# A Principled Optimal Transport Framework for Frame Selection in Long Video Understanding

**Ziwei Zhao**
Yizhun Medical AI Co., Ltd
Beijing, China
zhaozw998@gmail.com

**Yuhang Liu**
Yizhun Medical AI Co., Ltd
Beijing, China
yuhang.liu@yizhun-ai.com

## Abstract

Long-video understanding with multimodal LLMs is fundamentally constrained by the mismatch between video length and the model's limited visual context budget, making frame selection essential for practical reasoning. Existing query-aware methods typically rely on heuristic pipelines: they assign one-dimensional query-conditioned scores to frames and then apply sampling, ranking, or extraction rules to form the final subset. While effective in practice, such methods are generally not derived from an explicit optimization objective and struggle to capture the heterogeneous evidence required for complex reasoning. In this paper, we revisit long-video frame selection from a principled optimization perspective. We formulate it as a structured evidence-allocation problem and propose **OTFS**, an optimal transport framework that maps multiple evidence sources onto video frames. To reflect the asymmetry of this problem—important evidence should be preserved, whereas redundant frames need not absorb mass—we instantiate this view with a semi-unbalanced entropic optimal transport objective and an efficient solver, *OTFS-Sinkhorn*. Combined with an exact dynamic program for coverage-aware subset extraction, OTFS yields a practical two-stage training-free pipeline. We further show that, in the single-source case, our formulation induces a Gibbs-type distribution over frames, making Q-Frame's temperature-scaled distribution a special case of our framework while also providing a broader perspective for understanding related prior methods. Experiments on three long-video understanding benchmarks show that OTFS consistently outperforms strong training-free baselines. Codes are available at https://github.com/barry664/OTFS.

## 1 Introduction

A central challenge in long-video understanding with multimodal large language models (MLLMs) is that long videos contain far more visual content than current models can process within their limited visual context budget [23, 18, 3, 13, 48, 1, 46, 19, 28, 54, 44, 35, 30]. This mismatch between video length and model capacity makes visual compression necessary in practical systems [32, 21, 39, 27, 29, 41, 16, 36, 50, 52]. Among existing compression strategies, frame selection—that is, choosing a small subset of frames from the original video—has become a widely used approach. Because only the selected frames remain available to the downstream reasoner, frame selection largely determines which visual evidence can be preserved for long-video question answering and temporal reasoning.

Existing query-aware frame selection methods have substantially improved over uniform sampling, yet most still rely on heuristic frame-scoring and subset-construction procedures [34, 49, 53, 14]. As illustrated at the top of Figure 1, they typically derive a one-dimensional query-conditioned score over frames and then apply hand-designed sampling or extraction rules to form the final subset. Although Q-Frame [49], AKS [34], and DIG [14] differ in their scoring signals and extraction strategies, **they**

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

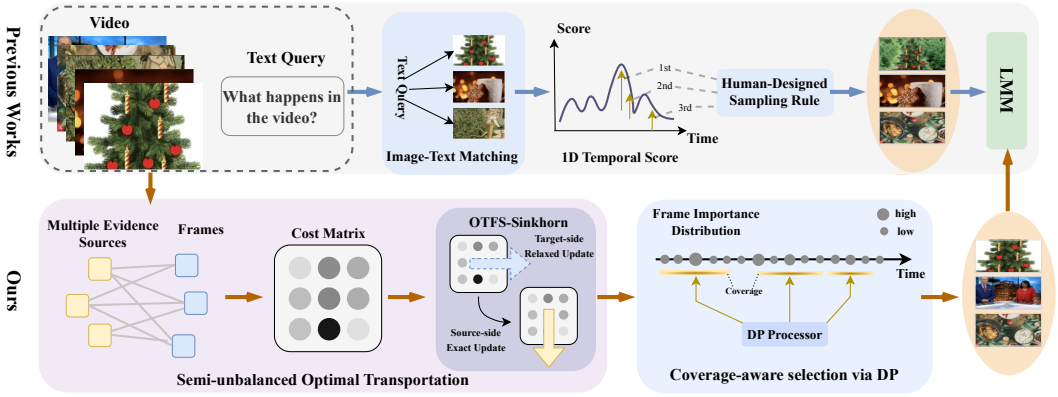



Figure 1: Comparison between prior training-free frame-selection procedures and our principled OT-based framework. Existing methods typically compute one-dimensional frame scores and then apply hand-designed sampling or extraction rules. In contrast, OTFS formulates frame selection as semi-unbalanced transport from evidence atoms to video frames, producing a soft frame-importance distribution that is converted into a coverage-aware subset by exact dynamic programming.

**are generally not derived as exact solvers of a clearly specified frame-selection objective**. As a result, it remains difficult to reason about optimality under an explicit mathematical target.

This limitation becomes more pronounced in complex long-video reasoning, where answering a question may require multiple types of visual evidence, such as direct semantic relevance, answer-oriented utility, and temporal continuity cues [34, 14]. In this setting, what matters is not only how important each frame appears in isolation, but how a limited frame budget should be allocated so that the selected subset collectively preserves the heterogeneous evidence required for downstream reasoning. This suggests viewing frame selection not merely as heuristic frame ranking, but as a structured evidence-allocation problem under budget constraints.

In this paper, we revisit long-video frame selection from a principled optimization perspective and propose **OTFS**, an **O**ptimal **T**ransport framework for **F**rame **S**election. Our key idea is to formulate frame selection as transporting evidence requirements onto video frames. Concretely, we represent query- or visually-induced cues as *evidence atoms* on the source side and video frames as candidate locations on the target side. The resulting transport plan captures how different evidence atoms allocate support to different frames, and its target marginal serves as a soft frame-importance distribution. In this way, frame importance is grounded in an explicit transport objective rather than query-aware scores alone.

A key observation is that this allocation problem is intrinsically asymmetric. On the source side, important evidence should not be omitted, since missing a crucial cue may directly harm downstream reasoning. On the target side, however, long videos contain many redundant or weakly relevant frames, which need not absorb mass. This asymmetry makes *semi-unbalanced entropic optimal transport* a natural fit [5, 4]: it preserves source-side evidence requirements while softly relaxing the target marginal, allowing mass to concentrate on evidence-bearing frames without artificially spreading it over redundant background content.

Based on this formulation, we instantiate a practical two-stage training-free pipeline, illustrated at the bottom of Figure 1. In the first stage, we specialize diagonal-scaling updates for entropic semi-unbalanced OT [6, 4, 26] to derive *OTFS-Sinkhorn*, which computes a globally consistent soft frame-importance distribution. In the second stage, we apply an exact dynamic program (DP) to extract a size-$k$ subset that optimizes the induced coverage-aware selection objective. Our contribution is not a new general theory of semi-unbalanced optimal transport itself. Rather, we show that long-video frame selection admits a principled transport-based formulation, and that this view leads to a clear and effective two-stage optimization pipeline.

Our formulation also clarifies the relationship to prior methods. In particular, when the source side collapses to a single evidence source, OTFS induces a Gibbs-type distribution over frames, making Q-Frame [49]'s temperature-scaled distribution a one-dimensional special case of our framework.

In contrast, AKS [34] and DIG [14] are not exact instances of our objective; they are better understood as related staged approaches to budgeted frame selection. Experiments on three long-video understanding benchmarks show that OTFS consistently outperforms strong baselines.

Our main contributions are summarized as follows:

1. We revisit long-video frame selection from a principled optimization perspective and formulate it as a structured evidence-allocation problem under limited visual budget.
2. We instantiate this view with a semi-unbalanced optimal transport objective and develop a practical two-stage training-free pipeline consisting of OTFS-Sinkhorn for soft evidence allocation and an exact dynamic program for coverage-aware subset extraction.
3. We show both analytically and empirically that this formulation is useful: in the single-source case, it induces a Gibbs-type distribution over frames, making Q-Frame's temperature-scaled distribution a direct one-dimensional special case of our framework; on long-video benchmarks, the full OTFS pipeline consistently outperforms strong baselines.

