LLM Decoder (Qwen-2) ❄/❄→🔥→🔥

<The key frames of this video are:>+<**Visual Tokens**>+<, and the motion features of the video are: >+<**Motion Tokens**>+<. Based on this, answer the following question.> + <Question>

🧩 Interleaved Input Token Sequences

<Position Embeddings>

<Visual Tokens>

<Motion Tokens>

+

<Text Tokens>

🖼 Vision Projector (FC) 🔥/❄→🔥→🔥

📽 Motion Projector (FC) ❄/🔥→🔥→🔥

T Text Embedding Layers ❄/❄→🔥→❄

👁 Vision Tower (SigLIP) 🔥/❄→🔥→🔥

🏃 Motion Extractor (SlowFast-R50) ❄/🔥→🔥→🔥

PROMPT **Prompt**

Key Frames (1fps)

Densely Extracted Frame Sequence

Stage-1+Stage-2

🎛 **VQA²-Scorers**

The quality of the video is

[high]	P_{High}
[good]	P_{Good}
[fair]	P_{Fair}
[Poor]	P_{Poor}
[low]	P_{Low}

$P_{quality\_level}$
Model output Logit of the specific [quality level]

Softmax Normalization

Weighted Summation

Predicted Score

Higher PLCC/SRCC 👍

Ground Truth

$$Q = \sum_{i=1}^{5} \omega_i \frac{e^{P_{quality\_levels[i]}}}{\sum_{i=1}^{5} e^{P_{quality\_levels[i]}}}$$

$\omega \xrightarrow{[1.0, 0.75, 0.5, 0.25, 0]}_{High \qquad\qquad Low}$

Stage-1+Stage-2+Stage-3

🤖 **VQA²-Assistant**

I can assist you in evaluating the overall quality or the severity of a specific distortion of a video, identifying the location and time where the distortion occurs, and providing effective post-processing or reshooting recommendations to enhance the quality. Feel free to ask!

