

Measuring Visual Understanding in Telecom domain: Performance Metrics for Image-to-UML conversion using VLMs

H. G. Ranjani

Ericsson R & D, Bangalore, India
ranjani.h.g@ericsson.com

Rutuja Prabhudesai

IIITB, Bangalore, India
rutuja10feb1999@gmail.com

Abstract

Telecom domain 3GPP documents are replete with images containing sequence diagrams. Advances in Vision-Language Large Models (VLMs) have eased conversion of such images to machine-readable PlantUML (*puml*) formats. However, there is a gap in evaluation of such conversions - existing works do not compare *puml* scripts for various components. In this work, we propose performance metrics to measure the effectiveness of such conversions. A dataset of sequence diagrams from 3GPP documents is chosen to be representative of domain-specific actual scenarios. We compare *puml* outputs from two VLMs — Claude Sonnet and GPT-4V — against manually created ground truth representations. We use version control tools to capture differences and introduce standard performance metrics to measure accuracies along various components: participant identification, message flow accuracy, sequence ordering, and grouping construct preservation. We demonstrate effectiveness of proposed metrics in quantifying conversion errors across various components of *puml* scripts. The results show that nodes, edges and messages are accurately captured. However, we observe that VLMs do not necessarily perform well on complex structures such as notes, box, groups. Our experiments and performance metrics indicates a need for better representation of these components in training data for fine-tuned VLMs.

1 Introduction

Sequence diagrams are widely used to represent signaling sequences and interactions among system components. However, these diagrams are often available only as static images within technical documents and scattered across versions and sections. This limits machine-readability of such sequences and their usability in tools that support analysis, simulation, automated verification and/or troubleshooting. We consider the telecom domain

as a case in point to illustrate some challenges using 3rd Generation Partnership Project (3GPP) specifications (3GPP, 2022). These are publicly available as word documents containing text, tables, equations and images (Roychowdhury et al., 2024a,b, 2025) including sequence diagrams (as images) within to illustrate procedural call flows across various network entities in various scenarios.

Recent advances in Vision-Language Large Models (VLMs) have enabled the extraction of structured information from images, including charts, tables and UML (Unified Modeling Language) diagrams. Several studies have proposed methods to extract UML components from visual representations, converting image-based diagrams into machine-readable formats using tools such as PlantUML to create *puml* scripts (PlantUML, 2025; PlantUMLGuide, 2023; Romeo et al., 2025). These approaches aid towards automating the conversion of legacy diagram archives into usable data. In this work, we use *puml* to refer to UML scripts (accessed or generated using PlantUML tools or equivalent ones). The work in (De Bari, 2024) uses LLMs to generate UML class diagrams. The diagrams are analyzed for syntactic, semantic, and pragmatic quality against that of human generated UML diagrams. In (Ye et al., 2024), flowchart images are converted to graphical structures using VLMs. Then, these structures are compared for the optimal representation format (*puml*, Mermaid or Graphviz) for improved performance in reasoning based question answering (QA) task. For evaluating the UML representation format, node-F1 and edge-F1 metrics have been considered. The work of (Axt, 2023) converts human sketches into UML diagrams using OpenCV libraries. The UML diagrams are evaluated based on precision and recall of classes, inheritances, and associations. The work (Conrardy and Cabot, 2024) also addresses converting human sketches into UML diagrams, by using VLMs. The approach is based on chain-of-

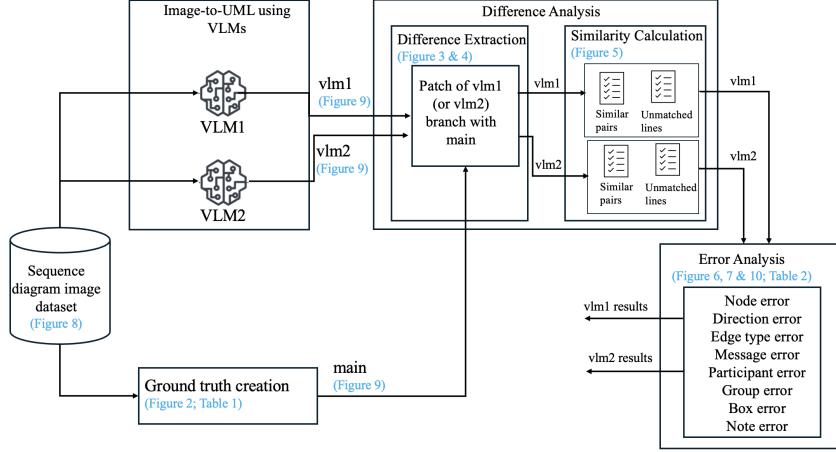


Figure 1: A block diagram depiction of proposed approach to compare image-to-UMLs outputs from two VLMs.

thoughts (CoT) via multiple prompts. They evaluate the approach through count of number of mistakes (including non-compilation, hallucinations, and similar errors). Another recent work (Bates et al., 2025) also leverages multi-modal LLMs to convert image based UML diagrams to *puml* format using fine-tuned VLMs. A synthetic image-based UML dataset was created. The generated *puml* was again visualized as an image and compared with the original image for visual fidelity using Structural Similarity Index Measure (SSIM) and use Bilingual Evaluation Understudy (BLEU) scores for semantic similarity.