## 2 Related Work

**Video MLLMs and long-video modeling.** Recent MLLMs have extended image-language models to video inputs and achieved strong progress in video understanding [18, 3, 13, 2, 44, 52]. Long videos, however, remain difficult because their visual token length often exceeds the effective context budget of current models. Existing methods address this limitation through long-context training or transfer [48, 21], memory or streaming mechanisms [32, 27], compact and hierarchical video representations [10, 39, 29, 16, 41, 31], redundancy reduction [36], or retrieval-, tree-, and agent-style reasoning [37, 38]. These approaches mainly improve how video MLLMs process or organize long inputs. OTFS instead focuses on a complementary question: given a fixed downstream MLLM and a limited visual budget, how should frames be selected in a principled way?

**Query-aware frame selection.** More closely related to our setting, recent work studies selecting a small query-conditioned subset of frames before downstream reasoning. AKS [34] combines query relevance with temporal coverage, Q-Frame [49] samples from a temperature-scaled score distribution, FOCUS [53] formulates training-free keyframe search, and DIG [14] combines query-type-dependent routing with reward-based utility estimation. Other recent related methods either perform inference-time query-aware frame selection or temporal search [33, 22, 42, 17, 43], or enhance video question answering through query-conditioned visual encoding, modular grounding, caption aggregation, or grid-based video representations [8, 24, 45, 12]. Most existing selectors still rely on scalar scoring and staged procedures, followed by top-$k$, sampling, routing, or hand-designed search rules. OTFS instead formulates frame selection as a principled evidence-allocation problem with an explicit optimization objective, and couples semi-unbalanced transport with exact coverage-aware subset extraction. This view also clarifies the relation to prior methods: Q-Frame [49] is the single-source limit of our formulation, while AKS [34] and DIG [14] are better understood as related staged selectors rather than exact solvers of a common objective.

**Optimal transport for long-video understanding.** Optimal transport (OT) has been explored for long-video modeling and compression. Lin [20] uses OT to mitigate noisy clip–caption and frame–word correspondences during video–language pre-training. OTT-Vid [11] applies OT between tokens in neighboring frames to estimate temporal compressibility and allocate token-level compression budgets. In contrast, OTFS employs semi-unbalanced OT to integrate heterogeneous evidence into query-aware frame importance.

## 3 Method

### 3.1 Overview

Our goal is to select a subset of $k$ frames from a long video that preserves the visual evidence most useful for answering the input query. We refer to the proposed method as **OTFS**. The method consists of two optimization stages. First, we solve a semi-unbalanced optimal transport problem between heterogeneous evidence atoms induced by the query and auxiliary visual cues and the candidate video

frames, yielding a soft global frame-importance distribution. Second, we convert this distribution into a size-$k$ subset through an exact dynamic program that accounts for temporal coverage.

This decomposition follows the structure of the problem: grounding heterogeneous evidence is a global allocation problem over all frames, whereas enforcing a fixed frame budget is a discrete subset optimization problem. OTFS handles these two roles separately through semi-unbalanced transport and exact combinatorial optimization.

### 3.2 Semi-Unbalanced Evidence Allocation Formulation

Let $q$ denote the input query and let $V = (x_1, \dots, x_M)$ be an ordered sequence of video frames. A frame selector outputs an index set

$$S \subseteq \{1, \dots, M\}, \qquad |S| = k, \tag{1}$$

whose corresponding frames are passed to the downstream MLLM. The objective is to preserve, under the budget $k$, the visual evidence required to answer $q$.

Rather than reducing the query to a single frame-score vector, we introduce a set of *evidence atoms*

$$\mathcal{Z} = \{z_1, \dots, z_N\}, \tag{2}$$

where each atom denotes an evidence requirement induced by the query or auxiliary visual cues. This abstraction allows heterogeneous evidence to be represented within a single optimization problem.

For each evidence atom $z_i$ and frame $x_j$, let $r_{ij} \in [0, 1]$ denote their compatibility, with larger values indicating stronger support. We convert compatibility into transport cost by

$$C_{ij} = 1 - r_{ij}, \tag{3}$$

which defines a cost matrix $C \in \mathbb{R}^{N \times M}$.

We treat the evidence atoms as the source domain and the video frames as the target domain. Let $\pi \in \mathbb{R}_+^{N \times M}$ be a transport plan, where $\pi_{ij}$ denotes how much mass of atom $z_i$ is assigned to frame $x_j$. Let $\mathbf{a} \in \Delta^N$ be a prior distribution over evidence atoms and $\mathbf{b} \in \Delta^M$ a prior distribution over frames, where $\Delta^d = \{u \in \mathbb{R}_+^d : \sum_\ell u_\ell = 1\}$ denotes the probability simplex.

We define the evidence-allocation objective as

$$\min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle - \varepsilon H(\pi) + \gamma \, \mathrm{KL}\big(\pi^\top \mathbf{1} \,\|\, \mathbf{b}\big) \quad \text{s.t.} \quad \pi \mathbf{1} = \mathbf{a}, \tag{4}$$

where $\langle C, \pi \rangle$ denotes the Frobenius inner product and

$$H(\pi) = -\sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij} (\log \pi_{ij} - 1) \tag{5}$$

is the entropy regularizer. Here, $\varepsilon > 0$ is the entropic regularization coefficient, and $\gamma$ controls the strength of the soft target-marginal regularization: larger $\gamma$ keeps the target marginal closer to $\mathbf{b}$, whereas smaller $\gamma$ allows mass to concentrate on evidence-supporting frames.

Equation (4) is a semi-unbalanced optimal transport formulation [6, 5, 4, 26]. The source marginal is enforced exactly, while the target marginal is only softly regularized. This asymmetry matches long-video frame selection: important evidence should not be discarded, whereas redundant or weakly relevant frames need not absorb mass. Given the optimal plan $\pi^\star$, we define the frame-importance distribution as the target marginal

$$\mathbf{w} = (\pi^\star)^\top \mathbf{1}, \qquad w_j = \sum_{i=1}^{N} \pi_{ij}^\star. \tag{6}$$

### 3.3 OTFS-Sinkhorn and the Single-Source Reduction

Eq. (4) instantiates entropic semi-unbalanced OT for the proposed evidence-allocation problem, where evidence atoms form the source side and video frames form the KL-relaxed target side. Specializing

the standard scaling characterization of semi-unbalanced OT [5, 4, 26] to this setting yields a simple solver, which we call OTFS-Sinkhorn.

Define the Gibbs kernel

$$K_{ij} = \exp(-C_{ij}/\varepsilon). \tag{7}$$

By standard results for entropically regularized semi-unbalanced OT, the optimizer admits the scaling form

$$\pi^\star = \mathrm{Diag}(\mathbf{u})\, K\, \mathrm{Diag}(\mathbf{v}), \tag{8}$$

for positive scaling vectors $\mathbf{u} \in \mathbb{R}^N_{++}$ and $\mathbf{v} \in \mathbb{R}^M_{++}$. Under this factorization, the scaling vectors satisfy

$$\mathbf{u} = \mathbf{a} \oslash (K\mathbf{v}), \tag{9}$$

$$\mathbf{v} = \left(\mathbf{b} \oslash (K^\top \mathbf{u})\right)^{\gamma/(\gamma+\varepsilon)}, \tag{10}$$

where $\oslash$ denotes element-wise division.

These updates define a simple fixed-point solver, which we refer to as *OTFS-Sinkhorn*. Starting from $\mathbf{v}^{(0)} = \mathbf{1}$, we iterate

$$\mathbf{u}^{(t)} \leftarrow \mathbf{a} \oslash (K\mathbf{v}^{(t-1)}), \tag{11}$$

$$\tilde{\mathbf{v}} \leftarrow \mathbf{b} \oslash (K^\top \mathbf{u}^{(t)}), \tag{12}$$

$$\mathbf{v}^{(t)} \leftarrow \tilde{\mathbf{v}}^{\,\gamma/(\gamma+\varepsilon)}, \tag{13}$$

until convergence, after which $\pi^\star$ is recovered through Eq. (8). Relative to balanced Sinkhorn, the distinctive term is the exponent $\gamma/(\gamma+\varepsilon)$ in Eq. (10), which arises from the KL-relaxed target marginal. A derivation specialized to our objective is provided in the appendix.

