, 2024) proposed the dedicated task of train ego-path detection and introduced TEP-Net, an end-to-end deep learning approach tailored to identifying the train's immediate path on railway tracks. Recent data-centric efforts such as Labels4Rails emphasize that annotation scarcity remains a major bottleneck in autonomous rail perception and that switch-aware rail labeling still needs better tooling and datasets (Hiebert et al., 2025). Another research (Jaß and Thomas, 2025) also highlights the safety motivation for robust railway track perception by studying architecture redundancy for rail segmentation.

These methods collectively show that railway ego-path detection is feasible, but they also reveal a persistent limitation: the dominant learning objective is still visual accuracy alone. Even when the downstream logic depends on switch structure, the model output does not expose the reasoning behind its decision. Our work targets this missing explanatory layer.

2.2 Vision-language Foundation Models for Grounded Perception

Large multimodal models have rapidly improved visual reasoning and instruction following. LLaVA established a strong recipe for visual instruction tuning by connecting a vision encoder to a large language model and training on language-image instruction data (Liu et al., 2023). In parallel, SAM introduced a promptable segmentation foundation model with strong zero-shot transfer across segmentation tasks (Kirillov et al., 2023). LISA unified these lines by augmenting a multimodal large language model with a dedicated `[SEG]` token whose hidden representation is decoded into a segmentation mask via SAM, thereby enabling reasoning segmentation.

This line has already begun to diversify. LISA++ improves the baseline reasoning-segmentation pipeline with stronger training data and visual chain-of-thought supervision (Yang et al., 2024). GSVA extends segmentation-oriented multimodal models to handle multiple or null referents through additional output tokens (Xia et al., 2024). See, Say, and Segment studies false-premise queries and shows that segmentation-capable multimodal models should not only segment but also reject invalid requests and respond helpfully in language (Wu et al., 2024). These works make it clear that segmentation-oriented multimodal language models are moving toward richer interaction and more reliable reasoning.

However, the existing literature is still centered on generic object grounding benchmarks such as referring expression segmentation and reasoning segmentation. To the best of our knowledge, no prior work has adapted reasoning segmentation to railway switch interpretation, where the target is not a generic object category but the physically valid train route, and where the explanation should explicitly refer to rail continuity, blade closure, and switch topology. This is the gap our work addresses.

3. Methodology

3.1 Task Formulation

We define reasoning-guided railway ego-path segmentation as follows. Given a forward-looking RGB image $I \in R^{H\times W\times 3}$ captured from a train-mounted camera and a natural-language query Q, the model predicts a binary mask $M \in R^{H\times W}$ corresponding to the train's valid forward route. When the query requests justification, the model additionally outputs a free-form explanation A grounded in the visible switch geometry. Unlike ordinary foreground segmentation, the target is not merely “rails” or “track.” It is the track bed of the route that remains physically valid for the train after considering switch state.

This task is especially relevant in switch scenes where multiple candidate paths are visible. In a switch region, the explanation must identify which blade is closed and which branch remains continuous. Moreover, the explanation must identify which upstream path remains connected into the downstream mainline.

3.2 LISA Background

Our method builds on LISA, which combines a LLaVA-style multimodal language backbone with the promptable mask generation capability of SAM. As illustrated in Fig. 1, LISA turns segmentation into a language-conditioned generation problem: the model reads an image-text conversation, decides whether the user is asking for a grounded mask, and, if so, emits a special token `[SEG]` that triggers the segmentation branch.

The `[SEG]` token therefore has two roles. First, it is a mask carrier. During generation, the hidden state at the position of