Evaluate on Q-bench-video Benchmark

Greater precision and relevance 👍

GPT judgement

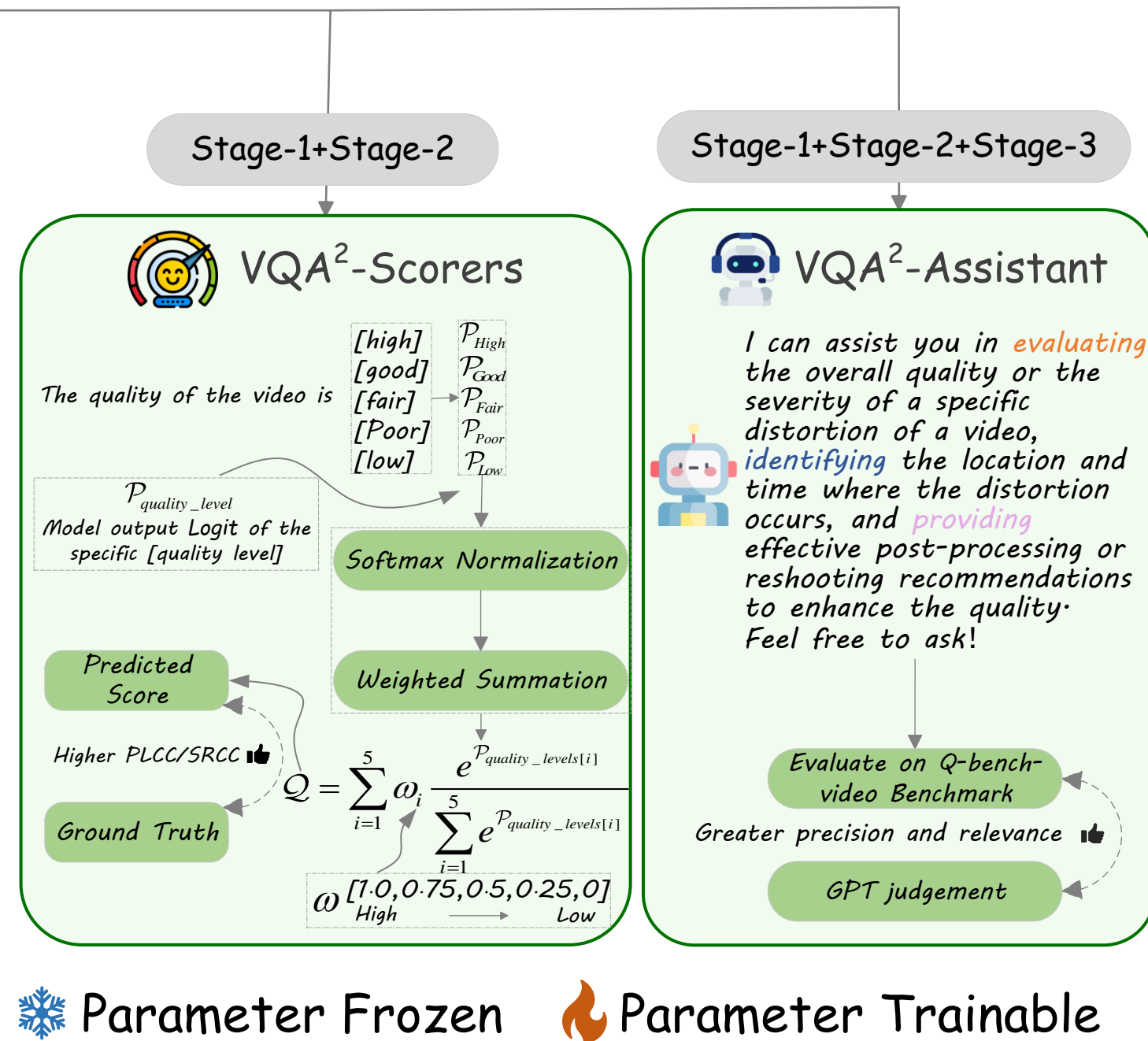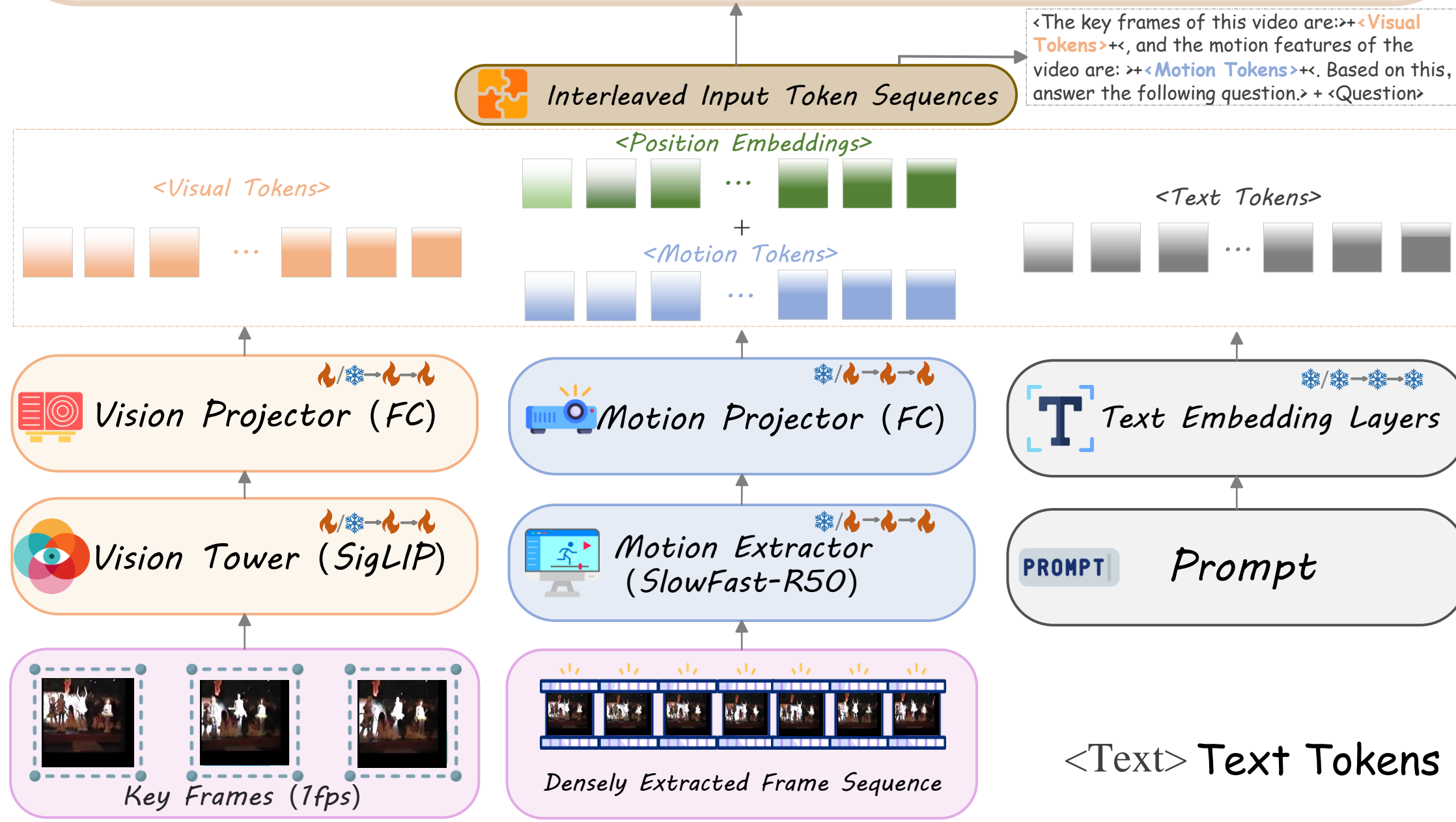<Text> **Text Tokens**   ❄ **Parameter Frozen**   🔥 **Parameter Trainable**