ReFoCUS: Reinforcement-guided Frame Optimization for Contextual Understanding

Supplementary Material

Appendix Contents

A. Details for *ReFoCUS*-393K

In this section, we elaborate on the data curation pipeline used to construct the *ReFoCUS*-393K dataset. The main manuscript describes our reward-variance filtering framework at a conceptual level, and we probe a pretrained video-LLM with multiple frame subsets and retain only those video-QA pairs that exhibit strong temporal sensitivity. Here, we make this procedure fully concrete by specifying (i) how long videos are decomposed into temporal segments, (ii) how each segment is paired with its complementary region, and (iii) how these intervals are turned into candidate frame subsets for reward computation. These implementation details precisely characterize the preprocessing that produces the training instances for policy learning.

A.1. Temporal Window-Complement Construction

Our goal in constructing frame subsets is to systematically contrast local evidence against the rest of the video, so that temporal sensitivity can be measured in a controlled way. To this end, we first tile each video with overlapping temporal windows that capture localized segments, and then pair each window with its complementary region, which serves as a counterfactual context where that local segment is removed.

Formally, for each video v with total length T frames, we divide it into overlapping temporal segments using a fixed window and stride. Specifically, we use a window size of $w = \lceil T/8 \rceil$ and a stride of $s = \lceil w/2 \rceil$, resulting in 8 temporal windows:

$$W_1{=}[0, w),\ W_2{=}[s, s{+}w),\ \ldots,\ W_8{=}[0, s){\cup}[7s, T). \quad (4)$$

The first seven windows behave like a sliding temporal crop that sweeps across the video with 50% overlap, while W_8 covers the remaining tail region and wraps back to the beginning if necessary. This construction ensures that long videos are covered in a near-uniform manner, while still providing sufficient overlap to capture events that straddle window boundaries.

As described in the main manuscript, we do not rely solely on localized segments. Instead, we explicitly construct a complementary region for each window to disentangle the contribution of a short temporal segment from that of the surrounding context. For each window W_i, we define its complementary range as $C_i = [0, T) \setminus W_i$, which contains all frames of the video except those in W_i. From each of W_i and C_i, we then uniformly sample $k = 32$ frames to

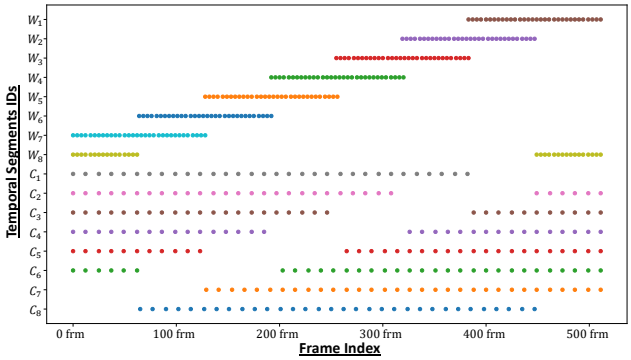


Figure 8: Visualization of the temporal segmentation and sampling strategy. We divide each video into 8 overlapping windows W_1 to W_8 (top), and for each window, define a complementary region C_i (bottom). We uniformly sample frames from both regions to construct 16 candidate subsets per QA pair.

construct a total of 16 candidate frame subsets:

$$\begin{cases} c_1, \ldots, c_8 & \text{from windows,} \\ c_9, \ldots, c_{16} & \text{from complements.} \end{cases} \tag{5}$$

In other words, each window W_i yields a local subset that focuses on a specific temporal segment and a complementary subset that retains the rest of the video while masking out that segment. By comparing the predictions of the pre-trained LMM across these 16 subsets, we can quantify how much the model's answer depends on specific temporal regions rather than on global, shortcut-style cues. Figure 8 provides a schematic illustration of this sampling scheme, where each W_i (temporal window) and C_i (complement region) is visualized as a distinct dashed line over the frame index axis.

A.2. Margin Computation over Temporal Regions

Given the 16 candidate frame subsets $\{c_j\}_{j=1}^{16}$ constructed for each video–question pair $\langle v, t \rangle$, we quantify temporal sensitivity using the same margin-based reward employed during policy training. For each subset c_j, the frozen reference LMM produces answer logits, from which we compute a prediction margin r_j that reflects how confidently the model favors the correct answer over the strongest distractor. Concretely, r_j is defined as the logit gap between the ground-truth option and the highest-scoring incorrect choice, consistent with the margin computation used in the main training pipeline. The reward variance for a given pair $\langle v, t \rangle$ is then defined as:

$$\mathrm{Var}(m) = \mathrm{Var}\big(\{r_j\}_{j=1}^{16}\big), \tag{6}$$

where the variance measures how strongly the model's answer confidence changes across different temporal regions of the same video. When the question can be resolved from global context or static cues, the margins r_j remain nearly identical across all subsets, resulting in low variance. Conversely, when the question depends on localized temporal events, informative and uninformative subsets yield markedly different margins, producing higher variance. As a result, $\mathrm{Var}(m)$ provides a direct and interpretable proxy for temporal grounding strength.