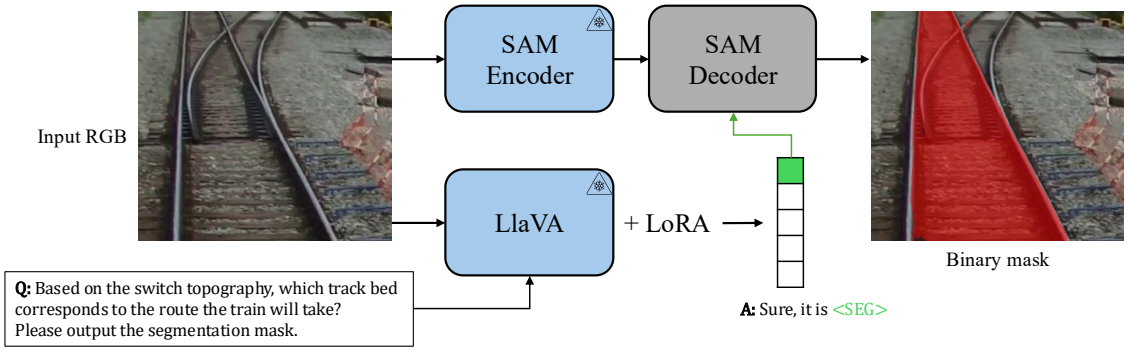

Figure 1. LISA architecture for reasoning segmentation (Lai et al., 2024). SAM encoder (Kirillov et al., 2023) and the vision backbone of LlaVA (Liu et al., 2023) are frozen throughout finetuning. Yet, LoRA (Hu et al., 2021) is applied to the attention projections of the language model, allowing the multimodal backbone to adapt its reasoning behavior while keeping most pretrained vision and fusion weights frozen.

[SEG] is passed through a learned projection module and converted into a text-conditioned prompt embedding for SAM's prompt encoder and mask decoder. This is the mechanism by which the language model specifies what should be segmented. Second, [SEG] acts as a mask-generation gate. The model only produces a segmentation mask when it actually emits this token in its textual response. If the token is absent, the system remains in pure language mode and no mask prompt embedding is extracted.

This gating behavior is important because original LISA is trained on a mixture of tasks rather than on segmentation alone. The original LISA model is jointly optimized on semantic segmentation datasets, referring segmentation datasets, visual question answering data (VQA), and the reasoning segmentation set. The reasoning segmentation portion teaches the model to handle implicit queries, world knowledge, explanatory answers, and reasoning-heavy segmentation requests using only 239 sample images. Seen from this perspective, [SEG] becomes the interface that lets one decoder support both “answer in language only” and “answer in language plus mask” behaviors within a single conversational model.

Our rail-specific finetuning does not incorporate VQA stream, and most railway prompts explicitly request ego-path segmentation. As a result, the model spends much more of its capacity on learning which route to mask than on deciding whether a mask should be produced at all. Still, the architectural mechanism remains intact, and it is important for understanding how the original model can alternate between text-only and mask-producing responses within the same dialogue interface.

3.3 Railway-specific Dataset Construction

To adapt LISA to railway ego-path reasoning, we assemble a specialized rail switch dataset from forward-looking railway imagery. Our reasoning segmentation dataset is a small pilot dataset derived from switch region frames of RailSem19. These frames include both close-up views that focus specifically on the switch area and wide-view, original images from RailSem19 that capture one or multiple switches. Note that the resolution of the cropped, close-up views are variable depending on the visibility of the switch region.

The current version of our reasoning segmentation repository contains 54 annotated training images. These images were extracted from the first 1,000 samples of RailSem19 with human

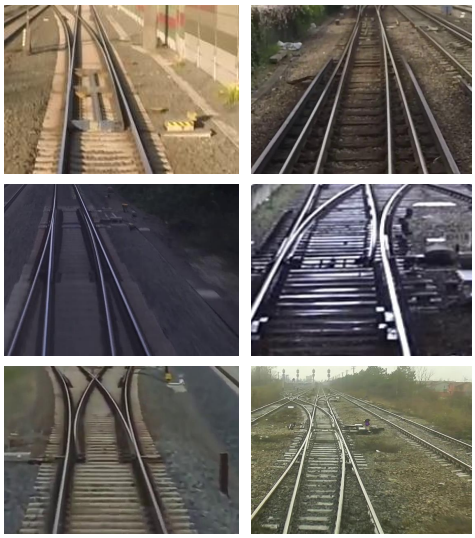

Figure 2. Sample raw images of the proposed rail reasoning segmentation dataset. These images are cropped from RailSem19 dataset (Zendel et al., 2019).

supervision. The annotated cases cover both turnout and merging geometries. Figure 2 depicts six sample images of our proposed reasoning dataset for switch region understanding. In the present pilot split, the prompts describe left-branch merges, right-branch merges, right-route turnout selections, left-route turnout selections, and straight-through continuity cases. This structure is important because it teaches the model that the correct segmentation depends on route logic rather than on a fixed visual template.