Despite these recent efforts, we observe a significant gap in terms of evaluations: currently there is no systematic approach to evaluate the correctness of such image-to-UML datasets. As a result, it is challenging to assess the accuracy of the converted *puml* representation in terms of efficiency of capture of various structural components of the original image. Thus, benchmarking of existing methods and the development of more robust systems is difficult. This is specifically true for complex *puml* representations where synchronous and asynchronous events, grouping of events are important information to be captured.

A wishlist of error metrics to compare two UML diagrams include errors pertaining to components such as participants, connector types, connector directions, messages passed, notes, sequences, groups amongst other syntactic and semantic components. In this work, we address this gap by introducing a set of metrics across these components to measure correctness of *puml* conversions.

The dataset chosen focuses on the telecom do-

main using 3GPP sequence diagrams (parsed from publicly available documents), where the sequence diagrams range from simple to complex and include synchronous and asynchronous events and contain many *puml* components listed above. We propose to measure the differences between two *puml* representations. Towards this, we manually curate ground truth and quantify *puml* output VLM for various components: participants, message flow, ordering, and grouping constructs.

The contributions of this work are: (i) compare VLM (Claude and GPT ¹) performances based on their ability to convert sequence diagram images to *puml* format (ii) propose use of version control tools for capture differences between *puml* format (iii) introduce metrics for various components to measure differences between various components of *puml* representations.

The manuscript is organized as follows: Section 2 details proposed approach, followed by experimental setup in Section 3 which includes dataset preparation, *puml* script generation and difference analysis. This is followed by the detailing of metrics introduced for error analysis in Section 2.3.2. Detailed analysis of the results is in Section 4 followed by concluding remarks in Section 5.

2 Proposed Approach

Figure 1 shows a block diagram representation of the proposed approach. As can be seen, there are 3 major steps:

- Image-to-UML conversion using VLMs.
- Difference analysis: extract the patch files

¹In this work, we use GPT-4 and GPT to refer to GPT4-V. Similarly, we use Claude to refer to Claude Sonnet.

between VLM outputs and ground truth. For analysis, the patch files are grouped into two groups (i) similar pairs and (ii) unmatched lines.

- Error analysis: introduce error metrics into various categories such as node-based, edge-based, message-based and other structural components (detailed in Section 2.2.2)

We detail each step in subsections below.

2.1 Image-to-UML conversion using VLMs

We parse images from the publicly available 3GPP standard documents. We manually categorize these images into two: (i) sequence diagrams and (ii) non-sequence diagrams. This image categorization can be automated with a fine-tuned classifier such as (Moreno et al., 2020; Soman et al., 2025), but is not the focus of this work. We curate a subset of these sequence diagrams for analysis in this work. More details on the dataset is in Section 3.1.

We consider two VLMs (Claude and GPT-4) for converting sequence diagram images to *puml* formats. We include details of the VLM models considered in Section 3.2.3.

2.2 Difference Analysis

The *puml* files are diagrams-as-code scripts. In order to capture and evaluate the differences between the VLM outputs and ground truth, we consider version control tools, due to the textual nature of the *puml* scripts. We obtain the differences between the ground truth and VLM outputs by first extracting differences and then performing a similarity calculation.

2.2.1 Difference extraction

There are two components involved in difference extraction:

- Repository creation: Here, we consider the ground truth *puml* scripts in the main branch of a git repository while each set of the VLM outputs with same name as corresponding ground truth files can be considered under a separate branch.
- Patch extraction: Git diff or patch files are generated for each VLM model output with respect to main branch. These captures document specific differences with respect to ground truth.

2.2.2 Similarity calculation

We propose a multi-step approach for similarity calculation using the git diff/patch files.

1. **Preprocess the diff/patch** file such that only relevant lines are considered for analysis:

- Lines containing arrows (connections between participants) such as “*PGWA -> SGW* : 2a. *Update Bearer Request (PGW Change Info)*”
- Lines with keywords such as “group”, “note”, “box”, “participant”, “actor”.
- Non-structural lines are excluded (e.g., “end”, “end group”, “end note”, “end box”)

2. **Pairing of lines** to align groups of lines and find optimal matching lines from the diff file.

For this, Levenshtein distance (Levenshtein, 1966) is calculated between every element of removed (starts with -) and added (starts with +) lines to quantify textual differences. Linear sum assignment implementation (Lin, 2025) of modified Jonker-Volgenant algorithm (Crouse, 2016) is applied to find optimal matching between removed and added lines, minimizing the total distance. This pairing approach identifies candidate lines in the model output which correspond to similar lines in the ground truth. Similar lines can be considered to be aligned based on minimal Levenshtein distance between candidate pairs.

3. **Post processing** output of the pairing of lines process includes:

- Unpaired lines from removed groups are considered as elements missing in model output.
- Unpaired lines from added groups are considered as insertions w.r.t. ground truth.
- Paired lines with differences are categorized either as substitutions or as a combination of additions and deletions, depending on the component, nature and extent of the change.

