Caption-Conditioned Temporal Frame Ordering

SNU AI Challenge Final Report

Lee Jae-Woon Chang Ji-Ho
Seoul National University, Team Mallangboksoonga

Abstract

We study caption-conditioned temporal ordering of four shuffled video frames, a 24-way permutation problem, under constraints of a single model, fully offline inference, and one NVIDIA RTX 3090 GPU with 24 GB of memory. Our system fine-tunes **Qwen3.6-27B using 4-bit QLoRA** and direct permutation generation. Because 48 of the backbone's 64 language-model layers use gated-delta linear attention, we apply LoRA to both conventional full-attention projections and the corresponding linear-attention projections while keeping the vision encoder frozen. At inference, perm-TTA@8 selects eight diverse display orders by farthest-first traversal under Kendall distance, maps every prediction back to the original frame indices, and aggregates them by plurality vote. The final system achieves Kaggle exact-match scores of 0.93542 on the Public Leaderboard and **0.92276 on the Private Leaderboard**; on our internal validation set, perm-TTA@8 improves exact match from 0.5903 to 0.6218, a gain of 3.15 percentage points. Teacher-forced scoring shows that the maximum-a-posteriori candidate nearly always matches greedy decoding, even on errors, so reranking and calibration are not the primary limitation. Of the 143 validation errors, 40.6% are single transpositions, and 67% of those swap temporally adjacent positions; the swapped frame pairs are also more visually similar than the overall pair baseline. These results identify **fine-grained visual temporal perception** as the principal bottleneck and motivate vision-tower adaptation, explicit temporal inputs, and higher-resolution processing.

1 Introduction

The task is to infer the temporal order of four shuffled frames using both visual evidence and a natural-language caption. Performance depends on two capabilities: a strong backbone that can distinguish subtle before–after relationships, and an inference procedure that can reduce individual ordering errors by combining multiple predictions. The competition constraints—a single model, a 24 GB GPU, offline execution, and no model ensemble—make both capabilities difficult to obtain simultaneously. Figure 1 summarizes the resulting system.

Our solution makes three main contributions:

(1) The largest practical single backbone under a 24 GB budget. A 27-billion-parameter model does not ordinarily fit on one 24 GB GPU. We use 4-bit NF4 quantization to reduce its memory footprint and load Qwen3.6-27B on a single RTX 3090. Its dense architecture activates all parameters during inference while remaining feasible after quantization; similarly sized mixture-of-experts alternatives did not satisfy the memory constraint.

(2) Architecture-aware LoRA layout. Qwen3.6 interleaves standard full-attention layers with gated-delta linear-attention layers. In this 64-layer backbone, 48 layers (75%) are linear-attention and only 16 are full-attention, so a conventional `q_proj/k_proj/v_proj/o_proj` layout would adapt the attention pathway of just those 16 layers and leave the attention of

the other 48 unadapted. We additionally adapt the linear-attention projections (`in_proj_qkv`, `in_proj_a`, `in_proj_b`, `in_proj_z`, `out_proj`); these account for 31.7% of the adapter's trainable parameters (74.1 M of 233.5 M) and are what let the 75% of layers that use linear attention receive attention-side adaptation.

(3) Constraint-compatible multi-view voting. We construct eight views by changing only the displayed order of the four frames. The views are selected to be mutually distant under Kendall distance. Display-order changes redistribute positional bias across frames, reducing error correlation between views. After mapping all predictions back to the original frame indices, we select the most frequent permutation.

2 Model and Training

2.1 Direct permutation generation

Two natural formulations are possible: score all 24 candidate permutations, or generate the answer directly. We adopt direct generation. The model receives the caption and four 640-pixel images labeled according to their displayed positions ("1", "2", "3", and "4"). It is trained to emit the chronological order as four index tokens, `p1 p2 p3 p4`. **The loss is applied only to these answer tokens**; prompt and image tokens are masked. This focuses supervision on the ordering decision rather than on reproducing the input format.

We also tested chain-of-thought targets that verbalized the reasoning before the answer. They reduced accuracy in this short perceptual task, so the final model generates only the permutation.

2.2 QLoRA configuration

The final model is checkpoint 600, corresponding to approximately 1.06 epochs of a planned two-epoch run. **The image encoder remains frozen, and only the language model is adapted.** Table 1 summarizes the training setup.

The training loss decreased steadily from 0.238 to 0.125 by step 600. The exact seed-42 training and validation split is included in the release bundle for reproducibility. Variants that also adapted the image encoder were evaluated but were not selected for the final submission.