Each image-level annotation contains six segmentation-oriented prompt paraphrases (Input Query Prompts in Fig. 3). These paraphrases are useful because they vary wording while preserving the same route logic, reducing overfitting to a single question form. We further extend the dataset with explanatory supervision. For each image, we provide a query asking why the segmented route is correct and an answer describing cues such as blade contact, visible rail gaps, or rail continuity (Explanatory answer text in Fig. 3). This supervision mirrors the explanatory mode originally supported by LISA, but it is specialized to railway switch semantics.

In addition to this railway reasoning segmentation, we utilize a

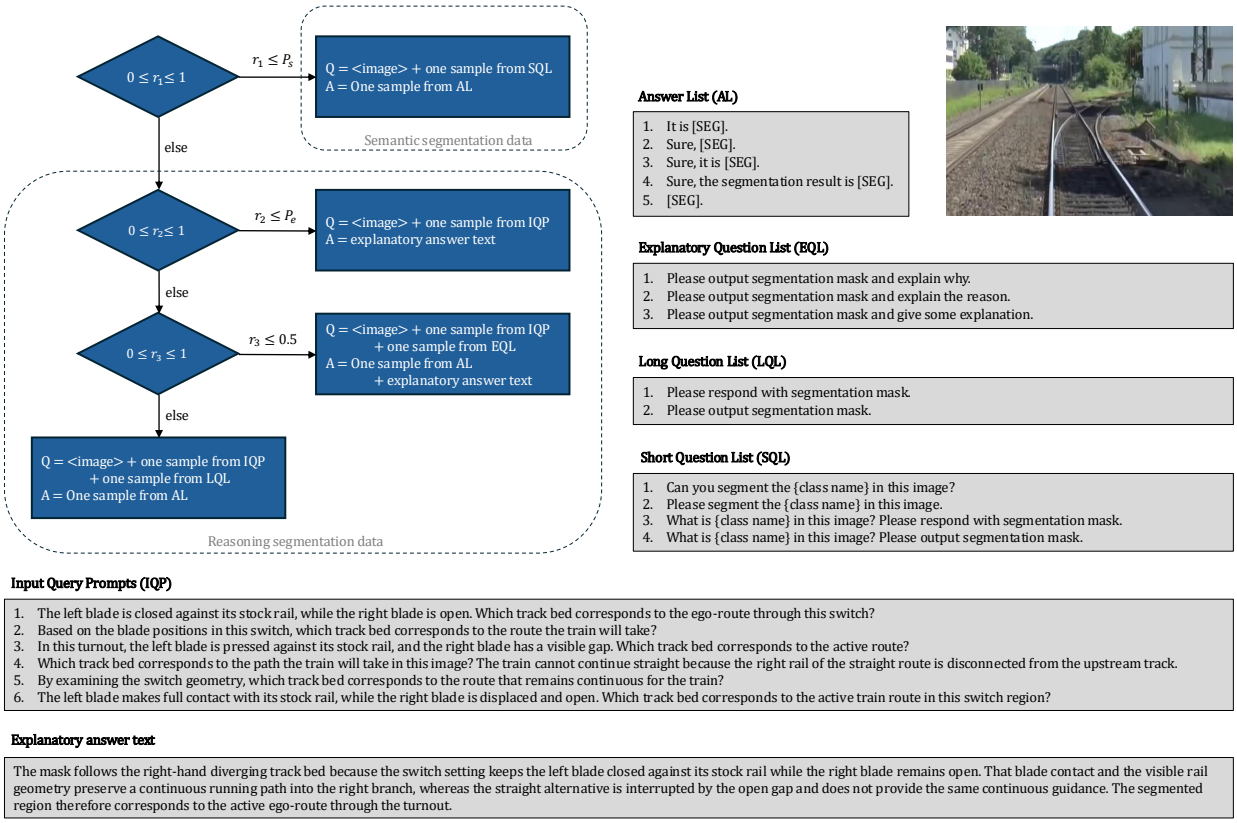

Figure 3. Flowchart for multimodal finetuning with LLaVA (Liu et al., 2023). For the sample image at the top right, the input query Q and ground-truth output answer A corresponding to semantic segmentation and reasoning-based segmentation tasks are illustrated. Random variables r_1–r_3 are used to select the training objective, after which an image is sampled from the corresponding dataset. The probabilities P_s and P_e control the likelihood of semantic segmentation and explanatory reasoning (without mask prediction), respectively.

semantic segmentation stream built from the first 6,000 original RailSem19 images. These images have a resolution of 1024×1024 and are cropped from the RailSem19 dataset. The reason for adding this extra supervision is that reasoning segmentation alone is too small and too switch-specific to teach the model robust rail context from scratch. Before the network can explain why a turnout is active or why a merge remains continuous, it must first learn the visual meaning of the basic railway classes themselves. We therefore used original labels of RailSem19 to provide dense class supervision for three scene elements: *rail*, *track bed*, and *background*.

