

Adapter-Style Vision-Language Models:

Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. A Vision Transformer already produces vector embeddings. Why can these embeddings not always be inserted unchanged into a decoder-only language model? Discuss embedding width, sequence length, representation compatibility, and preservation of spatial information.

Adapter-Style Vision-Language Models: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. A Vision Transformer already produces vector embeddings. Why can these embeddings not always be inserted unchanged into a decoder-only language model? Discuss embedding width, sequence length, representation compatibility, and preservation of spatial information.
2. Describe the path followed by an image from the input image to the LLaVA language model. Explain how the vision encoder creates patch features, how the projector transforms them, where the projected visual tokens are inserted, how the visual prefix differs from ordinary text context, and why later assistant tokens can use the visual information.

Adapter-Style Vision-Language Models: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. A Vision Transformer already produces vector embeddings. Why can these embeddings not always be inserted unchanged into a decoder-only language model? Discuss embedding width, sequence length, representation compatibility, and preservation of spatial information.
2. Describe the path followed by an image from the input image to the LLaVA language model. Explain how the vision encoder creates patch features, how the projector transforms them, where the projected visual tokens are inserted, how the visual prefix differs from ordinary text context, and why later assistant tokens can use the visual information.
3. What changes between LLaVA Stage 1 feature alignment and Stage 2 visual instruction tuning? For each stage, explain the training examples, conditioning inputs, supervised target tokens, trainable parameters, and behavior the model is expected to learn.

Adapter-Style Vision-Language Models: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. A Vision Transformer already produces vector embeddings. Why can these embeddings not always be inserted unchanged into a decoder-only language model? Discuss embedding width, sequence length, representation compatibility, and preservation of spatial information.
2. Describe the path followed by an image from the input image to the LLaVA language model. Explain how the vision encoder creates patch features, how the projector transforms them, where the projected visual tokens are inserted, how the visual prefix differs from ordinary text context, and why later assistant tokens can use the visual information.
3. What changes between LLaVA Stage 1 feature alignment and Stage 2 visual instruction tuning? For each stage, explain the training examples, conditioning inputs, supervised target tokens, trainable parameters, and behavior the model is expected to learn.
4. Suppose full visual instruction tuning produces a much higher evaluation score than feature alignment alone. What does this ablation demonstrate? Explain what each stage teaches, why caption generation does not guarantee diverse visual instruction following, what conclusion the performance gap supports, and what the ablation does not prove by itself.

Adapter-Style Vision-Language Models: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. A Vision Transformer already produces vector embeddings. Why can these embeddings not always be inserted unchanged into a decoder-only language model? Discuss embedding width, sequence length, representation compatibility, and preservation of spatial information.
2. Describe the path followed by an image from the input image to the LLaVA language model. Explain how the vision encoder creates patch features, how the projector transforms them, where the projected visual tokens are inserted, how the visual prefix differs from ordinary text context, and why later assistant tokens can use the visual information.
3. What changes between LLaVA Stage 1 feature alignment and Stage 2 visual instruction tuning? For each stage, explain the training examples, conditioning inputs, supervised target tokens, trainable parameters, and behavior the model is expected to learn.
4. Suppose full visual instruction tuning produces a much higher evaluation score than feature alignment alone. What does this ablation demonstrate? Explain what each stage teaches, why caption generation does not guarantee diverse visual instruction following, what conclusion the performance gap supports, and what the ablation does not prove by itself.
5. LLaVA places projected visual tokens before the user prompt and assistant answer in a decoder-only sequence. Explain why assistant tokens can attend to the earlier visual tokens, how the causal attention mask operates, why visual-token and user-prompt positions are normally excluded from the supervised loss, and how the answer-token loss can be written mathematically.