**Single-source Gibbs reduction.** The following special case is important for interpreting OTFS and relating it to prior query-aware selectors. Consider $N = 1$, with source marginal $a_1 = 1$, cost row $c_j = C_{1j}$, and a uniform frame prior $b_j = 1/M$ for all $j$. Then Eq. (4) reduces to

$$\min_{\pi \in \Delta^M} \sum_{j=1}^{M} \pi_j c_j + \varepsilon \sum_{j=1}^{M} \pi_j \log \pi_j + \gamma \sum_{j=1}^{M} \pi_j \log \frac{\pi_j}{b_j}, \tag{14}$$

whose unique solution is

$$\pi_j = \frac{\exp(-c_j/(\varepsilon+\gamma))}{\sum_{\ell=1}^{M} \exp(-c_\ell/(\varepsilon+\gamma))}. \tag{15}$$

The derivation is given in the appendix.

This result shows that one-dimensional frame scoring arises as a degenerate single-source limit of OTFS. In particular, the strategy in Q-Frame [49] is a direct one-dimensional special case: under a single evidence atom, a uniform frame prior, and a cost induced from the same scalar relevance signal, OTFS yields exactly the same Gibbs-form distribution as Q-Frame's temperature-scaled sampling rule. By contrast, AKS [34] and DIG [14] are not exact instances of Eq. (4), but can still be understood within the same broader view. AKS combines scalar query relevance with coverage-aware budget allocation, while DIG augments scalar or sparse utility-style signals with query-type-dependent routing and additional extraction logic. In this sense, Q-Frame corresponds to the single-source limit of OTFS, whereas AKS and DIG are better understood as related heuristic or staged procedures that preserve parts of the same query-aware, budget-aware intuition without being derived as exact solvers of a common optimization objective.

### 3.4 Instantiating the Evidence Atoms

The formulation itself is agnostic to how evidence atoms are constructed. In our experiments, we instantiate OTFS with three complementary evidence sources motivated by prior query-aware frame-selection methods [34, 49, 14]: semantic alignment to the query, answer-oriented utility, and temporal variation in the visual stream.

**Semantic relevance.** The first evidence source measures direct semantic alignment between the query $q$ and each frame $x_j$. Following prior query-aware selectors [34, 53], we use the image-text matching score from BLIP ITM [15] and define

$$r_j^{\text{sem}} = \text{BLIP}_{\text{ITM}}(q, x_j) \in [0, 1]. \tag{16}$$

**Utility reward.** We also include a reward-based utility signal inspired by DIG [14], which estimates the downstream usefulness of each frame for question answering. To keep the computation practical on long videos, we evaluate the reward only on a temporally sampled subset of frames and linearly interpolate the resulting scores to the remaining candidates. In practice, we follow the public DIG implementation to obtain the sampled frames and their corresponding reward values.

**Temporal transition.** The third evidence source captures temporal variation that may indicate event boundaries or visually salient state changes. Following DIG [14], we compute DINOv2 features [25] for adjacent frames and define

$$r_j^{\text{trans}} = 1 - \cos\big(\psi(x_j), \psi(x_{j-1})\big) \tag{17}$$

for $j \geq 2$, where $\psi(x_j)$ denotes the DINOv2 feature of frame $x_j$.

Using these three sources, we instantiate OTFS with $N = 3$ evidence atoms by setting

$$r_{1j} = r_j^{\text{sem}}, \qquad r_{2j} = r_j^{\text{util}}, \qquad r_{3j} = r_j^{\text{trans}}. \tag{18}$$

Before forming the transport objective, we normalize all three score sequences to $[0, 1]$ so that they are comparable across evidence sources. The resulting compatibility matrix $R = (r_{ij})$ is converted into the transport cost matrix $C$ through Eq. (3). More generally, OTFS can incorporate additional evidence atoms without changing the optimization framework itself.

### 3.5 Exact Coverage-Aware Budgeted Subset Extraction

As defined in Eq. (6), the optimal transport plan induces a frame-importance distribution $\mathbf{w}$. However, the final subset should not be determined by marginal importance alone. In long videos, high-importance frames often cluster around the same event, while downstream reasoning also benefits from sufficient temporal spread, as highlighted in [34]. To avoid the quadratic complexity of jointly optimizing evidence matching and temporal coverage in the framework of optimal transport, we adopt a principled two-stage relaxation. We treat subset extraction as a second-stage discrete optimization problem, using an exact dynamic program to enforce the frame budget and temporal coverage over the continuous OT marginal.

For an ordered subset $S = \{t_1 < \cdots < t_k\}$, we define the subset-extraction objective as maximizing

$$\mathcal{J}(S) = \sum_{\ell=1}^{k} w_{t_\ell} + \lambda \sum_{\ell=2}^{k} G(t_{\ell-1}, t_\ell), \qquad G(i, j) = \min(j - i, \delta), \tag{19}$$

where $\delta > 0$ denotes the desired minimum temporal gap between two consecutively selected frames. The coverage reward $G(i, j)$ increases with temporal distance and saturates once the gap reaches $\delta$, encouraging temporal spread without disproportionately favoring very distant frame pairs.

Because Eq. (19) has a chain-structured pairwise form, the globally optimal size-$k$ subset under this objective can be recovered exactly by dynamic programming. Let $\text{DP}[i, c]$ denote the maximum objective value achievable by selecting $c$ frames from $\{1, \ldots, i\}$ with frame $i$ fixed as the last selected frame. Then

$$\text{DP}[i, c] = w_i + \max_{j<i} \left(\text{DP}[j, c-1] + \lambda G(j, i)\right), \tag{20}$$

with initialization $\text{DP}[i, 1] = w_i$. The optimal subset is obtained by backtracking from $\max_i \text{DP}[i, k]$.

### 3.6 Algorithm Summary

The complete OTFS pipeline is summarized in Algorithm 1. Given a query and a long video, we first compute the evidence sources, construct the transport cost matrix, solve the semi-unbalanced OT problem via OTFS-Sinkhorn, and obtain frame-importance distribution from the target marginal. We then solve the coverage-aware dynamic program to recover the final subset of $k$ frames.

**Algorithm 1** OTFS for Long-Video Frame Selection

**Require:** Query $q$, frames $V = (x_1, \ldots, x_M)$, frame budget $k$, marginals $\mathbf{a}, \mathbf{b}$, transport hyperparameters $\varepsilon, \gamma$

Compute semantic scores $r_j^{\mathrm{sem}}$, utility scores $r_j^{\mathrm{util}}$, and transition scores $r_j^{\mathrm{trans}}$
Construct $R = (r_{ij})$ and $C = (1 - r_{ij})$
Form $K = \exp(-C/\varepsilon)$ and initialize $\mathbf{v}^{(0)} = \mathbf{1}$
**for** $t = 1, \ldots, T$ **do**
  $\mathbf{u}^{(t)} \leftarrow \mathbf{a} \oslash (K\mathbf{v}^{(t-1)})$
  $\tilde{\mathbf{v}} \leftarrow \mathbf{b} \oslash (K^\top \mathbf{u}^{(t)})$
  $\mathbf{v}^{(t)} \leftarrow \tilde{\mathbf{v}}^{\,\gamma/(\gamma+\varepsilon)}$
**end for**
Recover $\pi^\star = \mathrm{Diag}(\mathbf{u}^{(T)}) K \mathrm{Diag}(\mathbf{v}^{(T)})$
Compute frame weights $\mathbf{w} = (\pi^\star)^\top \mathbf{1}$
Solve Eq. (19) by dynamic programming and backtracking to obtain $S^\star$
**return** $S^\star$

# 4 Experiments

## 4.1 Experimental Settings

**Datasets.** We evaluate OTFS on three representative long-video understanding benchmarks: Video-MME [7], LongVideoBench [40], and MLVU [51]. Their video durations range from a few minutes to about one hour, covering diverse long-video settings. Following the official evaluation protocols, we report accuracy on the standard evaluation splits. For each downstream MLLM, all methods are compared under the same frame budget within the LMMs-Eval [47] framework. We do not use subtitle information in any experiment. Additional dataset statistics are deferred to the appendix.