An important implementation detail is that this semantic branch is still trained in LISA's language-conditioned format rather than as a classical multi-class pixel classifier. Although the RailSem19-derived annotations are organized as a three-class semantic labeling problem, the model does not predict a single multi-class mask with one softmax over all categories. Instead, the loader samples a class name, poses a prompt such as a standard segmentation question for that class, and converts the label map into a binary one-vs-all target for the queried class. This design is especially useful for our setting because it aligns the semantic pretraining stage with the later reasoning stage: both are expressed as text-conditioned binary mask prediction. As a result, the model learns the contextual vocabulary of railway scenes through dense semantic supervision while being asked to solve the harder reasoning problem of selecting the valid ego-route among several candidate paths in switch regions.

Figure 3 illustrates the input query Q and the ground-truth output answer A for our selected multimodal large language model (LLaVA) under both semantic segmentation and reasoning-based segmentation settings. The variables r_1–r_3 are randomly generated values used to determine the finetuning objective. Based on this selection, an image is then sampled from the corresponding dataset, as the semantic and reasoning segmentation tasks are constructed from different image sets. Here, P_s denotes the probability of performing semantic segmentation, while P_e represents the probability of generating explanatory reasoning without producing an output mask. Explanatory reasoning without an output mask is useful because it trains the model not only to emit `[SEG]` when appropriate, but also to generate free-form reasoning in cases where the query emphasizes explanation. In effect, the model learns that the route selection and the verbal rationale are two aligned outputs of the same switch interpretation process.

3.4 Finetuning Strategy

Similarly to the original approach of LISA, we keep most of the large pretrained visual and image-language alignment components fixed because they already contain broad visual knowledge that is useful for railway scenes. We then focus learning on the parts of the model that matter most for our task: the language-model attention updates that control how the model reasons over the prompt, the mask-generation decoder that converts the textual decision into a spatial prediction, and the small bridge that translates language features into segmentation prompts.