A.3. Variance-based Sample Filtering

To construct a training pool that emphasizes temporally grounded supervision, we filter each video–QA pair according to its reward variance. For every pair, we compute $\mathrm{Var}(m)$ over the 16 frame subsets and retain only those whose variance exceeds a threshold of $\tau = 0.2$. This threshold is selected empirically from the overall variance distribution, balancing the removal of flat, low-signal instances with the preservation of a sufficiently large set of temporally informative samples.

Applying this criterion to the full collection of 962K QA pairs removes examples that exhibit minimal changes in reward across subsets—typically globally answerable questions or instances dominated by dataset-level biases. The resulting filtered dataset contains approximately 393K QA pairs, which we use as the temporally grounded training pool for subsequent policy optimization.

A.4. Processing Setup

All video frames are pre-extracted to maximize inference throughput. For each video–question pair, the 16 frame subsets are generated by indexing into this cached frame pool, and their prediction margins are computed in parallel through a single batched forward pass of the reference LMM. Both raw logits and variance values are recorded for debugging and ablation analyses. This setup ensures stable reward estimation and supports reliable learning dynamics during downstream policy optimization.

B. Extended Details for *ReFoCUS*

B.1. Training Hyperparameters

During training, for each $\langle v, t \rangle$ instance, the policy generates N frame subsets of length T', which are used for learning. The policy model performs only a single update following each exploration step. We train the policy model based on 1.3B Video-MA^{2}mba and use an InternVL3-2B reward model, with AdamW (β_1=0.9, β_2=0.99), a batch size of 64 $\langle v, t \rangle$ pairs per update, and gradient clipping of 1.0. The learning rate is scheduled by linear warmup of 256 steps followed by cosine decay. Training is conducted in *bfloat16* precision under DeepSpeed ZeRO-1, with input frames resized to 384×384. Hyperparameters that vary under the Search-Space Scaling Curriculum are summarized in Tab. 5.

Table 5: Stage varying hyperparameters.

Stage	T'	N	β	$\text{LR}_{\text{initialized}}$	$\text{LR}_{\text{pretrained}}$
Stage 1	4	256	1×10^{-3}	6×10^{-5}	1×10^{-5}
Stage 2	8	128	3×10^{-4}	3×10^{-5}	1×10^{-5}
Stage 3	16	64	2×10^{-4}	2×10^{-5}	1×10^{-5}
Stage 4	32	32	1×10^{-4}	2×10^{-5}	1×10^{-5}

B.2. Frame Selection Process

Here, we provide the full implementation details of the autoregressive frame-selection mechanism summarized in Sec. 3.2. The model encodes the input video v and query t to produce frame embeddings F, each taken from the final output embedding of its corresponding visual sequence. The backbone's hidden representations (z_i and F) are transformed into query Q, key K, and value V through linear projection followed by RMS-normalization. Starting from `<sof>`, a query embedding q_i derived from the previously selected frame interacts with all frame keys K to produce the selection score $\mathbf{A}$. These scores define a probability distribution over candidate frames, from which a new frame index f_i is sampled without replacement. The embedding of the selected frame, $V[f_i]$, is then fed back as the selected embedding for the next selection, and this process continues until T' frames are chosen. The detailed autoregressive selection procedure is illustrated in Algorithm 1.