2.3 Error Analysis

Consider a sample *puml* sequence component “*PGWA -> SGW* : 2a. *Update Bearer Request (PGW Change Info)*”. Here, ‘PGWA’ and ‘SGW’ are considered as nodes, ‘->’ corresponds to edges, ‘2a. *Update Bearer Request (PGW Change Info)*’

corresponds to message element. In addition, there are other components such as box, group, participant and notes.

In this subsection, we describe the process considered for error categorization into various components, followed by the metrics introduced in this work to measure the categorized errors.

2.3.1 Error Categorization

For each paired line with Levenshtein distance > 0 (indicating a difference), and unpaired lines, the specific type of difference is determined through regex pattern matching and context analysis (such as presence of “->”, “.” patterns). Each difference is categorized into one of the pre-defined categories based on the specific nature of the error (e.g., arrow direction, message content, participant name). The categories are based on the specific element types:

- Node related errors (participant identification issues)
- Edge or connection errors (arrow direction, type)
- Message content errors
- Other structural element errors: notes, groups, boxes, participants².

2.3.2 Error Metrics

We quantify the differences between the ground truth and model output scripts for each of the categorized errors using the metrics introduced and detailed component-wise below:

- **Node related Metrics:** These metrics are closely associated with participants occurring in each sequence of the *puml* scripts.

- *Node Insertion rate:* Count of nodes in present in model output, but not in ground truth divided by total number of nodes in ground truth.
- *Node Deletion rate:* Count of nodes not present in model output, but present in ground truth divided by total number of nodes in ground truth.
- *Node Substitution rate:* Count of nodes with incorrect naming/representation divided by total number of nodes in ground truth. It is also associated with edit distance to quantify the incorrectness.

²We differentiate nodes and participants based on the context of their occurrence. Lines such as ‘participant PGWA’ contribute to participant category while lines such as ‘PGWA -> SGW : 2a. Update Bearer Request’ contributes to node category.

- **Edge/Connection Metrics:** These metrics are associated with connectors (or edges) and include:

- *Edge Direction change rate:* Count of arrows with incorrect direction in model output divided by total number of arrows in ground truth.
- *Edge Direction insertion rate:* Count of inserted arrows not in ground truth, but present in model output divided by total number of arrows in ground truth.
- *Edge Direction deletion rate:* Count of deleted arrows not in model output, but present in ground truth divided by total number of arrows in ground truth.
- *Edge Direction type change rate:* Count of arrows with incorrect type divided by total number of arrows in ground truth (e.g., solid vs. dashed i.e., ‘->’ vs. ‘-->’ representing synchronous message vs. asynchronous message)

- **Message related Metrics:** Most sequence diagrams considered show passing of messages between participants. Through these metrics, we can measure correctness of messages passed between participants.

- *Message insertion rate:* Count of inserted messages present in model output, not present in ground truth divided by total number of messages in ground truth.
- *Message deletion rate:* Count of messages present in ground truth, not present in model output divided by total number of messages in ground truth.
- *Message change rate:* Count of messages with non-exact matches in model output divided by total number of messages in ground truth.

- **Structural Element Metrics:** In addition to the nodes (participants), edges (connectors) and messages, there exist other structural elements in a complex *puml* diagram such as notes, groups, boxes.

- *Note Changes:* Rate of insertion, deletion, and substitutions of notes.
- *Group Changes:* Rate of insertion, deletion, and substitutions of groups.
- *Box Changes:* Rate of insertion, deletion, and substitution of boxes.

- *Participant Changes*: Rate of insertion, deletion, and substitution of participants.

3 Experimental Setup

In this section, we detail the setup considered for the experiments to measure effectiveness of VLMs for sequence diagram images-to-UML conversion.

3.1 Dataset Preparation

We parse 3GPP (Rel 18) documents (3GPP, 2022) for all the images in the word doc and docx files. The corresponding image-dataset comprises of ~ 14000 images. The images along with their captions are collected and labeled in accordance to the order of their occurrence in the documents. This dataset contains various categories corresponding to graphs, sequence diagrams, frequency diagrams, block diagrams and schematic diagrams. A sample set is shown in Figure 1 in Appendix A (available as Supplementary material)

These images are manually classified into sequence and non-sequence diagram categories; 32% of the images are sequence diagrams. This sequence diagram dataset, along with its corresponding captions corresponds to 4010 images. The total pixel count of these images ranges between 240×57 to 7548×6510 . This collection of sequence diagrams forms the dataset considered for further steps. We highlight that these images do not have the ground truth *puml* script readily available.

A sample representative subset of 50 sequence diagrams are selected from the complete dataset to create ground truth *puml* files. The selection criteria includes diversity of diagram features, including arrow types and styles, color schemes, note positioning, special features such as loops, alternative paths and participant representation styles. All results in this work pertain to these 50 sequence diagrams.

Although readers might presume that 50 files is a modest size dataset, we would like to highlight that the purpose of this work is to propose evaluation metrics considering associated complexities in comparing two *puml* scripts than evaluate the VLMs themselves on large datasets.

