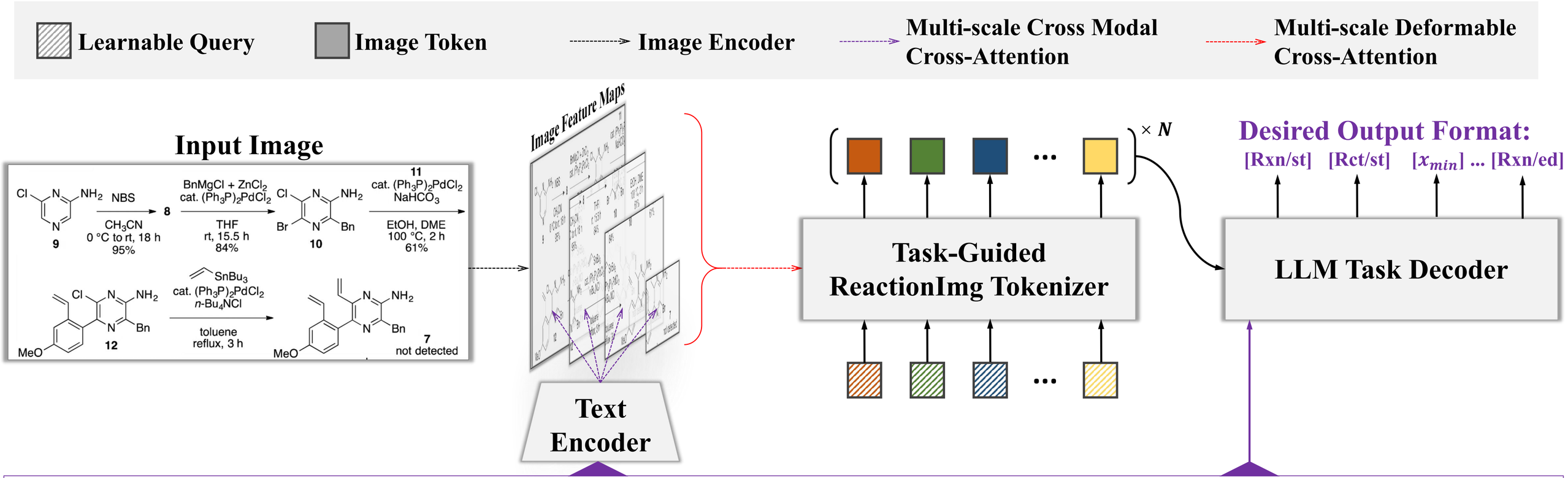
Learnable Query
Image Token
Image Encoder
Multi-scale Cross Modal Cross-Attention
Multi-scale Deformable Cross-Attention
Input Image
Image Feature Maps
Text Encoder
Task-Guided ReactionImg Tokenizer
× N
LLM Task Decoder
Desired Output Format:
[Rxn/st] [Rct/st] [x_{min}] ... [Rxn/ed]
Reaction Extraction Task: "Please list every reaction in this image <image> in detail. For each reaction, include the category and unique ID of each object, along with their coordinates [x1, y1, x2, y2]. Categories include Structure [Str] and Text [Txt]. Describe their roles in each reaction(<Rxn/st> to <Rxn/ed>)⋯ . Structured output format should be:[Rxn/st][Rct/st](object 1)...[Rct/ed][Cnd/st](object 2)...[Cnd/ed][Prd/st](object 3)...[Prd/ed][Rxn/ed],[Rxn/st]⋯ ."
Condition OCR and Role Identification Task: "For the given image <image>, what words are written in this text box<objs>?, And please indicate the condition role[Role] of each word in: solvent[Svt], agent[Agt], temperature[Agt], time [Time] and yield[Yld]. Structured output format should be:'Text content'[Role],⋯ ."
Task Instructions