Algorithm 1 *ReFoCUS* Selection

Require: video v, user query t, selection length T'

$F \leftarrow \text{Backbone}(v, t)$ ▷ get frame embedding
$K \leftarrow \text{norm}(F\,W_k^\top)$ ▷ for key
$V \leftarrow \text{norm}(F\,W_v^\top)$ ▷ for value
$\mathbf{e} \leftarrow$ `<sof>`
for $i = 1$ to T' **do**
  $\mathbf{z}_i \leftarrow \text{BackboneStep}(\mathbf{e})$
  $q_i \leftarrow \text{norm}(\mathbf{z}_i W_q^\top)$
  $\mathbf{A} = \frac{q_i\,K^\top}{\sqrt{d_{model}}}$ ▷ get selection score
  $\mathbf{A}[f_{<i}] \leftarrow -\infty$ ▷ prevent duplicate selection
  $f_i \sim \text{softmax}(\mathbf{A})$ ▷ equivalent to $\pi_\theta(\cdot \mid f_{<i}, v, t)$
  $\mathbf{e} \leftarrow V[f_i]$ ▷ get selected embedding
end for
return $\{f_i\}_{i=1}^{T'}$

B.3. Simplifying the Prediction Margin Reward

To derive the logit-based, numerically stable simplification of the margin reward from Eq. (1) introduced in Sec. 3.2, we expand the reward definition as follows. Each frame subset s_j generated by the policy π_θ is evaluated by the reward model r_φ, which outputs a confidence distribution over answer candidates. We identify the most competitive incorrect choice $\tilde{y}$ as:

$$\tilde{y} = \arg\max_{y \neq y^*} r_\varphi(y \mid s_j, v, t), \tag{7}$$

and compute the prediction margin between the correct answer y^* and this hardest negative $\tilde{y}$ as:

$$r_j = \frac{r_\varphi(y^* \mid s_j, v, t) - r_\varphi(\tilde{y} \mid s_j, v, t)}{r_\varphi(y^* \mid s_j, v, t) + r_\varphi(\tilde{y} \mid s_j, v, t)}. \tag{8}$$

Letting $z_j(y)$ denote the pre-softmax logits, this margin can be written more simply as:

$$r_j = \frac{e^{z_j(y^*)} - e^{z_j(\tilde{y})}}{e^{z_j(y^*)} + e^{z_j(\tilde{y})}} = \tanh\left(\frac{z_j(y^*) - z_j(\tilde{y})}{2}\right), \tag{9}$$

which avoids explicit probability normalization and yields a more numerically stable reward computation.

C. Additional Experiments

C.1. Evaluation on Extra OE Benchmarks

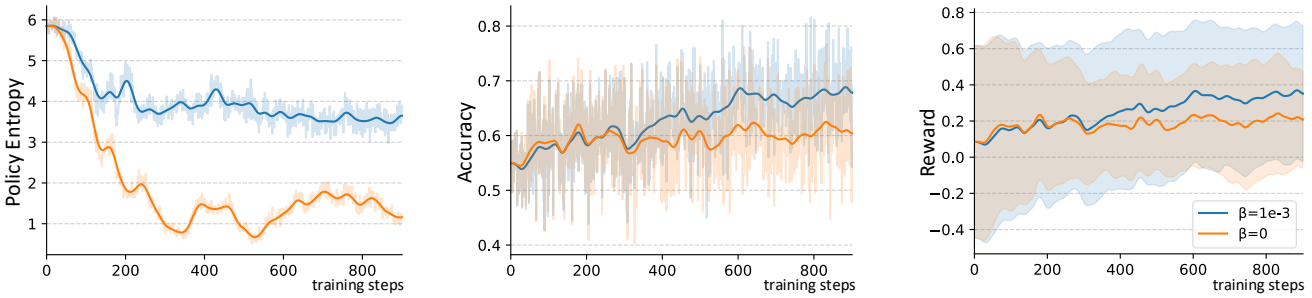
To examine whether *ReFoCUS* remains effective for open-ended settings, we report additional results on *ActivityNet-QA* [6] and *Video-ChatGPT* [32] in Tab. 6. As in the table, across 7–8B open-sourced models, *ReFoCUS* consistently improves scores on all open-ended evaluation setups. Similar to the improvements observed on the open-ended subset of *NExT-QA* in Tab. 1, these results confirm that *ReFoCUS* can generalize beyond multiple-choice training and remains effective in open-ended question answering with high transferability.

Table 6: Comparison between Baseline (uniform sampling) and *ReFoCUS* on *ActivityNet-QA* and *Video-ChatGPT*. Both Acc (for multiple-choice) and Score (for open-ended responses) are presented on a 0-100 scale.