3.2 PlantUML Script Generation

Here, we describe the ground truth preparation and approach for *puml* script generation using VLMs.

3.2.1 Ground Truth

The ground truth *puml* scripts are manually created for all 50 selected images. The resulting ground truth scripts serve as the reference for evaluation. The overall number of lines in ground truth *puml* script corresponds to ~ 2500 . The distribution of 50 files w.r.t. number of lines in ground truth is shown in Table 1.

Range of <i>puml</i> script lines	Count of <i>puml</i> files
1-20	10
21-30	13
31-40	5
41-50	13
51-100	9

Table 1: Distribution of *puml* files w.r.t. number of lines of script in ground truth.

3.2.2 VLM Prompt

The following prompt was used to generate *puml* scripts from the diagram images using VLM:

Generate puml script for given 3GPP standard call flow diagram of “[Caption of the image]” according to puml documentation. Please consider following important points:

1. *Correctly identify participants/actors.*
2. *Correctly identify the connection between the nodes using given arrows.*
3. *Correctly identify the arrow direction, start and end of the arrow.*
4. *Correctly identify text associated to the arrow.*
5. *If any text is in rectangles consider them as notes and write them in appropriate place.*
6. *Give numbering to each call sequentially.*
7. *Correctly identify color if any.*

3.2.3 VLM Models

The prompt described above was used with two VLMs:

- Claude 3.7 Sonnet model from Anthropic (Cla, 2025)
- GPT-4-Vision model from OpenAI (GPT, 2024)

VLMs is an evolving field with new models released quite frequently. At this juncture, we again

highlight that although other VLMs can be considered for comparison, the focus of this work is to establish an approach to evaluate *puml* scripts generated from VLMs and not to evaluate all the VLMs as such.

The generated *puml* scripts are rendered using the *puml* web server to manually visually verify syntactic correctness. We do not penalize VLM model outputs unnecessarily during error analysis. Hence, in scenarios where the scripts were not syntactically correct (and leads to not being able to generate the *puml* image), we identify and rectify minor issues such as introduction of spurious characters such as '#' and '–', replace elements identified as actors (by VLMs) as participants because the ground truth contains only participants, invalid arrow syntax such as '..>' to -->. In addition, unsupported note placements, overuse of participants in note overs are manually corrected and not counted towards errors. A few sample instances of such corrections (not counted towards errors) are depicted in Figure 2.

Syntactically Incorrect Instances	Corrected Version
participant UE#1	participant "UE#1" as UE1
actor P-CSCF	actor "P-CSCF" as PCSCF
note bottom of PCSCF NPLI: user location and/or UE timezone information	note over PCSCF, NPLI: user location and/or\nUE timezone information
UPF >> SUPF: Downlink User Plane data	UPF >> SUPF: Downlink User Plane data
UAA <> IMCN: 19. Dialog 1a, on-hold	UAA <> IMCN: 19. Dialog 1a, on-hold
note over App, MediaSessionHandler, MediaPlayer, MediaServer: 5GMS Service Announcement\nand Content Discovery	note over App, MediaServer: 5GMS Service Announcement\nand Content Discovery

Figure 2: Sample snapshot related to manual syntax corrections on VLM outputs with the errors and corrections highlighted in gray.

Sample sequence diagram image from 3GPP standard are depicted in supplementary material, Appendix B along with its corresponding visual renderings generated from *puml* scripts including the ground truth, Claude output, and GPT-4 output. Manually inserted annotations indicate some of the proposed metrics from the *puml* scripts.

3.3 Difference Analysis

We detail the experimental setup for repository creation and for patch extraction steps corresponding to Section 2.2 here.

3.3.1 Repository Structure

A Git repository is created to manage the different versions of *puml* scripts. Each sequence dia-



Figure 3: Snapshot of repository with three branches - main, claude and gpt

gram image is converted to its corresponding *puml* file within a folder corresponding to the document name. Three branches are created:

- main: containing the manually verified ground truth *puml* scripts
- claude: containing Claude generated *puml* scripts
- gpt: containing GPT-4 generated *puml* scripts

Figure 3 shows a sample snapshot of repository branches.

3.3.2 Difference Extraction

A Git diff analysis is performed to identify differences between models and ground truth. Towards this, two sets of patch/diff files are generated for each comparison to document specific differences:

- main (ground truth) with claude branch
- main (ground truth) with gpt branch

These patch/diff files are used for capturing and quantifying differences between ground truth and model outputs. A snapshot of diff file is shown in Figure 4. Figure 5 depicts an example of matched and unmatched lines from the depicted patch file .

4 Results and Analysis

We report the metrics considered for comparing *puml* files, aggregate them and analyze for quantifying behavior of models.

For the error analysis, we first aggregate error count at both file level and at overall dataset level.

