Dataset Collection

Real Videos

Synthetic Generation

QA Pairs Generation

Human Annotated QA

```
{
  "Question": "From the perspective of the player at
  bat, in which direction did the baseball go after
  bunting it?",
  "Answer": "To the left and away."
}
```

LLM Gen. MC

```
{
  "A": "To the left and away.", "C": "Not enough information.",
  "B": "To the right and away." "D": "Behind",
  "FF": "The baseball traveled to the left and away."
}
```

Human-in-the-loop Quality Review

✗ Ambiguous Videos

⚒ Ambiguous QA

⚒ Misannotations

⚒ Temporal Alignment

VLM4D-Bench

Dataset   VLM   Score

Real: 100
Synth: 100
―――――――
Total: 100