Model	LLM	ActivityNet-QA		VCGPT
		acc	score	score
VideoLLaMA3	7B	52.4	70.2	60.3
+ **ReFoCUS**		51.8 ↓-0.6	72.2 ↑2.0	62.7 ↑2.4
LLaVA-OV	7B	50.9	68.8	59.9
+ **ReFoCUS**		52.3 ↑1.4	69.7 ↑0.9	61.4 ↑1.5
InternVL3	8B	53.7	69.4	59.4
+ **ReFoCUS**		55.4 ↑1.7	70.4 ↑1.0	61.0 ↑1.6
InternVL3.5	8B	50.4	67.7	59.5
+ **ReFoCUS**		52.2 ↑1.8	68.8 ↑1.1	60.8 ↑1.3
Qwen3-VL	8B	50.8	65.9	60.3
+ **ReFoCUS**		52.4 ↑1.6	67.2 ↑1.3	62.3 ↑2.0

C.2. Low Entropy Degradation Analysis

To validate the effect of the entropy bonus $\mathcal{H}(\pi_\theta)$ introduced in Sec. 3.2, we retrained the policy under the Stage 1



(a) Policy Entropy $\mathcal{H}_\theta$ (b) Train Accuracy (c) Train reward with $\pm\sigma$

Figure 9: Ablation of entropy regularization. Eliminating the entropy coefficient β=0 leads to early entropy collapse and limits long-term performance improvement. Even a small entropy bonus coefficient β is sufficient to prevent this degradation.

configuration in Tab. 5 with β=0. Here, $\mathcal{H}(\pi_\theta)$ denotes the expected policy entropy across autoregressive frame-selection steps,

$$\mathcal{H}(\pi_\theta) = \mathbb{E}_{f_{<i}\sim\pi_\theta}\left[\mathcal{H}\big(\pi_\theta(\cdot \mid f_{<i}, v, t)\big)\right], \qquad (10)$$

which regularizes the frame-selection distribution to remain sufficiently stochastic and thus preserves the model's capacity for exploration. When the entropy regularization is removed (β=0), the policy entropy $\mathcal{H}(\pi_\theta)$ drops sharply in the early phase, leading to a rapid decline in explorability, the model's ability to meaningfully search diverse frame combinations, and causing premature convergence (Fig. 9 (a)). This degradation manifests as both mean reward and accuracy plateauing at a lower level (Fig. 9 (b)), accompanied by an early collapse of the standard deviation of reward (Fig. 9 (c)). Such a reduction in reward variability reflects a narrowing of the performance-improvement horizon, beyond which the policy can no longer discover higher-reward configurations. Additionally, the model begins to exhibit false confidence, consistently selecting frames with unwarranted certainty despite limited evidence, reinforcing suboptimal behaviors. In contrast, β=10^{-3} maintains adequate entropy and standard deviation throughout training, preventing false confidence and sustaining a broad performance-improvement horizon. This confirms that controlled entropy is essential for maintaining exploration potential while guiding stable convergence toward superior performance.

C.3. Ablation on Reward Formulation

To validate the necessity of our margin-based reward formulation (Eq. (1)), we compare it against a simpler binary reward variant that assigns a reward of 1 when the reward model predicts the correct answer and 0 otherwise. Both variants are trained under identical conditions. As shown in Tab. 7, the binary reward already provides a meaningful improvement over uniform sampling (62.2 → 64.6), confirming that even coarse feedback is sufficient to guide frame selection toward informative subsets. However, our margin-based design achieves a further gain (64.6 → 65.4), with the improvement especially pronounced in the medium and long video categories. This indicates that the continuous, magnitude-aware signal captures finer distinctions between frame subsets that the binary indicator cannot express. By encoding how confidently the reward model r_φ favors the correct answer over the strongest distractor, the margin reward provides richer gradient information to the policy, enabling it to discriminate between merely adequate and truly optimal frame compositions.

Table 7: Ablation on reward formulation. We compare our margin-based reward (Eq. (1)) against a binary variant (1 if the reward model predicts the correct answer, 0 otherwise) under the same training setup. We report the mean score of five 7–8B models on Video-MME with 32 frames.

Reward Design	**short**	**medium**	**long**	**overall**
Uniform sampling	73.8	61.0	51.9	62.2
ReFoCUS (binary)	**75.3**	63.3	55.2	64.6
ReFoCUS (margin, **ours**)	**75.3**	**64.1**	**56.8**	**65.4**

C.4. Ablation on Query Conditioning

To quantitatively verify that our policy genuinely conditions its frame selection on the input query rather than learning a fixed, query-agnostic selection pattern, we conduct a *cross-query interference* experiment on Video-MME. Specifically, we replace the correct query with a randomly sampled question from the same video during inference, while keeping all other conditions unchanged. As shown in Tab. 8, providing a random question leads to a significant performance drop (65.4 → 58.1), which falls even below the uniform sampling baseline (62.2). This result demonstrates that the policy actively filters out frames deemed irrelevant to the conditioning query: when guided by an incorrect question, it suppresses the visual evidence required for the actual answer, thereby degrading performance below that of uninformed uniform selection. The finding confirms that query conditioning plays a critical role in guiding effective frame selection and that *ReFoCUS* performs genuine

Table 8: Ablation on query conditioning. We evaluate the effect of providing an incorrect query during frame selection. "Random Question" replaces the correct query with a randomly sampled question from the same video. We report the mean score of five 7–8B models on Video-MME with 32 frames.