The overall dataset analysis provides a complete view of the model performance, while the file level analysis can provide more details of when the model doesn't perform well. For each file comparison, the following were calculated:

- Total count of elements in the ground truth (nodes, arrows, messages, notes, etc.)
- Raw counts of each error type (additions, deletions, substitutions) by category
- Percentage of each error type relative to the total count of relevant elements
- Error density per diagram (errors per element)

```

diff --git a/23007-i21_image_datarId30.puml b/23007-i21_image_datarId30.puml
index 7ff5101..91e9ac0 100644
--- a/23007-i21_image_datarId30.puml
+++ b/23007-i21_image_datarId30.puml
@@ -12,25 +12,23 @@ note over PGWA, PGWB
 to a PDN connection
 end note

-PGWA<--?: e.g. scale-in
-PGWB <--?: e.g. PCF reselects a\new PGW-C/SMF
+PGWA <- PGWA: e.g. scale-in
+PGWB <- PGWB: e.g. PCF reselects a\new PGW-C/SMF

-PGWA --> SGW: 2a. Update Bearer Request\n(PGW Change Info)
-SGW --> MME: 2a. Update Bearer Request\n(PGW Change Info)
-MME --> SGW: 3a. Update Bearer Response
-SGW --> PGWA: 3a. Update Bearer Response
+HSS --> MME: 2a. Update Bearer Request\n(PGW Change Info)
+HSS --> SGW: 2a. Update Bearer Request\n(PGW Change Info)
+MME --> HSS: 3a. Update Bearer Response
+SGW --> HSS: 3a. Update Bearer Response

-MME <- SGW: 2b. Update Bearer Request\n(PGW Change Info)
-SGW <- PGWB: 2b. Update Bearer Request\n(PGW Change Info)
-MME --> SGW: 3b. Update Bearer Response
-SGW --> PGWB: 3b. Update Bearer Response
+MME <- SGW: 2b. Update Bearer Request\n(PGW Change Info)
+MME --> SGW: 3b. Update Bearer Response

-MME --> SGW: 4. Create Session Request\n(PGW Change Indication)
+MME --> PGWB: 4. Create Session Request\n(PGW Change Indication)
| SGW --> PGWB: 4. Create Session Request\n(PGW Change Indication)
-PGWB --> SGW: 5. Create Session Response
-SGW --> MME: 5. Create Session Response
+PGWB --> MME: 5. Create Session Response
+PGWB --> SGW: 5. Create Session Response

-MME --> HSS: 6. Notify (PGW-B FQDN)
-MME <- HSS: 7. Notify response
+MME <- SGW: 6. Notify (PGW-B FQDN)
+MME --> SGW: 7. Notify response

@enduml
diff --git a/23122-i40_image_data_rId21.puml b/23122-i40_image_data_rId21.puml
index 79f80db..5b627cf 100644
--- a/23122-i40_image_data_rId21.puml
+++ b/23122-i40_image_data_rId21.puml
@@ -10,7 +10,7 @@ note over VPLMN
 2. Registration procedure initiation
 end note

-VPLMN --> HPLMN: Nudm_UECM_Registration request
+VPLMN --> HPLMN: Nudm_UECM_Registration request

```

Figure 4: An example of patch/diff file obtained by comparing model (claude/gpt branch) output with ground truth (main branch).

```

*** File: Data/Diff/23007-i21_image_datarId30.puml ===
--- Group 1 ---
Removed : 'PGWA<--?: e.g. scale-in'
Added : 'PGWA <- PGWA: e.g. scale-in'
Distance: 6
-----Removed : 'PGWB <--?: e.g. PCF reselects a\new PGW-C/SMF'
Added : 'PGWB <- PGWB: e.g. PCF reselects a\new PGW-C/SMF'
Distance: 5
-----Group 2 ---
Removed : 'PGWA --> SGW: 2a. Update Bearer Request\n(PGW Change Info)'
Added : 'HSS --> SGW: 2a. Update Bearer Request\n(PGW Change Info)'
Distance: 5
-----Removed : 'HSS --> MME: 2a. Update Bearer Request\n(PGW Change Info)'
Added : 'HSS --> MME: 2a. Update Bearer Request\n(PGW Change Info)'
Distance: 4
-----Removed : 'MME --> SGW: 3a. Update Bearer Response'
Added : 'MME --> HSS: 3a. Update Bearer Response'
Distance: 4
-----Removed : 'SGW --> PGWA: 3a. Update Bearer Response'
Added : 'SGW --> HSS: 3a. Update Bearer Response'
Distance: 5
-----Group 3 ---
Removed : 'MME <- SGW: 2b. Update Bearer Request\n(PGW Change Info)'
Added : 'MME <- SGW: 2b. Update Bearer Request\n(PGW Change Info)'
Distance: 1
-----Removed : 'MME --> SGW: 3b. Update Bearer Response'
Added : 'MME --> SGW: 3b. Update Bearer Response'
Distance: 1
-----Unmatched Removed Lines:
- SGW <- PGWB: 2b. Update Bearer Request\n(PGW Change Info)
-----Group 4 ---
Removed : 'MME --> SGW: 4. Create Session Request\n(PGW Change Indication)'
Added : 'MME --> PGWB: 4. Create Session Request\n(PGW Change Indication)'
Distance: 2
-----Group 5 ---
Removed : 'PGWB --> SGW: 5. Create Session Response'
Added : 'PGWB --> MME: 5. Create Session Response'
Distance: 1
-----Removed : 'SGW --> MME: 5. Create Session Response'
Added : 'PGWB --> MME: 5. Create Session Response'
Distance: 3
-----Group 6 ---
Removed : 'MME <- HSS: 6. Notify (PGW-B FQDN)'
Added : 'MME <- SGW: 6. Notify (PGW-B FQDN)'
Distance: 5
-----Removed : 'MME <- HSS: 7. Notify response'
Added : 'MME --> SGW: 7. Notify response'
Distance: 5
*** File: Data/Diff/23122-i40_image_data_rId21.puml ===

```

Figure 5: An example of distance calculation and classification of matched pairs and unmatched lines from patch/diff file shown in Fig. 4.