This choice reflects a practical assumption: some generic

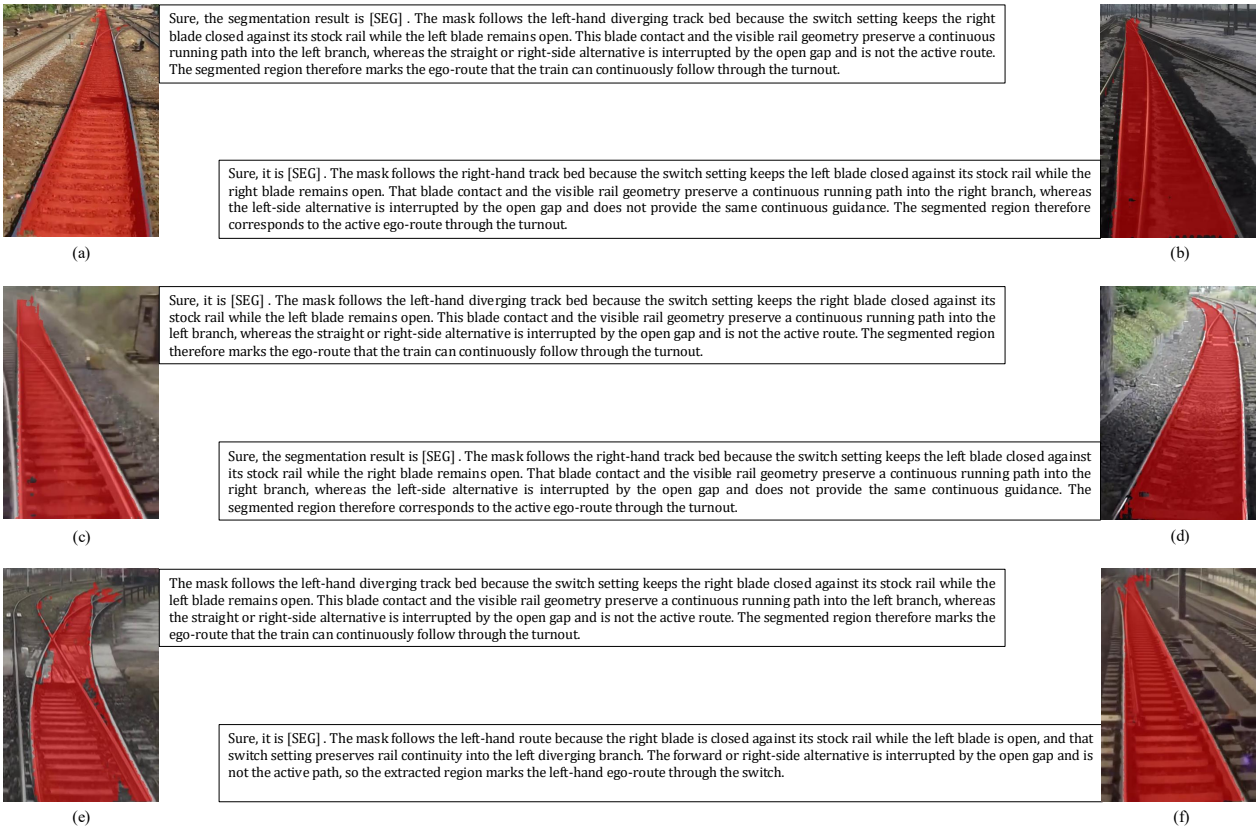

Figure 4. Qualitative results of our rail-finetuned LISA model on switch region samples using the query *“By examining rail continuity and switch geometry, segment the active ego-route and explain why.”* Each example shows the predicted mask overlay and explanation. (a)–(b) Successful cases. (c)–(d) Correct masks with inconsistent reasoning. (e)–(f) Mask failures despite partially correct explanations.

visual knowledge needed to recognize rails and scene context is already present in the pretrained backbone, while the critical missing capability lies in mapping railway-specific language cues to the correct segmentation target. LoRA adapters (Hu et al., 2021) are especially attractive here because railway finetuning data are limited, and full adaptation of all language-model weights would be unnecessarily expensive.

For optimization, the text decoder is supervised with an autoregressive language cross-entropy loss so that the model learns to produce the expected response format, including the appropriate use of `[SEG]` and, when enabled, the corresponding explanation text. In parallel, the predicted mask is supervised with two complementary pixel-level losses: a binary cross-entropy term and a Dice loss which encourages good overlap between the predicted and ground-truth masks. The final training objective is the weighted sum of the language loss and the two mask losses.

Our railway finetuning starts from the LISA-7B-v1 demo checkpoint released by the authors of LISA. This checkpoint is already trained on LISA’s original four-task mixture, namely semantic segmentation, referring segmentation, VQA, and reasoning segmentation. Here, the “7B” designation refers to the size of the underlying language backbone, i.e. the 7-billion-parameter LLaMA-based LLaVA model on which LISA’s smallest version is built.

