# Solution for Multiple-Choice VideoQA in the ECCV 2026 Perception Test Challenge

**Yuhan Zheng**[1] **Fengqiang Wan**[1] **Yang Yang**[1]
[1]Nanjing University of Science and Technology, Nanjing, China

**Abstract**

We present a multi-model ensemble approach for multiple-choice video question answering in the ECCV 2026 Perception Test Challenge. Our framework combines a fine-tuned Qwen3-VL-32B-Instruct model with two state-of-the-art commercial multimodal APIs, Doubao-Seed-2.0-Pro and Gemini 3.1 Pro Preview, to exploit their complementary capabilities in video understanding and multimodal reasoning. For model fine-tuning, we employ curriculum learning with hard-example mining to improve performance on challenging question types, and apply option shuffling augmentation to mitigate positional bias in multiple-choice prediction. For API-based inference, we design question-type-specific prompting strategies to improve reasoning over visual, temporal, spatial, and semantic information. We further integrate predictions from the fine-tuned and API-based models through a strategic ensemble framework, allowing different models to contribute according to their strengths across different question categories. Our final system achieves **91.40%** accuracy on the challenge test set, demonstrating the effectiveness of combining targeted fine-tuning, specialized prompting, and complementary multimodal models for robust video question answering.

## 1 Introduction

Multiple-choice video question answering (VideoQA) requires joint reasoning over visual appearance, temporal dynamics, spatial relations, object interactions, and semantics. The ECCV 2026 Perception Test Challenge evaluates these capabilities across diverse video reasoning scenarios.

Despite the strong performance of recent multimodal large language models (MLLMs), a single model often performs unevenly across question types. Moreover, fine-tuned models may suffer from option-position bias, while challenging temporal, spatial, and occlusion questions often require explicit reasoning.

We therefore build a hybrid framework combining a fine-tuned Qwen3-VL-32B-Instruct model with two commercial multimodal APIs, Doubao-Seed-2.0-Pro and Gemini 3.1 Pro Preview. Our main components are: (1) curriculum learning with hard-example mining; (2) option shuffling augmentation; (3) question-specific prompting and structured reasoning; and (4) validation-guided multi-model ensemble. The final system achieves 91.40% accuracy on the challenge test set.

## 2 Method

Our framework consists of three components: (1) fine-tuning Qwen3-VL-32B-Instruct; (2) API-based inference with Doubao-Seed-2.0-Pro and Gemini 3.1 Pro Preview; and (3) question-aware ensemble prediction.

### 2.1 Fine-Tuning Qwen3-VL

#### 2.1.1 Base Training

We fine-tune Qwen3-VL-32B-Instruct on the complete training set using instruction tuning. Each sample contains a video, a question, candidate answers, and the correct option.

```
{
  "messages": [
    {
      "role": "user",
      "content":
      "<video> {question}\n"
      "A. {opt1}\nB. {opt2}\nC. {opt3}\n"
      "D. {opt4}\nE. {opt5}"
    },
    {
      "role": "assistant",
      "content": "{letter}. {answer}"
    }
  ],
  "videos": ["{video_path}"] }
```

#### 2.1.2 Curriculum Learning

After base training, we evaluate performance across question types and identify weak categories:

$$T_{\text{hard}} = \{t : \text{Acc}(M_0, t) < \text{Acc}_{\text{avg}}(M_0)\}. \quad (1)$$

Training samples belonging to these categories form a hard subset $D_{\text{hard}}$. We then further fine-tune the base model:

$$M_{\text{final}} = \text{FineTune}(M_0, D_{\text{hard}}). \quad (2)$$

A smaller learning rate is used in this stage to improve difficult categories while preserving general capabilities.

#### 2.1.3 Option Shuffling

To reduce option-position bias, we randomly permute the $K$ candidate answers during training. Given options $\mathcal{O} = \{o_1, \ldots, o_K\}$ and correct index $i$, we sample a permutation $\pi$ and obtain

$$\mathcal{O}' = \{o_{\pi(1)}, \ldots, o_{\pi(K)}\}, \qquad i' = \pi^{-1}(i). \quad (3)$$

The model is trained using the shuffled options and updated target index $i'$.

### 2.2 API-Based Multimodal Inference

We additionally use two strong commercial MLLMs:

- **Doubao-Seed-2.0-Pro**, used for strong visual, spatial, and temporal reasoning;
- **Gemini 3.1 Pro Preview**, used for long-context video understanding and complex multimodal reasoning.

Their predictions complement the fine-tuned Qwen3-VL model and improve robustness across heterogeneous question categories.

#### 2.2.1 Question-Specific Prompting

We use lightweight question-specific instructions instead of a single generic prompt.

**Counting.** The model is instructed to inspect the entire video and count each valid occurrence independently.

**Temporal Reasoning.** The model first identifies relevant events and orders them chronologically before answering.

**Spatial Relations.** The model tracks objects relative to stable references and explicitly considers left/right, front/back, and above/below relations.

**Camera Motion.** The model selects a fixed reference object and distinguishes camera motion from foreground object motion.