Query Condition	short	medium	long	overall
Uniform sampling	73.8	61.0	51.9	62.2
ReFoCUS (Random Question)	66.0	55.1	53.1	58.1
ReFoCUS (Correct Question)	**75.3**	**64.1**	**56.8**	**65.4**

query-conditioned semantic grounding rather than relying on superficial visual saliency or fixed temporal heuristics.

D. Qualitative Case Study

We present qualitative comparisons across uniform sampling, MDP^3 [46], and our *ReFoCUS* policy in Figs. 10 to 15. Each example visualizes the selected frames along the timeline, enabling direct inspection of where each method concentrates its evidence. The uniform sampling selects frames at fixed temporal intervals regardless of the question, and therefore it often fails to include the critical temporal segments required for answering the query. MDP^3 shows stronger focus than uniform sampling, but its decisions are largely driven by low-level RGB similarity or visually salient fragments rather than query-conditioned semantics, resulting in inconsistent reasoning.

In contrast, *ReFoCUS* consistently selects compact yet semantically aligned frame subsets that directly correspond to the question intent. This query-conditioned behavior is clearly visible in the timestamp visualizations: frames cluster around the temporal regions that contain the decisive visual cues required to answer the question. Notably, for some cases exemplified in Fig. 15, *"What do the expanding red lines on the map in the first few minutes of the video stand for?"*, both uniform sampling and MDP^3 fail to identify or densely sample the early-video interval mentioned in the query. *ReFoCUS*, however, accurately localizes this segment, gathers the relevant evidence concentrated in the first few minutes, and consequently provides the correct answer. These examples demonstrate that the learned policy performs genuine semantic grounding rather than heuristic or appearance-based selection.

E. Failure Case and Limitation

Although *ReFoCUS* demonstrates robust query-conditioned frame selection across diverse scenarios, it also has some limitations. As illustrated in Figs. 16 and 17, since the selection process operates under a fixed frame budget, the policy sometimes uses most of its budget too early in the video. When this happens, it may focus on initial frame segments and miss later events that are still relevant to the question. Such cases mostly appear in long videos that contain multiple separated temporal cues, where early commitment prevents the policy from observing the full temporal context.

While these errors are relatively infrequent, they highlight an inherent trade-off between early commitment and global temporal coverage. Future work may explore more adaptable selection strategies or hierarchical policies that dynamically reallocate the budget based on the observed evidence, potentially mitigating premature focus and improving global temporal awareness.

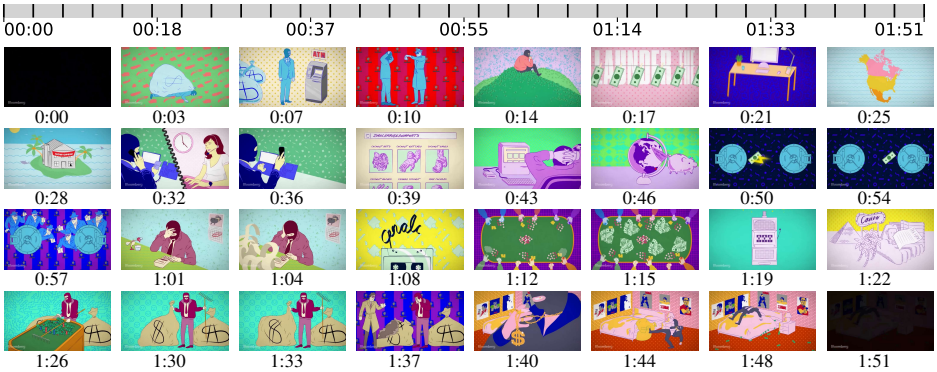
E.1. Discussion and Limitation

While *ReFoCUS* introduces a new perspective by shifting policy optimization from output-level textual behavior to input-level visual grounding, several limitations remain. Similar to other RL-based methods, our training requires substantial computational cost due to repeated autoregressive sampling and reward evaluation. Moreover, the learned policy can reflect biases or blind spots of the reward model, since its selection behavior is shaped by the model's visual scoring patterns. Despite these challenges, our results show that modeling selection behavior directly at the input level can yield semantically coherent and contextually meaningful frame subsets. We believe that this direction opens new opportunities for aligning LMM behavior not only by how they respond, but also by how they perceive and prioritize visual information.