In terms of training configuration, we finetuned the model in bfloat16 precision on our HPC cluster using 8 parallel NVIDIA L40 GPUs distributed over two compute nodes. We use a per-GPU batch size of 2 together with gradient accumulation over 10 steps, corresponding to an effective batch size of 160 samples per optimizer update. For data mixing, we use a sample-rate ratio of 45:1 between the semantic segmentation branch and the railway reasoning-segmentation branch ($P_s = 0.98$). In addition, the explanatory ratio is set 0.2, meaning that 20% of eligible railway reasoning samples are converted into text-only explanatory supervision during training ($P_e = 0.2$). The number of epochs here is set to 3, with 500 iterations per epoch. The reader is encouraged to visit our Github repository should they seek the remaining configurations.

3.5 Inference Workflow

At inference time, the user provides a raw railway image and a natural-language prompt such as: (i) Segment the train’s valid forward path in this image, (ii) Which route is active in this switch? Please output segmentation mask, and (iii) Please output the segmentation mask and explain why this is the valid ego-route.

The model generates a text response that may include `[SEG]`, and the hidden state associated with `[SEG]` is decoded into the predicted ego-path mask. If the prompt requests explanation, the language decoder also produces a textual rationale. In the railway setting, a desirable explanation should mention exactly the cues a human operator would inspect, such as which blade is pressed against its stock rail, which competing branch is interrupted by an open gap, or which approach connects continuously into the shared mainline.

	”By examining rail continuity and switch geometry, segment the active ego-route.”		*”Segment the track bed in this image.”*	
	CIoU	GIoU	CIoU	GIoU
LISA (Lai et al., 2024)	52.3	54.1	24.8	25.2
Rail-finetuned LISA	81.7	83.4	81.9	83.0

Table 1. Comparison of ego-path segmentation performance between the original LISA demo checkpoint and the rail-finetuned LISA under two prompt formulations, reported in terms of %CIoU and %GIoU.

4. Experimental Results

We used the final 2,500 images of the RailSem19 dataset for testing and validation. Ego-path ground-truth annotations are obtained from those provided by the TEP-Net authors. The testing and validation images are subsequently cropped to a region of interest (ROI), defined as a bounding box that encloses the ego-path with a predefined offset.

Figure 4 illustrates the mask prediction results of our rail-adapted LISA model given the query *“By examining rail continuity and switch geometry, segment the active ego-route and explain why.”* The figure presents six representative switch region samples, selected to highlight both the strengths and limitations of the model, thereby motivating directions for future improvement.

One notable observation from Fig. 4 is the presence of noisy and blobby masks across most samples, indicating potential room for improvement in semantic segmentation quality. Figures 4(a)–(b) demonstrate successful cases, with accurate ego-path mask predictions and coherent explanations that correctly contrast the selected route against alternative tracks. A minor discrepancy is observed in Fig. 4(b), where a merging switch is incorrectly described as a turnout.

Figures 4(c)–(d) present cases where the ego-path mask is correctly identified, but the accompanying explanations are inconsistent with the visual evidence. In these examples, the reasoning contradicts both the predicted mask and the underlying switch geometry, with errors in inferred train direction and blade–stock contact.

Figures 4(e)–(f) illustrate failure cases in mask prediction. Although the generated explanations are correct, the ego-route masks are either misclassified or insufficiently distinguished from competing track candidates.

Table 1 evaluates ego-route segmentation under two different prompt formulations. Specifically, the comparison contrasts a reasoning-oriented prompt *“By examining rail continuity and switch geometry, segment the active ego-route”* with a simpler prompt *“Segment the track bed in this image”* that is more instruction-style.

The results reveal a clear pattern of the original LISA demo checkpoint: its zero-shot performance is highly sensitive to prompt phrasing. While it achieves moderate performance under the reasoning-guided prompt (52.3 CIoU and 54.1 GIoU), it degrades significantly under the simpler prompt (24.8 CIoU and 25.2 GIoU).