	Node	Direction change	Direction type	Message	Box	Group	Note	Participants
Ground truth count								
	1736	881	881	873	19	39	229	278
Error analysis for Claude output								
Insertion (%)	13.02	12.71	0.00	13.06	42.11	0.00	17.03	0.72
Deletion (%)	15.78	15.44	0.00	14.89	52.63	69.23	17.90	2.88
Substitution (%)	12.56	6.02	10.90	11.23	21.05	15.38	31.88	8.63
Error analysis for GPT-4 output								
Insertion (%)	19.76	19.41	0.00	19.70	0.00	69.23	42.36	1.44
Deletion (%)	18.66	18.50	0.00	17.87	100.00	76.92	64.19	5.76
Substitution (%)	34.10	7.95	16.00	39.29	0.00	5.13	35.37	40.64

Table 2: Statistics of components of *puml* in Ground truth and error analysis metrics for the same using Claude and GPT-4 output files.

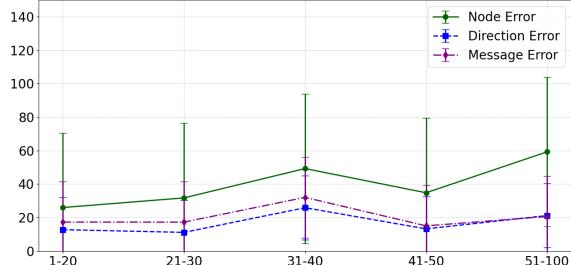


Figure 6: Error bars for node, message, direction error rates in Claude outputs based on number of lines in ground truth files.

Table 2 depicts the percentage of insertion, deletion and substitution rates measured for Claude and GPT-4 models w.r.t. the ground truth. We observe that Claude outputs have lesser number of insertion, deletion and substitution rate than GPT-4 outputs across almost all components of *puml*³. The direction type errors are mostly related to substitution because they correspond to synchronous being categorized as asynchronous or vice-versa.

It is worth noting that there are higher errors in both VLMs outputs with respect to structural elements such as box, group and notes. This indicates that it might be required to fine-tune VLMs for such tasks to reduce error rate across these components.

We further analyze the VLM outputs using these metrics in terms of number of lines of the *puml* script. The errors are accumulated across insertion, deletion and substitution categories at a file level and calculated as percentage of total. Figure 6 and Figure 7 depict the same. We observe that the trend of percent of error in Claude increases with increasing number of lines of script. This is expected because when the sequence diagram is longer, there is less likelihood of retaining the visual context and it is possible that there are more errors. With GPT-4, however, the error rate shows a decreasing trend. This, although is not intuitive, hints that GPT-4 retains higher visual context in more complex sequence diagrams over that of simpler one. This needs further investigation. In summary, the overall performance seen from Claude model is much better than that of GPT4.

5 Conclusions

It is possible to convert images to *puml* scripts. We have explored the use of VLMs on limited set of sequence diagrams from publicly available 3GPP

³Claude performance is statistically significant over GPT4 at $p < .05$ except for box and group components.

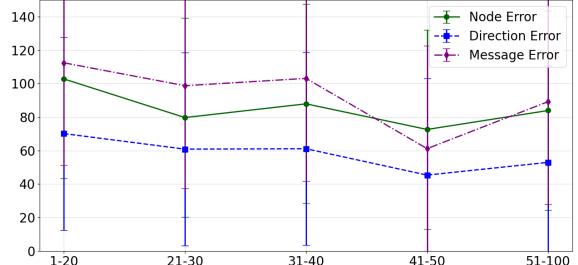


Figure 7: Error bars for node, message, direction error rates in GPT-4 outputs based on number of lines in ground truth files.

documents. These have applications in telecom network analysis, simulation, and automated verification systems.

In this work, we highlight the lack of systematic evaluation of image-to-UML conversion using VLMs. We propose to use version control tools to capture the differences in *puml* representations between ground truth and VLM outputs. We analyze the patch files, align them to be able to capture effectiveness of the *puml* conversion. We propose a set of performance metrics to measure the effectiveness of image-to-uml conversion across various components (viz. nodes, edges, messages, participants, box, groups and notes). We observe that Claude model is more effective than GPT-4 in the *puml* conversion for the considered dataset.

The errors are concentrated on complex components such as box, groups, notes. It is expected that a fine-tuning of VLMs focused on sequence diagrams to improve effectiveness for such components. To realize the same, it is important to ensure that training set has these components included appropriately.

