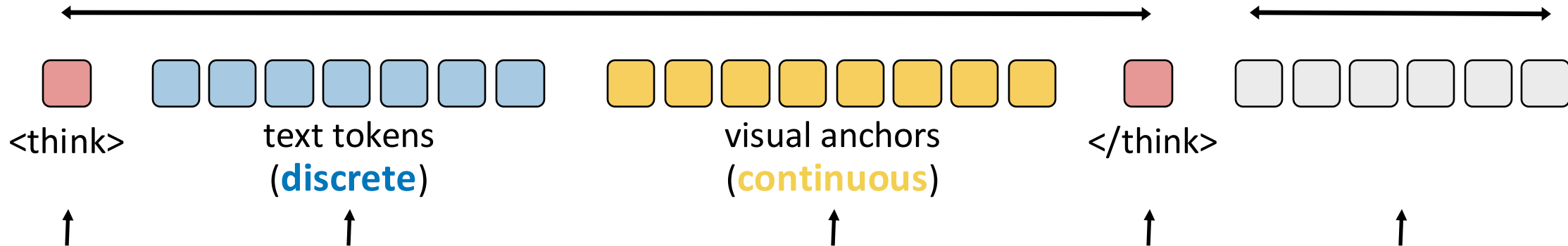How many clouds within the image?

Chain-of-**Multimodal-Thoughts** (CoMT)

Answers

<think>

text tokens (**discrete**)

visual anchors (**continuous**)

</think>

Vision-Language Models (VLMs)