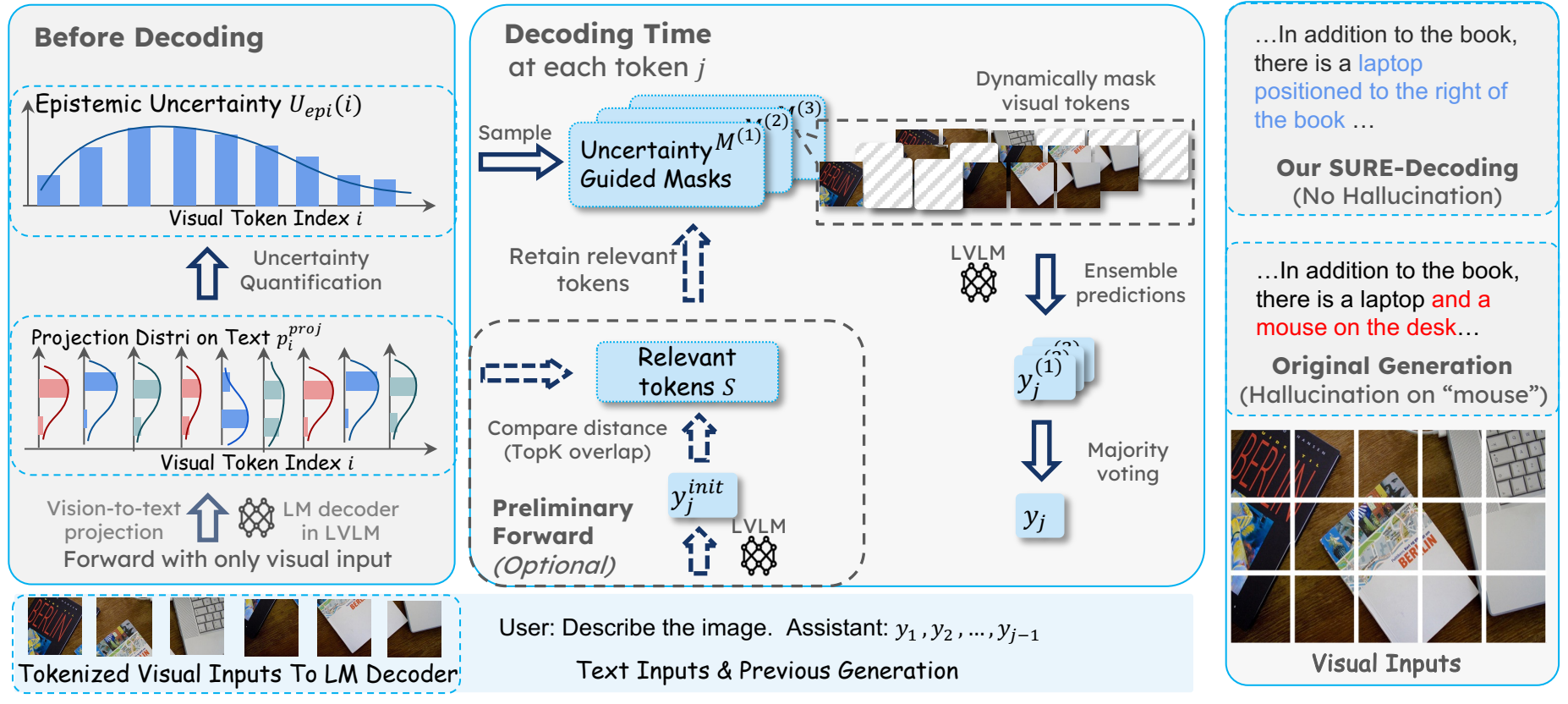
Before Decoding
Epistemic Uncertainty $U_{epi}(i)$
Visual Token Index i
Uncertainty Quantification
Projection Distri on Text p_i^{proj}
Visual Token Index i
Vision-to-text projection
LM decoder in LVLM
Forward with only visual input
Decoding Time
at each token j
Sample
Uncertainty Guided Masks $M^{(1)}$ $^{(2)}$ $^{(3)}$
Dynamically mask visual tokens
Retain relevant tokens
LVLM
Ensemble predictions
Relevant tokens S
Compare distance (TopK overlap)
Preliminary Forward (Optional)
y_j^{init}
LVLM
$y_j^{(1)}$
Majority voting
y_j
…In addition to the book, there is a laptop positioned to the right of the book …
Our SURE-Decoding
(No Hallucination)
…In addition to the book, there is a laptop and a mouse on the desk…
Original Generation
(Hallucination on “mouse”)
Visual Inputs
Tokenized Visual Inputs To LM Decoder
User: Describe the image. Assistant: $y_1, y_2, \ldots, y_{j-1}$
Text Inputs & Previous Generation