We also observe that errors for Claude increases with increasing number of lines in the script. This is expected as retaining longer visual context may be challenging. However, GPT-4 shows that performance is not much impacted by the number of lines in the script. This is unexpected and necessitates a detailed further analysis.

In this work, we have focused on simple prompts for the VLMs. Future experiments can include advanced prompts to introduce chain-of-thought approach for image-to-UML conversions. While this work has used limited and focused number sequence diagrams from publicly available 3GPP specifications, the proposed set of performance metrics are agnostic to the domain, source and dataset size of sequence diagram images.

References

2024. GPT4(OpenAI) Vision. <https://openai.com/index/gpt-4/>.

2025. Claude 3.7 Sonnet from Anthropic. <https://www.anthropic.com/news/clause-3-7-sonnet>.

2025. Linear Sum Assignment. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linear_sum_assignment.html.

3GPP. 2022. 3GPP release 18. Technical report. Accessed: 2024-05-19.

Monique Axt. 2023. Transformation of sketchy UML Class Diagrams into formalPlantUML models.

Averi Bates, Ryan Vavricka, Shane Carleton, Ruosi Shao, and Chongle Pan. 2025. Unified modeling language code generation from diagram images using multimodal large language models. *Machine Learning with Applications*, page 100660.

Aaron Conrardy and Jordi Cabot. 2024. From image to UML: first results of image based UML diagram generation using LLMs. *arXiv preprint arXiv:2404.11376*.

David F. Crouse. 2016. *On implementing 2D rectangular assignment algorithms*. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696.

Daniele De Bari. 2024. *Evaluating large language models in software design: A comparative analysis of UML class diagram generation*. Ph.D. thesis, Politecnico di Torino.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Valentín Moreno, Gonzalo Génova, Manuela Alejandres, and Anabel Fraga. 2020. Automatic classification of web images as uml static diagrams using machine learning techniques. *Applied Sciences*, 10(7):2406.

PlantUML. 2025. Open-source tool that uses simple textual descriptions to draw beautiful UML diagrams. <https://plantuml.com/>.

PlantUMLGuide. 2023. Drawing UML with PlantUML: PlantUML Language Reference Guide. https://pdf.plantuml.net/PlantUML_Language_Reference_Guide_en.pdf.

Joseph Romeo, Marco Raglanti, Nagy Csaba, and Michele Lanza. 2025. UML is back. Or is it? Investigating the Past, Present, and Future of UML in Open Source Software. In *ICSE 2025 47th International Conference on Software Engineering*.

Sujoy Roychowdhury, HG Ranjani, Sumit Soman, Nishtha Paul, Subhadip Bandyopadhyay, and Siddhanth Iyengar. 2025. Intelligibility of text-to-speech systems for mathematical expressions. *arXiv preprint arXiv:2506.11086*.

Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024a. Evaluation of RAG metrics for question answering in the telecom domain. *arXiv preprint arXiv:2407.12873*.

Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Avantika Sharma, Neeraj Gunda, and Sai Krishna Bala. 2024b. Evaluation of Table Representations to Answer Questions from Tables in Documents: A Case Study using 3GPP Specifications. *arXiv preprint arXiv:2408.17008*.

Sumit Soman, HG Ranjani, Sujoy Roychowdhury, VDSN Sastry, Akshat Jain, Pranav Gangrade, and Ayaaz Khan. 2025. A Graph-based Approach for Multi-Modal Question Answering from Flowcharts in Telecom Documents. *KDD 2025 Workshop on Structured Knowledge for Large Language Models*.

Junyi Ye, Ankan Dash, Wenpeng Yin, and Guiling Wang. 2024. Beyond End-to-End VLMs: Leveraging Intermediate Text Representations for Superior Flowchart Understanding. *arXiv preprint arXiv:2412.16420*.

NOTE TO REVIEWERS: We have included the appendix section along with the submission as there was no explicit provision/option/portal field in to include the supplementary material inspite of request for clarifications to org team. Our submission otherwise conforms to the required number of pages (8 pages + 1 for reference).

Appendix A 3GPP image categories

A sample snapshot of various categories of images present in 3GPP standards are shown in Figure 8.

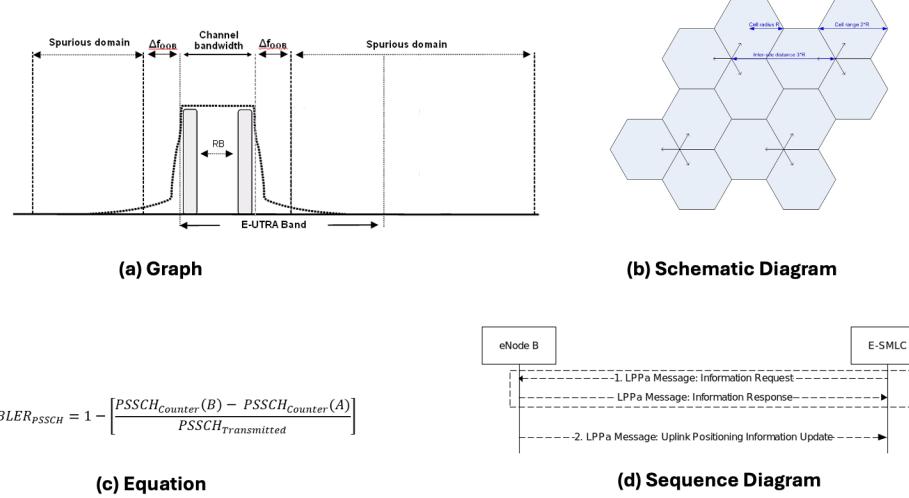


Figure 8: Representative images from various categories from the 3GPP dataset.

