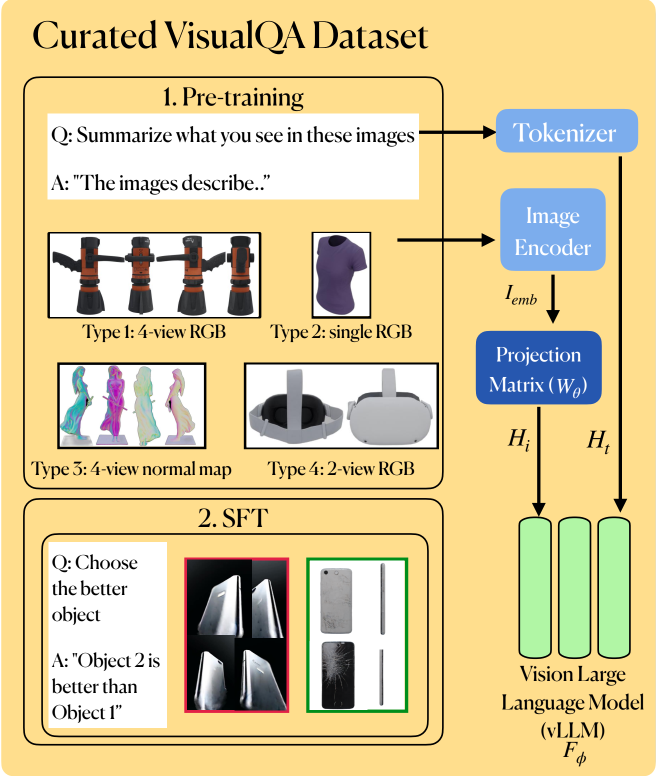
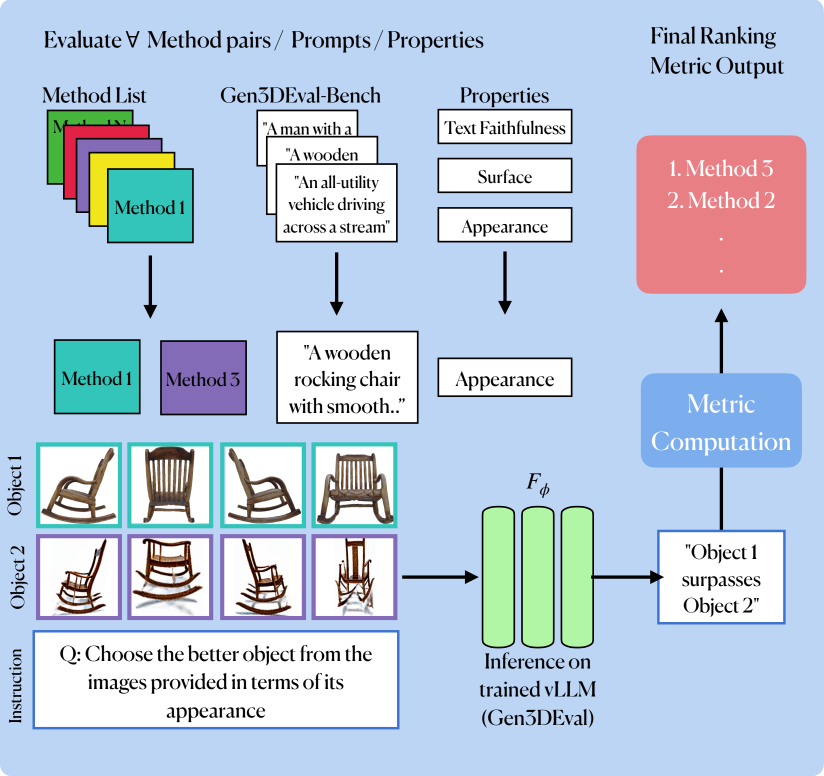

Curated VisualQA Dataset
1. Pre-training
Q: Summarize what you see in these images
A: "The images describe.."
Type 1: 4-view RGB
Type 2: single RGB
Type 3: 4-view normal map
Type 4: 2-view RGB
Tokenizer
Image Encoder
I_{emb}
Projection Matrix (W_θ)
H_i
H_t
2. SFT
Q: Choose the better object
A: "Object 2 is better than Object 1"
Vision Large Language Model (vLLM)
F_ϕ

Stage 1: Train the vLLM (Gen3DEval) in two stages - pre-training and supervised fine-tuning (SFT).


Evaluate ∀ Method pairs / Prompts / Properties
Final Ranking Metric Output
Method List
Method 1
Gen3DEval-Bench
"A man with a
"A wooden
"An all-utility vehicle driving across a stream"
Properties
Text Faithfulness
Surface
Appearance
1. Method 3
2. Method 2
.
.
Method 1
Method 3
"A wooden rocking chair with smooth.."
Appearance
Metric Computation
Object 1
Object 2
F_ϕ
Instruction
Q: Choose the better object from the images provided in terms of its appearance
Inference on trained vLLM (Gen3DEval)
"Object 1 surpasses Object 2"

Stage 2: Use the trained vLLM (Gen3DEval) to generate a ranking metric.