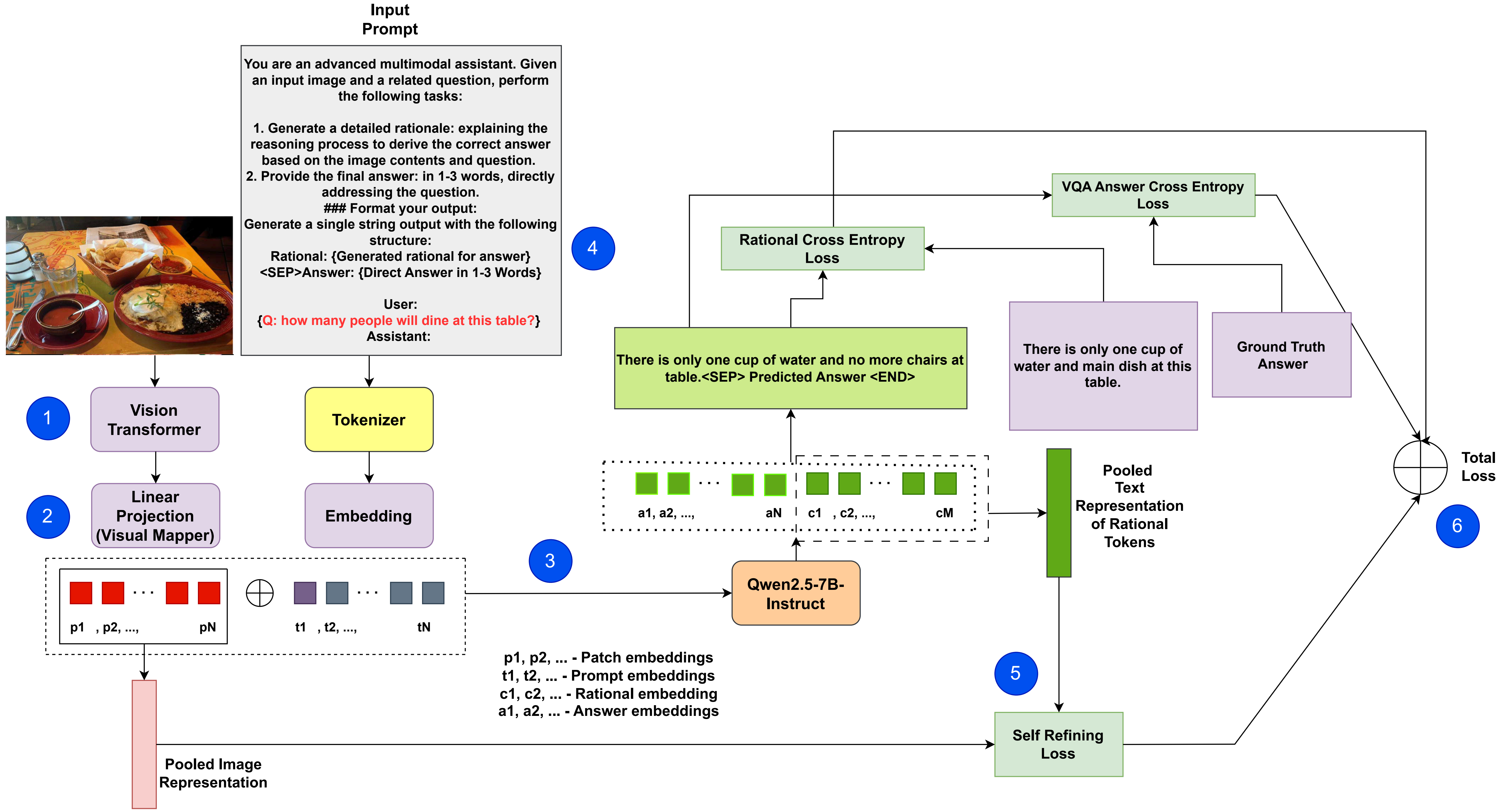

Input
Prompt
1
Vision Transformer
2
Linear Projection (Visual Mapper)
Tokenizer
Embedding
p1 , p2, ..., pN
t1 , t2, ..., tN
3
Qwen2.5-7B-Instruct
a1, a2, ..., aN
c1 , c2, ..., cM
There is only one cup of water and no more chairs at table.<SEP> Predicted Answer <END>
Rational Cross Entropy Loss
VQA Answer Cross Entropy Loss
There is only one cup of water and main dish at this table.
Ground Truth Answer
Pooled Text Representation of Rational Tokens
Pooled Image Representation
5
Self Refining Loss
Total Loss
6
4

You are an advanced multimodal assistant. Given an input image and a related question, perform the following tasks:

1. Generate a detailed rationale: explaining the reasoning process to derive the correct answer based on the image contents and question.
2. Provide the final answer: in 1-3 words, directly addressing the question.

Format your output:
Generate a single string output with the following structure:
Rational: {Generated rational for answer}
<SEP>Answer: {Direct Answer in 1-3 Words}

User:
{Q: how many people will dine at this table?}
Assistant:

p1, p2, ... - Patch embeddings
t1, t2, ... - Prompt embeddings
c1, c2, ... - Rational embedding
a1, a2, ... - Answer embeddings