Appendix B Sample sequence diagrams from 3GPP specifications

Figure 9 shows sample sequence diagram seen in 3GPP standard and it's equivalent image constructed from various puml scripts (viz. ground truth scripts and scripts from two VLMs models as output.). Figure 10 shows the corresponding *puml* scripts.

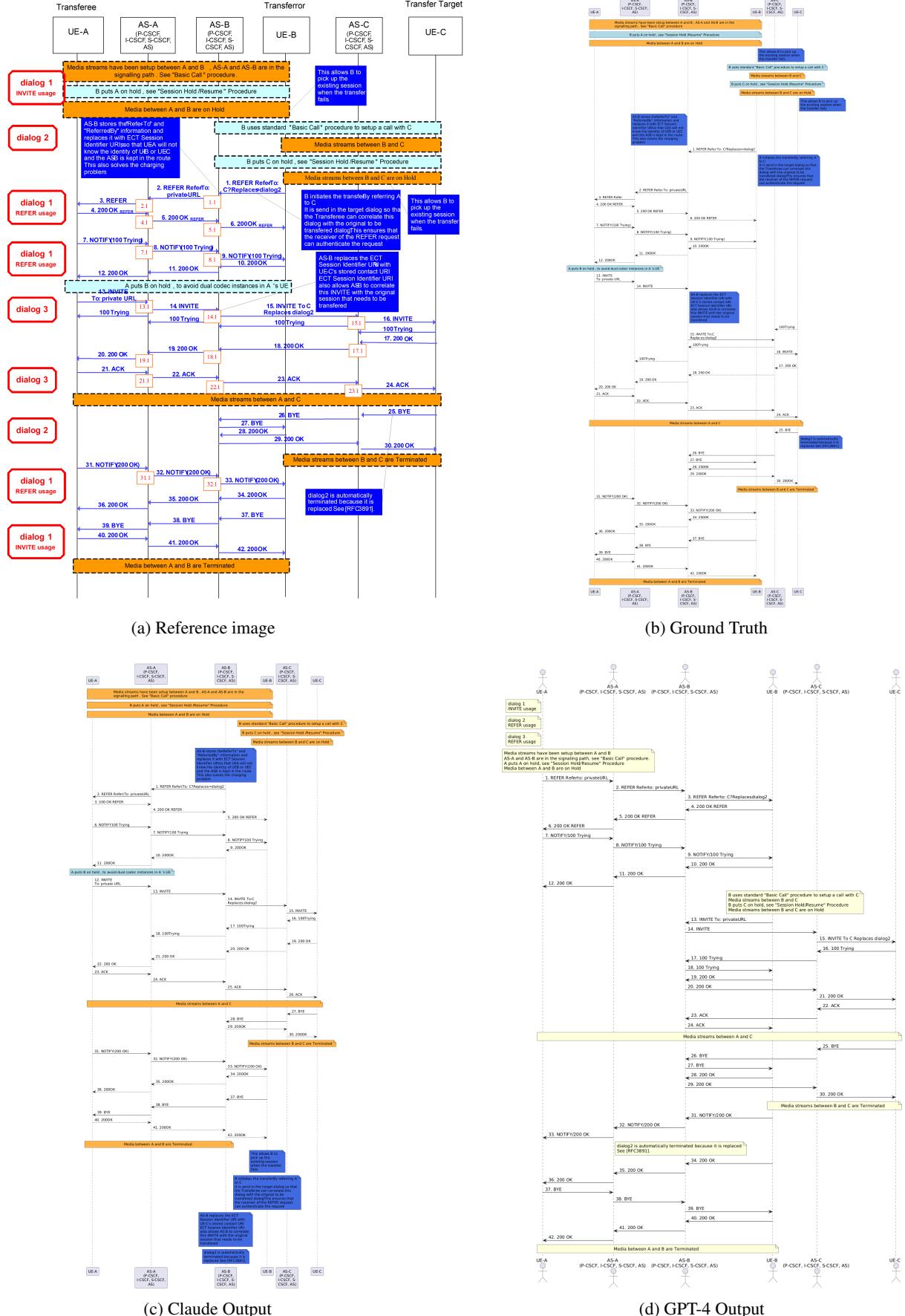


Figure 9: Reference image from the 3GPP standard dataset, along with manually created ground truth and outputs from Claude and GPT-4, as visualized on the puml web server.

Figure 10: Comparison of ground truth, Claude and GPT-4 *puml* script with metrics