**Implementation details.** We instantiate the selector on top of two video-capable MLLM backbones, Qwen2.5-VL-7B [1] and Qwen3-VL-4B [2], and evaluate each method under two budgets, $k \in \{32, 64\}$. Unless otherwise specified, all compared methods use the same prompt format, decoding setup, and downstream evaluation protocol, so performance differences are attributable to the frame selector itself. For OTFS, we use three evidence atoms corresponding to semantic relevance, utility reward, and temporal transition, with source prior $\mathbf{a} \propto (0.5, 0.4, 0.1)$ and a uniform frame-side prior $\mathbf{b}$ with $b_j = 1/M$ for $j = 1, \ldots, M$. We set the transport hyperparameters to $\gamma = 0.1$ and $\varepsilon = 0.05$ in all experiments, and run OTFS-Sinkhorn for 100 iterations by default. Following AKS [34], we sample the input video at 1 fps to construct the candidate-frame stream. More details are provided in the appendix.

## 4.2 Main Results

We compare OTFS against uniform sampling and four representative training-free frame-selection methods: AKS [34], Q-Frame [49], FOCUS [53], and DIG [14]. All baseline results are reproduced using their publicly released codebases under the same evaluation framework. Tables 1 and 2 report the results on the two downstream MLLM backbones.

OTFS achieves the strongest overall performance across both backbones and both frame budgets. It obtains the best result in 11 out of 12 backbone-budget-dataset settings, and in the remaining case is within 0.03 points of the strongest baseline. On Qwen2.5-VL-7B, OTFS improves over the best baseline by $1.11$, $1.64$, and $0.35$ points on Video-MME, LongVideoBench, and MLVU at $k = 32$, and by $0.07$ and $0.43$ points on LongVideoBench and MLVU at $k = 64$. On Qwen3-VL-4B, the gains are larger: OTFS surpasses the strongest baseline by $1.18$, $1.94$, and $3.83$ points at $k = 32$, and by $0.48$, $0.45$, and $2.70$ points at $k = 64$.

These results support the main claim of the paper. Compared with prior training-free baselines, OTFS benefits from formulating frame selection as a principled evidence-allocation problem with an explicit optimization objective, and from enforcing the final frame budget through exact subset extraction. The gains persist across datasets, backbones, and frame budgets, suggesting that the advantage of

Table 1: Main results with Qwen2.5-VL-7B.

| | 32 Frames | | | 64 Frames | | |
|---|---|---|---|---|---|---|
| Method | Video-MME | LongVideoBench | MLVU | Video-MME | LongVideoBench | MLVU |
| Uniform Sampling | 60.89 | 58.56 | 58.69 | 62.56 | 59.84 | 60.24 |
| Q-Frame [49] | 60.77 | 59.01 | 60.89 | 62.92 | 60.88 | 63.64 |
| FOCUS [53] | 61.29 | 60.35 | 64.29 | 62.00 | 61.70 | 64.22 |
| DIG [14] | 62.11 | 61.85 | 65.49 | **64.25** | 63.28 | 66.43 |
| AKS [34] | 62.44 | 62.83 | 65.39 | 63.81 | 63.87 | 64.96 |
| OTFS (ours) | **63.55** | **64.47** | **65.84** | 64.22 | **63.94** | **66.86** |

Table 2: Main results with Qwen3-VL-4B.

| | 32 Frames | | | 64 Frames | | |
|---|---|---|---|---|---|---|
| Method | Video-MME | LongVideoBench | MLVU | Video-MME | LongVideoBench | MLVU |
| Uniform Sampling | 61.77 | 60.73 | 61.20 | 64.18 | 62.75 | 64.47 |
| Q-Frame [49] | 62.48 | 61.55 | 63.33 | 65.22 | 64.32 | 66.84 |
| FOCUS [53] | 61.11 | 62.75 | 66.54 | 62.51 | 63.79 | 67.96 |
| DIG [14] | 63.22 | 64.02 | 66.97 | 65.48 | 66.19 | 67.56 |
| AKS [34] | 63.96 | 65.37 | 68.27 | 65.55 | 66.71 | 69.27 |
| OTFS (ours) | **65.14** | **67.31** | **72.10** | **66.03** | **67.16** | **71.97** |

OTFS is structural rather than tied to a particular benchmark or model configuration. A qualitative visualization is provided in Figure 2.

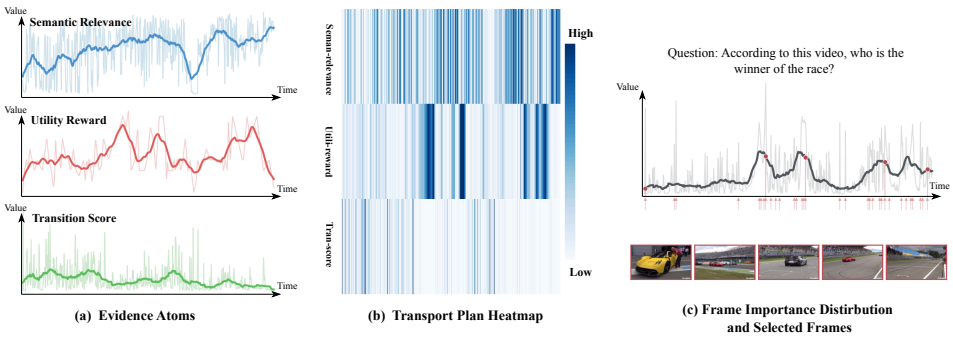
### 4.3 Ablation Study

The ablations in Table 3 are conducted on Qwen3-VL-4B with $k = 32$. We vary one component at a time while keeping the others fixed in order to isolate the contributions of three design choices in OTFS: multi-source evidence modeling, semi-unbalanced transport geometry, and exact budgeted subset extraction. Additional hyperparameter sensitivity results are deferred to the appendix.

**Effect of multi-source evidence modeling.** We first examine how the choice of evidence sources affects performance. Using semantic relevance alone already gives a competitive selector, confirming that direct query-frame alignment is a strong signal. Adding the utility reward further improves performance on all three benchmarks, indicating that semantic alignment and downstream answer usefulness are complementary rather than interchangeable. Incorporating temporal transition yields the strongest overall variant, especially on Video-MME and MLVU, which suggests that temporal variation provides additional evidence not captured by query-conditioned scoring alone. Overall, the full three-source instantiation performs best because it combines semantic, utility, and temporal cues within the same global evidence-allocation objective.

**Effect of transport geometry.** We isolate the effect of the transport formulation by keeping the evidence atoms and DP extractor fixed. The tuned-fusion baseline uses the same three normalized evidence atoms and the same DP extractor as OTFS, but replaces transport with a grid-searched scalar weighted sum. We also compare with balanced OT, which enforces hard marginal constraints on both evidence and frame sides. Semi-unbalanced OT achieves the strongest overall performance, indicating that the gains are not merely due to multi-signal fusion or the DP extractor. Unlike scalar fusion, which collapses heterogeneous cues into a single frame score before selection, OTFS preserves source-level evidence structure through the transport plan and then induces a target-side frame-importance marginal. This better matches long-video frame selection, where important evidence should be preserved while redundant frames need not absorb mass.

**Effect of the final subset extractor.** Finally, we study the role of the discrete subset extractor. With the transport-induced frame weights fixed, we compare the proposed dynamic programming solver against deterministic top-$k$ selection and Gumbel-Max sampling [9]. Dynamic programming performs best on all three benchmarks. This result shows that final performance depends not only on




(a) Evidence Atoms (b) Transport Plan Heatmap (c) Frame Importance Distirbution and Selected Frames

Figure 2: Left: Three evidence atoms over time. Middle: The optimal transport plan $\pi^*$, shown as a heatmap from evidence atoms to video frames. Different evidence sources assign mass to distinct temporal regions, highlighting their complementary roles in grounding the query. Right: The resulting frame-importance marginal $\mathbf{w}$ and the final subset selected by dynamic programming (red arrows indicate all selected frames, and red dots mark five representative examples).