3 Inference

For each example, perm-TTA@8 proceeds as follows:

1. Prepare eight display orders. The first is the original input order; the remaining seven are selected by farthest-first traversal to maximize their Kendall distance from the original order and from one another.
2. For every view, provide the caption and reordered frames to the model and greedily generate four answer indices, with a maximum of 16 new tokens.
3. Convert each output from the current display coordinates back to the original frame indices. Without this **inverse mapping**, predictions from different views are not comparable.
4. Choose the most frequent mapped permutation by **plurality vote**. Ties are resolved by the earliest occurrence among the eight views.

The eight views are evaluated in two batches of four using left padding. Although vLLM would normally accelerate repeated inference, it does not support LoRA on the linear-attention modules of this backbone. We therefore use Transformers with bitsandbytes 4-bit loading and PEFT, completing the full test set on one RTX 3090 within 24 hours.

4 Results and Error Analysis

4.1 Main results

The internal validation split contains a comparatively high proportion of sparse, short captions, whereas the leaderboard distribution contains mostly caption-rich examples. This distribution

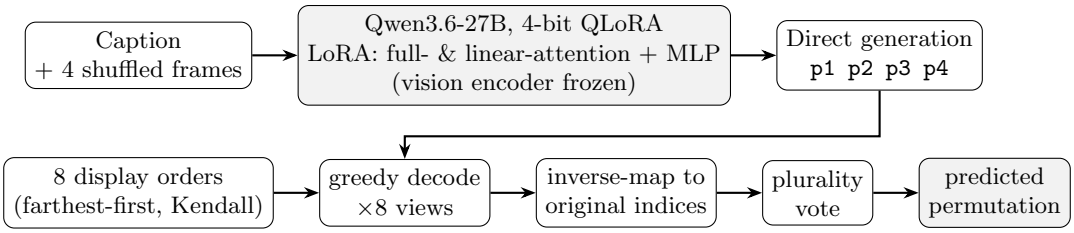

Figure 1: End-to-end pipeline. A single Qwen3.6-27B backbone (4-bit QLoRA, adapting both full- and linear-attention projections; the vision encoder is frozen) generates an order directly. At test time, perm-TTA@8 runs the same model on eight farthest-first display permutations and aggregates the inverse-mapped predictions by plurality vote.

Table 1: Final training configuration.

Component	Setting
Backbone	Qwen3.6-27B, 4-bit NF4
LoRA	$r = 32$, $\alpha = 64$, dropout $= 0.05$
Trainable component	Language model only; frozen image encoder
Micro-batch / accumulation	2 / 8 (effective batch size 16)
Learning rate	5×10^{-5}, cosine decay, 3% warm-up
Maximum sequence length	2,560 tokens
Image resolution	640 px
Data augmentation	None
Random seed	42
Selected checkpoint	Step 600 (approximately 1.06 epochs)

difference explains the lower absolute validation score. The controlled comparison at checkpoint 600 remains informative: perm-TTA@8 improves validation exact match by **3.15 percentage points** over greedy decoding. It is the only test-time modification used in the final submission.

4.2 A perceptual bottleneck

Selection is not the bottleneck. We scored all 24 candidate orderings with teacher-forced log probabilities and took the maximum-a-posteriori (MAP) candidate. The MAP candidate was almost always identical to the greedy output, including on incorrect examples. Thus the model assigns its highest probability to the wrong order even when the correct order is among the candidates; simple reranking or probability calibration over the candidate set cannot repair these cases.

Error structure. We analyzed the checkpoint-600 validation predictions (382 examples, 143 errors). For each error we computed the Kendall-tau distance between the predicted and gold orderings and isolated *single transpositions* — errors in which exactly one pair of positions is swapped while the rest of the sequence is correct. Single transpositions form the largest error category, 40.6% (58/143). Among them, swaps of *temporally adjacent* positions dominate: positions 1–2, 2–3, and 3–4 together account for 67% of single transpositions, while the non-adjacent 1–3 swap accounts for only 10%. The model thus preserves the overall event structure but most often confuses the order of temporally neighboring events.

Table 2: Exact-match performance of the final checkpoint. A dash denotes an unsubmitted configuration.

Inference method	Internal validation EM	Public LB EM	Private LB EM
Greedy decoding	0.5903	0.92495	–
perm-TTA@8	**0.6218**	**0.93542**	**0.92276**

Swapped frames are visually harder to distinguish. To test whether these confusions are visually grounded, for each single-transposition error we measured the mean absolute pixel difference between the two frames whose order was swapped (64×64 grayscale, normalized to $[0, 1]$) and compared it against the same measure computed over all six frame pairs of the erroneous examples. Swapped pairs are more similar than the baseline (median difference 0.177 vs. 0.193; 59% of swapped pairs fall below the overall-pair median). The effect is modest but consistent: erroneous orderings concentrate on frame pairs that are harder to separate visually.