In contrast, the rail-finetuned LISA demonstrates consistently strong performance across both prompts, achieving approximately 82–83% in both CIoU and GIoU regardless of prompt formulation. At first glance, this may appear as a limitation, since explicitly referencing the ego-route does not seem to affect performance. However, this behavior is likely due to the evaluation setup. The testing images consist of close-up switch regions that are largely dominated by track beds. Therefore, the dataset may not be sufficiently challenging to reveal the true impact of prompt variation on model performance.

Even before large-scale experimentation, the proposed formulation is useful for two reasons. First, it reframes railway ego-path perception from a silent geometry extraction problem into an explainable decision-making problem. Second, it provides a low-data path for specialization of a powerful pretrained multimodal system. This is particularly notable in our setting, where the reasoning capability is adapted using only 54 annotated samples and the smallest variant of LISA, yet still yields promising improvements. This is important because railway datasets are usually much smaller than road-scene benchmarks, and switch annotations are especially expensive to curate.

5. Future Work

Several limitations of the proposed approach should be acknowledged. These limitations form the basis for the future research directions outlined below.

First, the current training objective relies on conventional semantic segmentation losses, which do not explicitly enforce structural continuity or topological constraints of railway tracks. This limitation may contribute to the observed noisy and spatially diffuse mask predictions, particularly in complex switch regions. We leave the integration of such constraints as an important direction for future work.

Second, standard IoU-based metrics do not adequately report the correctness of ego-path predictions, particularly in scenarios involving overlapping track candidates. Due to substantial spatial overlap between alternative track beds, an incorrect route prediction in switch regions may still yield an IoU exceeding 50%, thus overstating model performance. This highlights the need for more task-specific evaluation metrics that directly measure route selection success rate. Developing such metrics constitutes an important direction for future work. However, existing railway datasets do not explicitly annotate switch regions, which presents an additional challenge. Addressing this gap by generating dedicated switch region labels for one or more public benchmarks is therefore another key direction for future research.

Third, we have not quantitatively evaluated the generated language explanations in terms of fluency, correctness of blade–stock interaction understanding, switch type recognition, rail continuity reasoning, train traversal prediction, or susceptibility to hallucination. Empirical observations indicate that the model can produce plausible explanations even when the corresponding mask predictions are incorrect. This discrepancy highlights

a gap between linguistic fluency and factual grounding. Therefore, future work should incorporate evaluation protocols that assess alignment between the generated text, visual evidence, and predicted outputs.

Another limitation of the current version is the absence of comprehensive ablation experiments to disentangle the contribution of each component in our training framework. In particular, we do not explicitly quantify the impact of incorporating semantic segmentation alongside reasoning-based supervision. Furthermore, it remains uncertain whether explanatory supervision improves route selection accuracy. The robustness of the model across different switch geometries, such as turnout versus merging configurations, is also not evaluated. Additionally, the effect of prompt design, i.e., the use of diverse paraphrased prompts versus a single, fixed formulation has not been explored. Conducting such ablation studies would guide more effective design choices in future work.

Moreover, direct comparison with existing state-of-the-art methods is not included in the current study. Existing approaches are typically trained and evaluated on conventional ego-path or track detection benchmarks, whereas our method relies on a newly constructed dataset with paired segmentation and reasoning supervision. As a result, a fair comparison is not straightforward at this stage. Therefore, establishing standardized evaluation protocols remains an important direction for future work.

Due to data scarcity, the model is not explicitly exposed to detailed rail network topology during training, including components such as switches, stock rails, switch blades, rail frogs, diamond crossings, and upstream/downstream path structure. Instead, it is required to infer these complex relationships indirectly through textual prompts. This places a substantial burden on the model’s reasoning capabilities and may limit its robustness.

Finally, in multi-track scenes where the image is captured beyond the switch region, the unique ego-path becomes inherently ambiguous once the blade–stock interaction is no longer visible. In such cases, the model is prone to hallucinating an ego-path mask, often accompanied by uninformative or inconsistent explanations, since these scenarios are not represented in the training data. This limitation suggests that incorporating temporal information from consecutive frames should be a critical direction for future work.