Table 3: Ablation study of OTFS on Qwen3-VL-4B with $k = 32$.

| Variant | Video-MME | LongVideoBench | MLVU |
|---|---|---|---|
| *Evidence modeling* | | | |
| Semantic only | 63.88 | 65.14 | 69.06 |
| Semantic + Utility | 65.03 | **67.46** | 71.79 |
| Semantic + Utility + Transition (full) | **65.14** | 67.31 | **72.10** |
| *Transport geometry* | | | |
| Tuned score fusion | 64.18 | 65.81 | 70.23 |
| Balanced OT | 64.62 | 65.59 | 70.42 |
| Semi-unbalanced OT (full) | **65.14** | **67.31** | **72.10** |
| *Subset extraction* | | | |
| Top-$k$ | 63.74 | 66.49 | 71.60 |
| Gumbel-Max sampling | 64.62 | 65.89 | 71.09 |
| Dynamic programming (full) | **65.14** | **67.31** | **72.10** |

the transport-derived frame-importance distribution, but also on how the frame budget is enforced. Exact optimization of the coverage-aware extraction objective yields better subsets than heuristic ranking or stochastic sampling.

### 4.4 Complexity Analysis

Let $N$ denote the number of evidence atoms, $M$ the number of candidate video frames, $T$ the number of OTFS-Sinkhorn iterations, and $k$ the final frame budget. The transport stage performs iterative matrix scaling on an $N \times M$ transport matrix and therefore costs $\mathcal{O}(TNM)$. The subsequent dynamic program has complexity $\mathcal{O}(M^2k)$ under the recurrence in Eq. (20). Hence the overall complexity of OTFS is

$$\mathcal{O}(TNM + M^2k). \tag{21}$$

In our default instantiation, $N = 3$, so the transport stage scales linearly with the number of candidate frames. The DP stage has quadratic dependence on $M$ in its naive form, but it operates only on frame-level weights and a simple chain-structured coverage term, making its overhead modest relative to downstream MLLM inference. Moreover, the DP implementation can be accelerated to $\mathcal{O}(Mk)$ for this saturated-gap reward by maintaining prefix maxima.

Overall, once the evidence scores are available, the transport-and-extraction module adds only a small additional cost beyond evidence extraction. Additional efficiency analysis is deferred to the appendix.

Table 4: Robustness of OTFS to source noise on Qwen3-VL-4B with $k = 32$.

| Source | Noise ratio $\rho$ | Video-MME | LongVideoBench | MLVU |
|---|---|---|---|---|
| Semantic | 0.0 | 65.14 | 67.31 | 72.10 |
| | 0.1 | 65.14 | 67.09 | 71.47 |
| | 0.3 | 64.37 | 66.19 | 70.62 |
| | 0.5 | 64.04 | 67.09 | 70.16 |
| Utility | 0.0 | 65.14 | 67.31 | 72.10 |
| | 0.1 | 65.63 | 67.46 | 72.15 |
| | 0.3 | 64.37 | 67.84 | 72.50 |
| | 0.5 | 64.52 | 66.12 | 71.59 |
| Transition | 0.0 | 65.14 | 67.31 | 72.10 |
| | 0.1 | 65.11 | 67.31 | 71.42 |
| | 0.3 | 64.96 | 67.17 | 71.41 |
| | 0.5 | 65.07 | 67.09 | 71.06 |

### 4.5 Robustness to Corrupted Evidence

The performance of OTFS is ultimately bounded by the quality of the underlying evidence extractors. However, this does not necessarily mean that a local error in one evidence source will be directly or disproportionately amplified into the final frame subset.

To directly quantify this behavior, we independently corrupt each evidence source by mixing its atom-score vector with uniform random noise:

$$\widetilde{\mathbf{r}}^{(s)} = (1-\rho)\mathbf{r}^{(s)} + \rho\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{U}(0,1), \tag{22}$$

where $\rho \in \{0, 0.1, 0.3, 0.5\}$ controls the corruption strength, and $\rho = 0$ corresponds to the original noise-free setting. We apply the corruption separately to the semantic, utility, and transition sources while keeping the other two sources unchanged.

The overall results in Table 4 show limited sensitivity under low-to-moderate corruption and clearer degradation under severe corruption. OTFS is not equivalent to directly ranking frames by a single noisy score. In the transport stage, each evidence source contributes a fixed total mass determined by its source prior, preventing an individual source from contributing unrestricted weight to the final frame marginal. Entropic and marginal regularization further smooth the allocation over candidate frames. The subsequent DP stage then selects frames under the cardinality and temporal-coverage constraints, rather than directly taking the frames with the largest atom scores. Therefore, errors in individual atoms can affect the final subset, but they are not propagated through an unconstrained one-to-one mapping.

## 5 Conclusion

We introduced **OTFS**, a training-free and principled framework that formulates long-video frame selection as a semi-unbalanced optimal transport problem from heterogeneous evidence atoms to candidate video frames. By grounding frame importance in an explicit evidence-allocation objective, OTFS provides a structured alternative to prior hand-designed sampling procedures under the redundancy structure of long videos. Our formulation also clarifies how temperature-scaled one-dimensional scoring emerges as a single-source special case, placing several prior methods within a broader perspective. To obtain the final subset, OTFS couples transport-based evidence allocation with an exact dynamic program for coverage-aware budgeted extraction. Empirically, this combination yields consistent improvements across multiple long-video benchmarks, downstream MLLM backbones, and frame-budget regimes. We hope OTFS can serve as a useful step toward more principled visual compression and evidence selection for long-context multimodal reasoning.

## Acknowledgments and Disclosure of Funding

**Funding.** This work was supported by internal computational resources provided by Yizhun Medical AI Co., Ltd. No external funding was received for this research.

**Competing interests.** We declare no additional competing financial interests beyond the disclosed institutional affiliations.

## References


[1] Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[4] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314): 2563–2609, 2018.

[5] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274 (11):3090–3123, 2018.

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[7] Chaoyou Fu et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[8] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871, 2024.

[9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkE3y85ee`.

[10] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

[11] Minseok Kang, Minhyeok Lee, Jungho Lee, Minjung Kim, Donghyeong Kim, Dayeon Lee, Heeseung Choi, Ig-jae Kim, and Sangyoun Lee. Ott-vid: Optimal transport temporal token compression for video large language models. *arXiv preprint arXiv:2605.11803*, 2026.

[12] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 12:193057–193075, 2024. doi: 10.1109/ACCESS.2024.3517625.

[13] Bo Li et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=zKv8qULV6n`.

[14] Jialuo Li et al. Divide, then ground: Adapting frame selection to query types for long-form video understanding. *arXiv preprint arXiv:2512.04000*, 2025.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[16] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=MUjdNcfNPv`.

[17] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024.

[18] Bin Lin et al. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[19] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. Mm-vid: Advancing video understanding with gpt-4v(ision). *arXiv preprint arXiv:2310.19773*, 2023.

[20] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Xi Peng, et al. Multi-granularity correspondence learning from long-term noisy videos. In *International conference on learning representations*, pages 44188–44207, 2024.

[21] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *International Journal of Computer Vision*, 134(3), 2026. doi: 10.1007/s11263-025-02620-2.

[22] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3318–3327, 2025.

[23] Muhammad Maaz et al. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[24] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024.

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

[26] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Publishers, 2019.

[27] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 119336–119360, 2024.

[28] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[29] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54582–54599, 2025.

[30] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=Y2RW9EVwhT`.

[31] Min Shi, Shihao Wang, Chieh-Yun Chen, Jitesh Jain, Kai Wang, Junjun Xiong, Guilin Liu, Zhiding Yu, and Humphrey Shi. Slow-fast architecture for video multi-modal large language models. *arXiv preprint arXiv:2504.01328*, 2025.

[32] Enxin Song et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[33] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24090–24101, 2025.

[34] Xi Tang et al. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[35] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356, 2024.

[36] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024.

[37] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, volume 15138 of *Lecture Notes in Computer Science*, pages 58–76. Springer, 2024.

[38] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3282, 2025.

[39] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *Computer Vision – ECCV 2024*, volume 15091 of *Lecture Notes in Computer Science*, pages 453–470. Springer, 2024.

[40] Haoning Wu et al. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 28828–28857, 2024.

[41] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

[42] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591, 2025.

[43] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025.

[44] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[45] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, 2024.

[46] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023.

[47] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

[48] Peiyuan Zhang et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[49] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[50] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=ZJo6Radbqq`.

[51] Junjie Zhou et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701, 2025.