Interpretation. The three measurements agree: MAP equals greedy, adjacent-position confusions dominate, and swapped frames are visually more similar. Together, these findings point to visual temporal perception, rather than candidate selection, as the limiting factor. Perm-TTA partially mitigates the errors by decorrelating positional biases across input views, but it does not remove the underlying perceptual ambiguity.

5 Alternative Approaches

The diagnosis in Section 4 predicts which interventions should help: those that strengthen visual perception, not those that rerank or reaggregate a fixed candidate set. We tested this prediction across seven families of methods; Table 3 summarizes each hypothesis and its outcome. Metrics come from different checkpoints and subsets and should be read within each row ("LB": Kaggle Public Leaderboard; "proxy": exact match on a caption-rich subset; "val": internal split).

Two findings are decisive. First, reranking and aggregation cannot beat plurality voting because the MAP candidate coincides with greedy decoding even on errors, and language-space preference optimization only degrades greedy accuracy through likelihood displacement—both confirm that candidate selection is not the bottleneck. Second, every intervention that left the visual pathway unchanged failed, while the single lever that helped (input-permutation TTA) works by decorrelating errors rather than by improving perception. The one family the diagnosis marks as promising—directly strengthening perception—is also the one we could not fully realize within the time and compute budget, which is exactly where Section 6 locates the remaining gap.

6 Strengths and Limitations

Strengths. The system satisfies all competition constraints and is reproducible: a single 27B model runs in 4-bit precision on a 24 GB RTX 3090, without vLLM, and processes the test set within 24 hours. It uses no model ensemble; the only multi-view component is rule-compliant test-time augmentation. Permutation TTA also addresses display-order bias without fitting a separate aggregator to the leaderboard distribution.

Limitations. The final Private Leaderboard score is 0.92276, compared with 0.93542 on the Public Leaderboard, a decrease of 1.266 percentage points. This public–private gap indicates that some ordering errors persist under the held-out evaluation distribution. Our experiments consistently locate the main limitation in **visual perception rather than in prob-**

Table 3: Alternative approaches grouped by hypothesis. Metrics are within-experiment (see text); the final system scores 0.93542 on the Public LB and 0.92276 on the Private LB.

Hypothesis	Methods tested	Result	Verdict
Better candidate selection / aggregation	log-prob MAP, Kemeny median, Borda, log-prob weighting, MBR, checkpoint soup	selection/soup 0.92844 LB	MAP ≡ greedy; selection is not the bottleneck
Preference / ranking fine-tuning	CPO/DPO objectives; reversed, random, and adjacent-transposition negatives; warm start	MAP still ≡ greedy	likelihood displacement; ineffective in language space
Stronger visual perception	vision-tower LoRA, temporal-order aux. loss, video input (M-RoPE + frame replication), 896 px	long training degraded	promising per diagnosis but unrealized (future work)
Larger / other backbone	InternVL3-8B (frozen ViT), 32B direct generation	0.888 / 0.819	no gain from frozen swap or scale
Verbose reasoning targets	text-only CoT, distilled vision-CoT	0.612 / 0.741	reasoning hurts a short perceptual task
Alternative input representation	2 × 2 grid, pairwise + min-inversion sort, optical-flow, clause–frame (SigLIP2)	0.42 / 0.10 / 0.22	no usable temporal signal
Test-time diversity	temperature ensemble; input-permutation TTA; dual-TTA	perm-TTA adopted; dual-TTA 0.913 (8B)	input-level decorrelation is the only lever

ability calibration or aggregation. The final submission is consequently a conservative combination of language-model-only adaptation and perm-TTA. Closing the remaining gap will likely require direct improvement of the visual pathway through vision-tower training, explicit temporal encoding, or higher-resolution inputs.

7 Conclusion

We fit Qwen3.6-27B on the competition hardware using 4-bit QLoRA, adapting both standard and hybrid linear-attention modules, and train it to generate frame permutations directly. Farthest-first perm-TTA@8 with plurality voting reduces confident ordering errors and yields Kaggle exact-match scores of 0.93542 on the Public Leaderboard and **0.92276 on the Private Leaderboard**. Beyond the scores themselves, the experiments establish that the principal bottleneck is fine-grained visual perception. This diagnosis narrows future work to vision-tower adaptation, explicit temporal inputs, and higher-resolution processing rather than further language-space reranking.