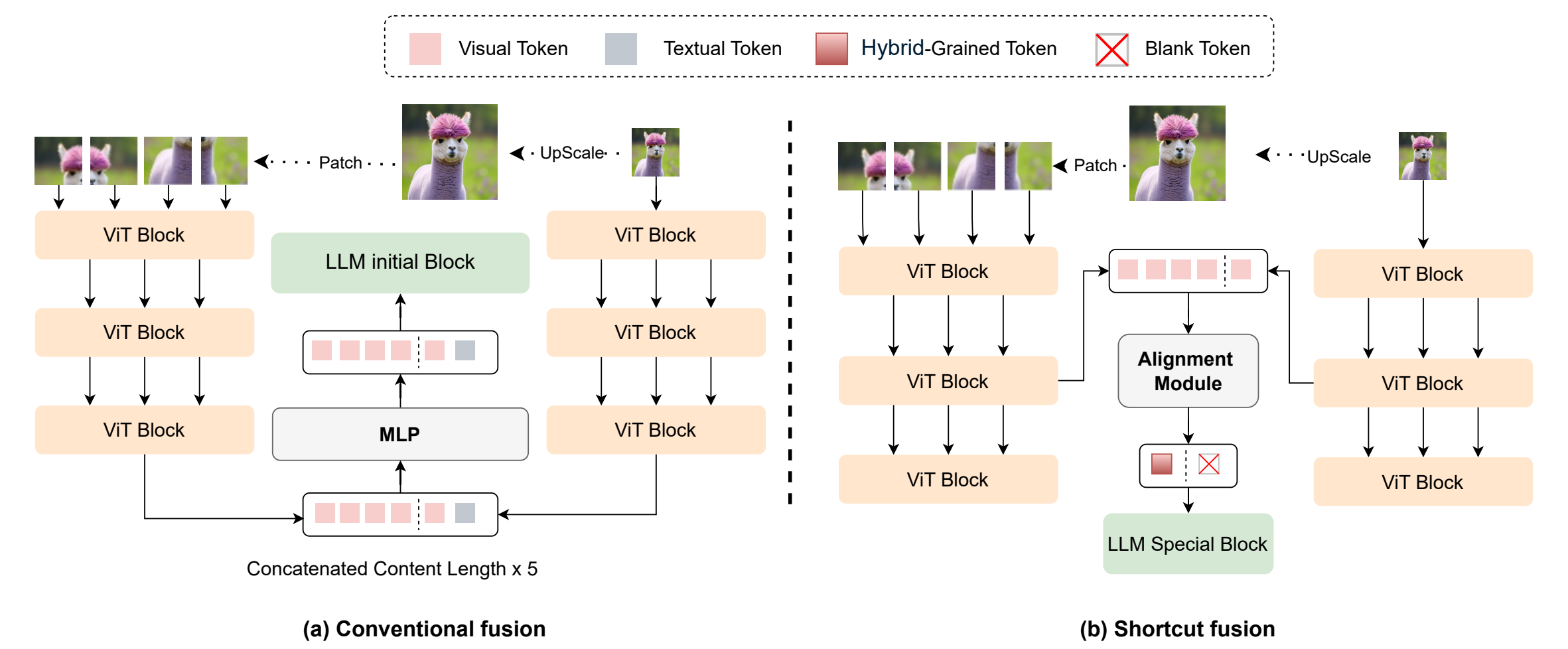
Visual Token
Textual Token
Hybrid-Grained Token
Blank Token
Patch
UpScale
ViT Block
ViT Block
ViT Block
LLM initial Block
MLP
ViT Block
ViT Block
ViT Block
Concatenated Content Length x 5
(a) Conventional fusion
Patch
UpScale
ViT Block
ViT Block
ViT Block
Alignment Module
LLM Special Block
ViT Block
ViT Block
ViT Block
(b) Shortcut fusion