While railway ego-path extraction may appear structurally simple compared to general scene understanding, we intentionally built on a large vision-language foundation model to move beyond purely geometric or pixel-level reasoning. More importantly, this study represents the initial stage of a larger research agenda centered on autonomous railway safety. Reliable ego-path extraction enables defining an ROI aligned with the train’s physically valid trajectory. Building on this ROI, future work will extend toward context-aware behavior prediction and collision risk estimation, where the integration of visual perception with common-sense knowledge encoded in large language models may provide advantages in interpreting dynamic scenes and anticipating hazards. In this sense, the current work is not only about improving segmentation performance, but about exploring a scalable pathway toward more interpretable and context-aware railway autonomy systems.

6. Conclusion

This paper introduced reasoning-guided ego-path segmentation as a new formulation for switch region understanding in autonomous train perception. Rather than treating route estimation as a purely visual problem, we framed it as a joint task in which the model identifies the ego-route and, when prompted, provides a textual justification. To support this formulation, we adapted LISA vision-language foundation model to the railway domain using rail-specific finetuning data. Compared with the original LISA demo checkpoint, the rail-finetuned model achieved substantially stronger ego-path segmentation performance and showed robustness to prompt phrasing. Although the model can often produce accurate masks and plausible explanations, the generated reasoning is not always consistent with the visual evidence, and mask quality can still degrade in challenging scenes.

Overall, the study shows that vision-language foundation models can serve as a useful basis for more interpretable railway perception. However, the current work should be viewed as an initial step rather than a complete solution. Broader evaluation, stronger reasoning verification, more comprehensive ablations, and richer supervision over railway topology will be necessary to determine how reliably such models can support safety-critical rail applications.

References

Ghorbanalivakili, M., Kang, J., Sohn, G., Beach, D., Marin, V., 2023. TPE-Net: Track point extraction and association network for rail path proposal generation. *Proceedings of the IEEE 19th International Conference on Automation Science and Engineering (CASE)*.

Ghorbanalivakili, M., Sohn, G., 2025. TRIT-Net: Triplet-based railway instance tracing network using attraction field representation. *Proceedings of the 22nd Conference on Robots and Vision (CRV)*.

Guo, H., Wei, X., Ji, Y., Ge, C., Tang, Q., Cai, K., 2026. Track2Net: A fast lightweight model with keypoint alignment and track anchors identification for railway line tracking. *IEEE Transactions on Intelligent Transportation Systems*, 27(1), 433–447. https://doi.org/10.1109/TITS.2025.3628655.

Hiebert, T., Hofstetter, F., Thomas, C., Mushtaq, S., Kaan, E., Parameswaran, B., 2025. Labels4Rails: A railway image annotation tool and associated reference dataset. *Data*, 10(12). https://doi.org/10.3390/data10120210.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. LoRA: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685.

Jaß, P., Thomas, C., 2025. Using N-version architectures for railway segmentation with deep neural networks. *Machine Learning and Knowledge Extraction*, 7(2). https://doi.org/10.3390/make7020049.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J., 2024. LISA: Reasoning segmentation via large language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Laurent, T., 2024. Train ego-path detection on railway tracks using end-to-end deep learning. https://doi.org/10.48550/arXiv.2403.13094.

Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023. Visual instruction tuning. *37th Conference on Neural Information Processing Systems (NeurIPS)*.

Olivier, B., Guo, F., Qian, Y., Connolly, D., 2025. A review of computer vision for railways. *IEEE Transactions on Intelligent Transportation Systems*, 26. https://doi.org/10.1109/TITS.2025.3552011.

Wu, T.-H., Biamby, G., Chan, D., Dunlap, L., Gupta, R., Wang, X., Gonzalez, J., Darrell, T., 2024. See, Say, and Segment: Teaching LMMs to overcome false premises. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G., 2024. GSVA: Generalized segmentation via multimodal large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J., 2024. LISA++: An improved baseline for reasoning segmentation with large language model. https://arxiv.org/abs/2312.17240.

Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., Beleznai, C., 2019. RailSem19: A dataset for semantic rail scene understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.