#### 2.2.2 Two-Stage Occlusion Reasoning

For occlusion and viewpoint questions, we decompose reasoning into two stages.

**Stage 1: Visual extraction.** The model determines the relevant objects, viewpoint, target trajectory, and final position from the camera perspective.

**Stage 2: Viewpoint transformation.** The extracted position is transformed according to the requested viewpoint:

$$\text{Position}_{\text{target}} = \mathcal{T}\left(\text{Position}_{\text{camera}}, v\right), \quad (4)$$

where $v$ denotes the estimated viewpoint.

#### 2.2.3 Test-Time Augmentation

For selected difficult samples, we evaluate multiple temporal sampling or resolution variants. For model $m$, the final TTA prediction is obtained by majority voting:

$$\hat{y}_m = \arg\max_y \sum_{a \in \mathcal{A}} \mathbb{I}\left[f_m(x_a) = y\right]. \quad (5)$$

### 2.3 Question-Aware Multi-Model Ensemble

We categorize each question into coarse types such as counting, temporal, spatial, occlusion, camera motion, and general reasoning.

Let

$$\mathcal{M} = \{M_{\text{Qwen}}, M_{\text{Doubao}}, M_{\text{Gemini}}\}.$$

For question type $t$, each model is assigned a weight based on its validation accuracy:

$$w_{m,t} = \frac{\exp\left(\text{Acc}_{m,t}\,/\tau\right)}{\sum_{j \in \mathcal{M}} \exp\left(\text{Acc}_{j,t}\,/\tau\right)}. \quad (6)$$

The final prediction is obtained by weighted voting:

$$\hat{y} = \arg\max_y \sum_{m \in \mathcal{M}} w_{m,t} \mathbb{I}[y_m = y]. \quad (7)$$

This strategy exploits the complementary strengths of Qwen3-VL-32B-Instruct, Doubao-Seed-2.0-Pro, and Gemini 3.1 Pro Preview. The final ensemble achieves 91.40% accuracy on the ECCV 2026 Perception Test Challenge test set.

## 3 Experiments

### 3.1 Dataset and Metric

We follow the official training, validation, and test splits of the ECCV 2026 Perception Test Challenge. The task consists of multiple-choice VideoQA problems covering diverse visual, temporal, spatial, and semantic reasoning scenarios. We use accuracy (%) as the evaluation metric.

### 3.2 Implementation Details

**Fine-Tuning.** We fine-tune Qwen3-VL-32B-Instruct using MS-Swift. The base training uses a learning rate of $1 \times 10^{-4}$, followed by curriculum learning with a reduced learning rate of $5 \times 10^{-5}$. Option shuffling is applied with a probability of 0.8 to mitigate positional bias.

**API Inference.** We use Doubao-Seed-2.0-Pro and Gemini 3.1 Pro Preview for API-based multimodal inference. Question-specific prompts and structured reasoning are applied according to the question category. For difficult samples, test-time augmentation is additionally employed to improve prediction robustness.

**Parameter Settings.** For Qwen3-VL fine-tuning, we use the AdamW optimizer with a global batch size of 8 and train for 2,000 iterations in the base stage, followed by 1,000 iterations on the hard-example subset. For API inference, the temperature is set to 0.1 and the maximum output length is limited to 4,096 tokens. For test-time augmentation, we use multiple temporal sampling offsets and input resolutions, and aggregate the resulting predictions by majority voting.

### 3.3 Main Results

Table 1 summarizes the models used in our final system. The three models provide complementary capabilities, and their predictions are integrated using the validation-guided ensemble described in Section 6.

The final ensemble achieves 91.40% test accuracy. The improvement benefits from the complementary strengths of the three models, together with targeted fine-tuning, question-specific prompting, and validation-guided ensemble prediction.

Table 1: Performance of our final system on the ECCV 2026 Perception Test Challenge.

| **Method** | **Test Acc.** |
|---|---|
| Qwen3-VL-32B-Instruct | 81.62% |
| Doubao-Seed-2.0-Pro | 78.43% |
| Gemini 3.1 Pro Preview | 82.56% |
| **Ensemble (Ours)** | **91.40%** |

### 3.4 Qualitative Analysis

We observe that structured reasoning is particularly helpful for occlusion, spatial, and temporal questions. Decomposing complex problems into visual information extraction and subsequent reasoning reduces errors caused by viewpoint ambiguity or incomplete temporal tracking. Nevertheless, repeated-action counting remains challenging, as models may occasionally miss short or visually similar events.

## 4 Conclusion

We present a hybrid multi-model framework for multiple-choice VideoQA in the ECCV 2026 Perception Test Challenge. Our approach combines curriculum-based fine-tuning and option shuffling for Qwen3-VL-32B-Instruct with API-based inference from Doubao-Seed-2.0-Pro and Gemini 3.1 Pro Preview. Question-specific prompting, structured reasoning, and validation-guided ensemble prediction further improve robustness across diverse question types.

Our final system achieves 91.40% accuracy on the challenge test set, demonstrating the effectiveness of combining complementary multimodal models for video question answering.