[52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[53] Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026.

[54] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18891–18901, 2025.

# Appendix

The appendix is organized as follows:

- Section A provides additional theoretical results and proofs for the semi-unbalanced OT formulation and the dynamic-programming-based subset selector.
- Section B presents additional dataset details used in our experiments.
- Section C reports hyperparameter sensitivity analyses.
- Section D provides additional efficiency analysis.
- Section E discusses the two-stage OT–DP decomposition.
- Section F discusses the limitations of the current framework.
- Section G discusses broader impacts.

## A Additional Theoretical Results and Proofs

In this appendix, we provide the derivations and proofs underlying the OTFS formulation. We first specialize the standard scaling characterization of entropic semi-unbalanced optimal transport [6, 5, 4, 26] to our evidence-allocation objective, which yields the OTFS-Sinkhorn updates used in the main text. We then analyze the task-specific consequence of this formulation: the single-source Gibbs reduction, which clarifies the connection between OTFS and one-dimensional frame-scoring rules. We also discuss the optimality of the dynamic-programming subset selector for the proposed coverage-aware extraction objective. Firstly, recall the optimization problem:

$$\min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle - \varepsilon H(\pi) + \gamma \, \mathrm{KL}\big(\pi^\top \mathbf{1} \,\|\, \mathbf{b}\big) \quad \text{s.t.} \quad \pi \mathbf{1} = \mathbf{a}, \tag{23}$$

where

$$H(\pi) = -\sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij} (\log \pi_{ij} - 1).$$

### A.1 Existence and Uniqueness

Assume $\varepsilon > 0$, $\gamma > 0$, $C_{ij} < \infty$ for all $i, j$, and $\mathbf{a} \in \Delta^N$, $\mathbf{b} \in \Delta^M$ have strictly positive entries. Under these conditions, the optimization problem in Eq. (23) admits a unique minimizer.

To see this, define the feasible set

$$\mathcal{U} = \left\{ \pi \in \mathbb{R}_+^{N \times M} : \pi \mathbf{1} = \mathbf{a} \right\}.$$

Since $\mathbf{a} \in \Delta^N$, the set $\mathcal{U}$ is nonempty, closed, and convex. Moreover, the row-sum constraint and nonnegativity imply that $\mathcal{U}$ is bounded, hence compact.

The objective in Eq. (23) consists of three parts. The transport cost $\langle C, \pi \rangle$ is linear in $\pi$. The negative entropy term $-\varepsilon H(\pi)$ is strictly convex on $\mathbb{R}_+^{N \times M}$ when $\varepsilon > 0$. The KL term

$$\mathrm{KL}(\pi^\top \mathbf{1} \,\|\, \mathbf{b})$$

is convex in the frame marginal $\pi^\top \mathbf{1}$ and therefore convex in $\pi$. Since $\mathbf{b}$ has strictly positive entries, this term is well defined over the feasible set.

Therefore, the overall objective is strictly convex on the convex feasible set $\mathcal{U}$, which implies uniqueness of the minimizer. Existence follows from lower semicontinuity of the objective together with compactness of $\mathcal{U}$.

### A.2 Scaling Form of the Optimal Plan

We next specialize the standard diagonal-scaling form for entropically regularized semi-unbalanced OT [6, 5, 4, 26] to our objective.

Let

$$K_{ij} = \exp(-C_{ij}/\varepsilon).$$

Then the unique minimizer $\pi^\star$ of Eq. (23) admits the factorized form

$$\pi^\star = \mathrm{Diag}(\mathbf{u})\, K\, \mathrm{Diag}(\mathbf{v})$$

for some positive vectors $\mathbf{u} \in \mathbb{R}^N_{++}$ and $\mathbf{v} \in \mathbb{R}^M_{++}$.

This factorization follows directly from the first-order optimality conditions. Introduce Lagrange multipliers $\alpha_1, \ldots, \alpha_N$ for the row constraints $\pi\mathbf{1} = \mathbf{a}$. The Lagrangian is

$$\mathcal{L}(\pi, \alpha) = \sum_{i,j} \pi_{ij} C_{ij} + \varepsilon \sum_{i,j} \pi_{ij} \log \pi_{ij} - \varepsilon \sum_{i,j} \pi_{ij} + \gamma\, \mathrm{KL}(\pi^\top \mathbf{1} \,\|\, \mathbf{b}) + \sum_i \alpha_i \left( a_i - \sum_j \pi_{ij} \right).$$

Let

$$w_j = \sum_i \pi_{ij}.$$

Differentiating $\mathcal{L}$ with respect to $\pi_{ij}$ gives

$$C_{ij} + \varepsilon \log \pi_{ij} + \gamma \log \frac{w_j}{b_j} - \alpha_i = 0.$$

Rearranging,

$$\pi_{ij} = \exp(\alpha_i/\varepsilon) \exp(-C_{ij}/\varepsilon) \exp\left( -\frac{\gamma}{\varepsilon} \log \frac{w_j}{b_j} \right).$$

Define

$$u_i = \exp(\alpha_i/\varepsilon), \qquad v_j = \exp\left( -\frac{\gamma}{\varepsilon} \log \frac{w_j}{b_j} \right).$$

Then

$$\pi_{ij} = u_i K_{ij} v_j,$$

which yields the stated diagonal-scaling form.

### A.3 Fixed-Point Equations and OTFS-Sinkhorn

Under the factorization above, the corresponding scaling vectors satisfy the fixed-point equations

$$\pi^\star = \mathrm{Diag}(\mathbf{u})\, K\, \mathrm{Diag}(\mathbf{v}), \qquad K = \exp(-C/\varepsilon).$$

$$\mathbf{u} = \frac{\mathbf{a}}{K\mathbf{v}}, \tag{24}$$

$$\mathbf{v} = \left( \frac{\mathbf{b}}{K^\top \mathbf{u}} \right)^{\frac{\gamma}{\gamma+\varepsilon}}, \tag{25}$$

where all operations are elementwise. These are the standard scaling updates for this class of semi-unbalanced entropic OT problems and directly motivate the OTFS-Sinkhorn solver used in the main text.

The first equation comes from the hard source-side constraint:

$$\pi^\star \mathbf{1} = \mathrm{Diag}(\mathbf{u}) K \mathbf{v} = \mathbf{a},$$

which immediately gives Eq. (24). For the second equation, let

$$\mathbf{w} = (\pi^\star)^\top \mathbf{1} = \mathrm{Diag}(\mathbf{v}) K^\top \mathbf{u},$$

so that $w_j = v_j (K^\top \mathbf{u})_j$. From the stationarity condition derived above,

$$v_j = \left( \frac{b_j}{w_j} \right)^{\gamma/\varepsilon}.$$

Substituting $w_j = v_j(K^\top \mathbf{u})_j$ yields

$$v_j = \left( \frac{b_j}{v_j(K^\top \mathbf{u})_j} \right)^{\gamma/\varepsilon}.$$

Equivalently,

$$v_j^{1+\gamma/\varepsilon} = \left( \frac{b_j}{(K^\top \mathbf{u})_j} \right)^{\gamma/\varepsilon}.$$

Taking both sides to the power $\varepsilon/(\gamma+\varepsilon)$ gives

$$v_j = \left( \frac{b_j}{(K^\top \mathbf{u})_j} \right)^{\frac{\gamma}{\gamma+\varepsilon}},$$

which is Eq. (25).

### A.4 Single-Source Gibbs Reduction

We now record an elementary single-source specialization of Eq. (23). This reduction is included for interpretive purposes: it makes explicit how temperature-scaled one-dimensional frame scoring arises from our objective in a degenerate regime.

Consider the single-source case $N = 1$, with source marginal $a_1 = 1$, cost row

$$c = (c_1, \ldots, c_M),$$

and frame prior $\mathbf{b} \in \Delta^M$. Then the transport plan is simply a row vector

$$\pi = (\pi_1, \ldots, \pi_M), \qquad \sum_{j=1}^{M} \pi_j = 1,$$

so Eq. (23) reduces to

$$\min_{\pi \in \Delta^M} \sum_{j=1}^{M} \pi_j c_j + \varepsilon \sum_{j=1}^{M} \pi_j \log \pi_j + \gamma \sum_{j=1}^{M} \pi_j \log \frac{\pi_j}{b_j}. \tag{26}$$

Expanding the KL term gives

$$\gamma \sum_{j=1}^{M} \pi_j \log \frac{\pi_j}{b_j} = \gamma \sum_{j=1}^{M} \pi_j \log \pi_j - \gamma \sum_{j=1}^{M} \pi_j \log b_j.$$

Hence the optimizer remains exponential in the negative cost, modulated by the prior $\mathbf{b}$; in this sense, the solution is Gibbs-type.

