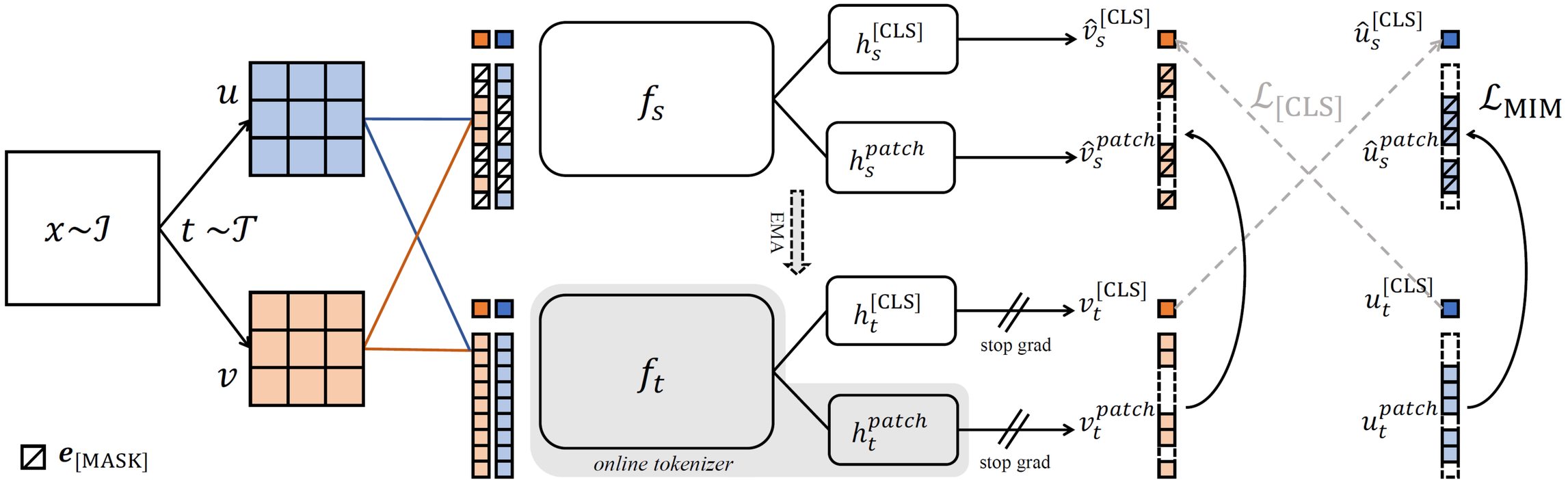

Figure 3: **Overview of iBOT framework, performing masked image modeling with an *online tokenizer*.** Given two views $\boldsymbol{u}$ and $\boldsymbol{v}$ of an image $\boldsymbol{x}$, each view is passed through a teacher network $h_t \circ f_t$ and a student network $h_s \circ f_s$. iBOT minimizes two losses. The first loss $\mathcal{L}_{\texttt{[CLS]}}$ is self-distillation between cross-view `[CLS]` tokens. The second loss $\mathcal{L}_{\text{MIM}}$ is self-distillation between in-view patch tokens, with some tokens masked and replaced by $\boldsymbol{e}_{\texttt{[MASK]}}$ for the student network. The objective is to reconstruct the masked tokens with the teacher networks' outputs as supervision.