If the frame prior is uniform, namely

$$b_j = \frac{1}{M}, \qquad j = 1, \ldots, M,$$

then

$$\sum_{j=1}^{M} \pi_j \log \frac{\pi_j}{b_j} = \sum_{j=1}^{M} \pi_j \log \pi_j + \log M,$$

because $\sum_{j=1}^{M} \pi_j = 1$. The additive constant $\gamma \log M$ does not affect the optimizer, so Eq. (26) becomes equivalent to

$$\min_{\pi \in \Delta^M} \sum_{j=1}^{M} \pi_j c_j + (\varepsilon + \gamma) \sum_{j=1}^{M} \pi_j \log \pi_j.$$

Introducing a Lagrange multiplier for the simplex constraint gives

$$c_j + (\varepsilon + \gamma)(\log \pi_j + 1) + \tau = 0,$$

which implies

$$\pi_j = \frac{\exp(-c_j/(\varepsilon+\gamma))}{\sum_{\ell=1}^{M}\exp(-c_\ell/(\varepsilon+\gamma))}. \tag{27}$$

If the KL term is removed, the same calculation yields the standard softmax

$$\pi_j = \frac{\exp(-c_j/\varepsilon)}{\sum_{\ell=1}^{M}\exp(-c_\ell/\varepsilon)}.$$

This form immediately explains the connection to temperature-scaled frame scoring. If the single-source cost is defined by

$$c_j = 1 - r_j,$$

where $r_j$ is a scalar frame score, then Eq. (27) becomes

$$\pi_j \propto \exp(r_j/(\varepsilon+\gamma)),$$

because the shared factor $\exp(-1/(\varepsilon+\gamma))$ cancels during normalization. In the KL-free case, this further reduces to

$$\pi_j \propto \exp(r_j/\varepsilon).$$

Therefore, temperature-scaled frame scoring can be interpreted as a single-source specialization of Eq. (23).

### A.5 Optimality of the Dynamic Programming Selector

We now justify the recursion used in the second-stage coverage-aware subset selection.

Recall the objective

$$\mathcal{J}(S) = \sum_{\ell=1}^{k} w_{t_\ell} + \lambda \sum_{\ell=2}^{k} G(t_{\ell-1}, t_\ell), \tag{28}$$

for an ordered subset

$$S = \{t_1 < t_2 < \cdots < t_k\}.$$

Define

$$\mathrm{DP}[i, c]$$

as the optimal value of Eq. (28) when selecting $c$ frames from $\{1, \ldots, i\}$ with frame $i$ selected as the last frame. Then the recursion

$$\mathrm{DP}[i, c] = w_i + \max_{j<i}\Big(\mathrm{DP}[j, c-1] + \lambda G(j, i)\Big)$$

holds, with initialization

$$\mathrm{DP}[i, 1] = w_i.$$

The recursion follows from optimal substructure. Any optimal subset contributing to $\mathrm{DP}[i, c]$ must end with some predecessor $j < i$ immediately before frame $i$. Its objective therefore decomposes into the best value for selecting $c-1$ frames ending at $j$, namely $\mathrm{DP}[j, c-1]$, plus the incremental contribution of appending frame $i$, namely $w_i + \lambda G(j, i)$. Therefore the best subset ending at $i$ is obtained by choosing the predecessor $j < i$ that maximizes

$$\mathrm{DP}[j, c-1] + \lambda G(j, i),$$

which yields the stated recurrence. For $c = 1$, the only feasible subset ending at $i$ is $\{i\}$, whose value is $w_i$.

## B Additional Dataset Details

**Video-MME.** Video-MME [7] is a comprehensive benchmark for evaluating MLLMs on video analysis. It contains 900 manually curated videos with 2,700 human-annotated question-answer pairs, covering diverse domains and temporal regimes from short clips to long videos. In particular, its video durations range from 11 seconds to 1 hour, spanning short-, medium-, and long-term video understanding. In this work, we only use the visual stream and do not provide subtitle information to the model, so that the comparison focuses on frame selection under a visual-only setting.

**LongVideoBench.** LongVideoBench [40] is a long-context interleaved video-language benchmark designed to evaluate detailed reasoning over long inputs. It contains 3,763 web-collected videos and 6,678 human-annotated multiple-choice questions in 17 fine-grained categories. The videos can be as long as one hour, making the benchmark particularly suitable for testing whether a frame selector can preserve sparse but answer-critical evidence. We do not use subtitle information in our experiments.

**MLVU.** MLVU [51] is a multi-task long-video understanding benchmark covering diverse video genres such as movies, surveillance footage, egocentric videos, cartoons, and game videos. It is designed to evaluate several aspects of long-video understanding through nine task categories, including both multiple-choice and open-ended settings. The dataset is split into a development set containing 2,593 questions and a test set with 509 questions. For evaluation purposes, we exclusively employ multiple-choice questions from the development set, while excluding all open-ended questions.

# C Hyperparameter Sensitivity

We analyze the sensitivity of OTFS to its transport and subset-selection hyperparameters on Qwen3-VL-4B with $k = 32$. Results are reported in Tables 5 and 6.

**Transport hyperparameters.** Table 5 studies the source prior $\mathbf{a}$, entropic regularization $\varepsilon$, and marginal relaxation strength $\gamma$. Overall, OTFS is fairly stable across a broad range of settings on both Video-MME and LongVideoBench. The default source prior $(0.5, 0.4, 0.1)$ gives the best or near-best performance, suggesting that a moderate emphasis on semantic and utility signals with a smaller transition prior provides a good balance. For $\varepsilon$, intermediate values perform best, while overly small or overly large regularization leads to slight degradation, consistent with the trade-off between overly sharp and overly diffuse transport plans. A similar trend is observed for $\gamma$: moderate relaxation works best, whereas overly large $\gamma$ weakens the desired evidence-preserving behavior of the semi-unbalanced formulation.

**Dynamic-programming hyperparameters.** We set the trade-off coefficient $\lambda$ to $0.5$ on Video-MME and to $0.005$ on LongVideoBench and MLVU. The reason is that Video-MME includes more global questions and therefore benefits more from stronger temporal coverage, whereas LongVideoBench and MLVU contain more locally grounded questions, for which an overly large $\lambda$ may over-disperse the selected frames and weaken local evidence preservation. The temporal gap parameter $\delta$ is fixed to $4$ for all datasets.

Table 6 studies the trade-off coefficient $\lambda$ and temporal gap $\delta$ in the DP stage. Setting $\lambda = 0$, which removes the coverage term, consistently underperforms the default setting, confirming the importance of temporal coverage in the final subset selection. At the same time, excessively large $\lambda$ also hurts performance, indicating that over-emphasizing coverage may sacrifice evidence quality. The method is also reasonably robust to $\delta$, with $\delta = 4$ performing best overall in our setup and nearby values remaining competitive.

Overall, these results suggest that OTFS does not rely on fragile hyperparameter tuning, and the default settings used in the main experiments lie in stable regions of the parameter space.

Table 5: Hyperparameter sensitivity on Qwen3-VL-4B with $k = 32$. Default settings are *.

| | Source prior $\mathbf{a}$ | | | | Entropic regularization $\varepsilon$ | | | | Marginal relaxation $\gamma$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $(0.4, 0.5, 0.1)$ | $(0.5, 0.4, 0.1)^*$ | $(0.6, 0.3, 0.1)$ | $(0.3, 0.6, 0.1)$ | 0.02 | 0.05* | 0.1 | 0.2 | 0.02 | 0.05 | 0.1* | 0.2 |
| Video-MME | 64.92 | 65.14 | 64.81 | 64.77 | 64.96 | 65.14 | 64.85 | 65.25 | 65.00 | 65.00 | 65.14 | 64.48 |
| LongVideoBench | 67.31 | 67.31 | 66.86 | 67.01 | 66.79 | 67.31 | 66.94 | 65.96 | 67.31 | 66.79 | 67.31 | 66.41 |

Table 6: Sensitivity to dynamic-programming hyperparameters on Qwen3-VL-4B with $k = 32$.

| | Trade-off coefficient $\lambda$ | | | | | | Temporal gap $\delta$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 0 | 0.005 | 0.01 | 0.02 | 0.05 | 0.5 | 2 | 4* | 6 | 8 |
| Video-MME | 63.74 | 65.25 | 65.18 | 64.88 | 65.00 | 65.14 | 64.11 | 65.14 | 64.22 | 65.11 |
| LongVideoBench | 66.49 | 67.31 | 66.41 | 66.11 | 65.44 | 64.77 | 67.09 | 67.31 | 65.22 | 66.04 |

## D Efficiency Analysis

We analyze efficiency by separating the optimization core of OTFS from the full evidence-extraction pipeline. As discussed in Sec 4.4, the OTFS solver itself is lightweight: the Sinkhorn-based transport step and the subsequent DP-based subset extraction introduce only a small additional overhead once the evidence scores are available. This is also confirmed empirically in Table 7, where the standalone Sinkhorn+DP stage requires only 5.8, 9.2, and 16.1 minutes on LongVideoBench, MLVU, and Video-MME, respectively. The wall-clock time of full OTFS is mainly determined by evidence extraction, while the transport-and-extraction module contributes only a small fraction of the total cost.

To study a more practical operating point, we further reduce the candidate-frame rate from 1 fps to 0.5 fps. As shown in Table 7, this substantially lowers the preprocessing cost, bringing the overall runtime much closer to AKS, while Table 8 shows that the corresponding performance drop remains modest across all three benchmarks. This result suggests that the main computational bottleneck of the current system is not the OT-based optimization core, but the cost of obtaining rich evidence signals at dense temporal resolution.

Overall, these results support two conclusions. First, OTFS does not introduce a heavy optimization bottleneck beyond evidence extraction. Second, the practical efficiency of the full pipeline can be improved further through cheaper evidence estimators or lower candidate-frame rates, without changing the OTFS formulation itself. All experiments are conducted on 8 NVIDIA 4090 GPUs.

Table 7: Efficiency comparison on three datasets. We report standard LMM inference time (minutes) under different frame budgets and additional key frame selection time (minutes) for different methods.

| | LMM Inference Time | | Key Frame Selection Time | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | 32 Frames | 64 Frames | Q-Frame | AKS | DIG | OTFS (1 fps) | OTFS (0.5 fps) | Sinkhorn+DP |
| LongVideoBench | 27.3 | 53.6 | 658.3 | 660.5 | 300.9 | 971.2 | 634.1 | 5.8 |
| MLVU | 43.5 | 81.5 | 902.3 | 906.0 | 567.1 | 1487.6 | 1033.4 | 9.2 |
| Video-MME | 48.2 | 94.7 | 1856.1 | 1861.3 | 789.7 | 2676.1 | 1743.2 | 16.1 |

Table 8: Results with different fps.

| | 32 Frames | | |
|---|---|---|---|
| Method | Video-MME | LongVideoBench | MLVU |
| OTFS (1 fps) | 65.14 | 67.31 | 72.10 |
| OTFS (0.5 fps) | 65.07 | 67.01 | 71.55 |

## E Discussion on the two-stage OT–DP decomposition.

**Why the OT stage does not explicitly include $k$.** The decomposition is deliberate and separates two roles. The OT stage estimates a global, budget-independent evidence allocation over all candidate frames, whereas the DP stage imposes the specific cardinality constraint and temporal structure required by the final frame budget. As a result, the same transport allocation can be reused for different frame budgets, such as $k = 32$ and $k = 64$, without recomputing the evidence scores or the transport plan.

A direct joint formulation would require binary support variables $y_j \in \{0, 1\}$, for example,

$$\sum_{j=1}^{M} y_j = k, \qquad \pi_{ij} \le a_i y_j,$$

together with a temporal coverage reward defined over the selected support. Such a formulation would become a mixed discrete–continuous optimization problem. It would no longer retain the convex transport subproblem or admit the efficient Sinkhorn scaling updates used by OTFS. The proposed two-stage pipeline is therefore a tractable decomposition that separates global evidence grounding from exact budgeted subset extraction.

**Why the DP stage uses the target marginal.** Let

$$w_j = \sum_{i=1}^{N} \pi^{\star}_{ij}$$

be the target marginal of the optimal transport plan. For the additive evidence retention term used in our current subset objective, we have

$$\sum_{j \in S} w_j = \sum_{j \in S} \sum_{i=1}^{N} \pi^{\star}_{ij} = \sum_{i=1}^{N} \sum_{j \in S} \pi^{\star}_{ij}.$$

Therefore, under the current objective, the marginal $\mathbf{w}$ is sufficient for measuring the total transport mass retained by the selected subset. The DP does not need the full transport matrix to optimize this particular additive quantity.

# F Limitations

OTFS provides a principled optimization framework for frame selection, but its empirical performance still depends on the quality and task coverage of the underlying evidence atoms. In our current instantiation, these atoms are derived from semantic relevance, utility reward, and temporal transition. If these estimators are noisy, biased, or fail to capture the information truly required by a question, the transport stage may propagate such errors into the final subset. This issue can be more pronounced for questions requiring abstract temporal reasoning or subtle evidence not well captured by the current evidence design.

In addition, OTFS optimizes a frame-selection objective rather than the downstream reasoning objective of the MLLM itself. Thus, the selected subset should be interpreted as optimal with respect to the proposed transport-and-extraction objective, rather than guaranteed to be globally optimal for final question-answering accuracy. Relatedly, while the OTFS optimization core is lightweight, the runtime of the full system is currently dominated by evidence computation. Improving evidence quality and efficiency remains an important direction for future work.

# G Broader Impacts

This work studies long-video frame selection for multimodal large language models. A more efficient and principled selector can reduce the computational cost of long-context video understanding, which may improve the accessibility of video-language systems and lower the cost of deploying them in research and industry. In this sense, OTFS may contribute to more efficient multimodal reasoning over long videos.

At the same time, improved video understanding may also enable harmful applications, including large-scale surveillance, privacy intrusion, or more effective analysis of sensitive personal media. In addition, frame selection errors may systematically omit important evidence, which can be particularly problematic in high-stakes settings such as education, healthcare, or safety monitoring. Since OTFS is a training-free method built on upstream evidence signals, biases or failures in those signals may propagate into the selected subset and subsequently affect downstream reasoning.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The last few sentences in **Abstract**; the last paragraph of **Introduction**.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix F

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Appendix A

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the datasets, backbones, frame budgets, evidence sources, hyperparameters, evaluation framework, baselines, ablations, and additional sensitivity/efficiency analyses. This provides sufficient information to understand and reproduce the main experimental protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper uses public datasets and provides substantial implementation details, but anonymized code and full reproduction scripts are not included at submission time. We will release code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the evaluation datasets, backbones, budgets, evidence sources, transport hyperparameters, frame sampling rate, and comparison protocol in the main paper, with additional dataset details and hyperparameter sensitivity analyses provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The experimental setup in this study involved extensive testing across multiple datasets and with various types of backbone networks, consistently demonstrating improvements through our method. This extensive validation across diverse conditions ensures the reproducibility and reliability of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper includes complexity analysis in the main paper, efficiency measurements and compute resources in appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work does not violate the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G discusses the societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or other asset that would require special safeguards for controlled release. The work studies an inference-time frame-selection method built on existing public benchmarks and pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets and codes used in the paper have been cited in their original literature.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The code is not publicly available at this time. We will make them available later.

    Guidelines:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human-subject research or new data collection from participants, so IRB approval is not applicable

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM/VLM components are part of the experimental pipeline and evaluation setting, including the use of video-capable MLLM backbones for downstream evaluation and a reward-based utility signal obtained following the public DIG implementation. These usages are described in Sections 3.4 and 4.1.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.