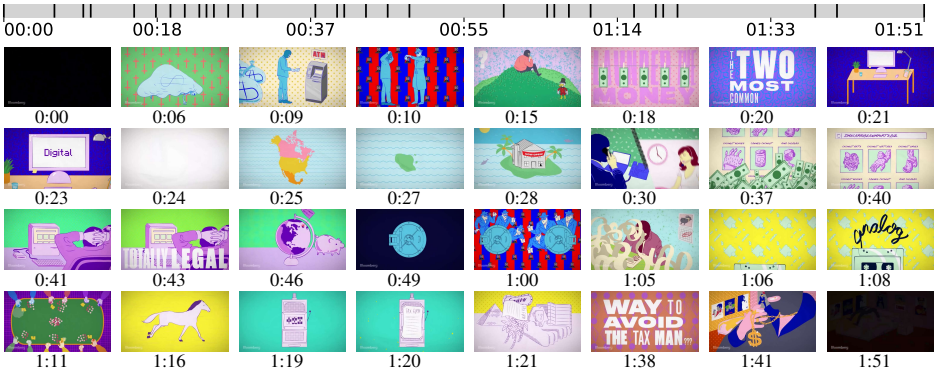
Question:
Which time is displayed on the clock in the video?

Options:
- A. 08:00.
- B. 03:00.
- C. 11:00.
- D. 05:00.

Uniform **Ans: C. 11:00.**



MDP^3 **Ans: B. 03:00.**



ReFoCUS **Ans: D. 05:00.**

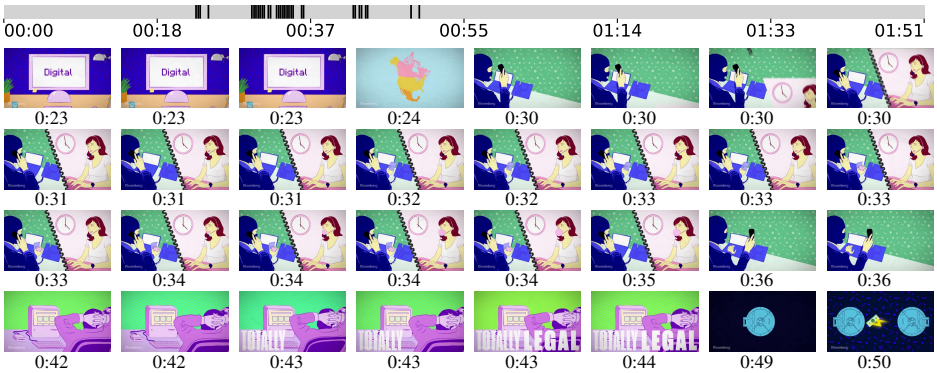


Figure 10: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

Question:
What is the total number of people in the video?

Options:
- A. 7.
- B. 6.
- C. 5.
- D. 8.

Uniform **Ans: B. 6.**

MDP^3 **Ans: B. 6.**

ReFoCUS **Ans: A. 7.**

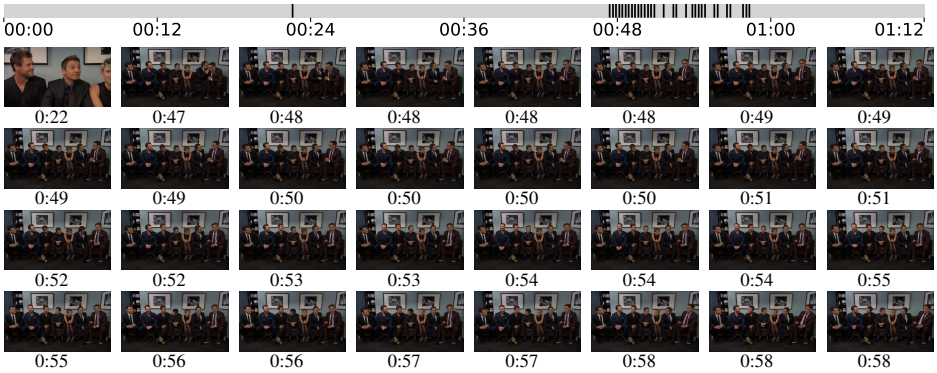
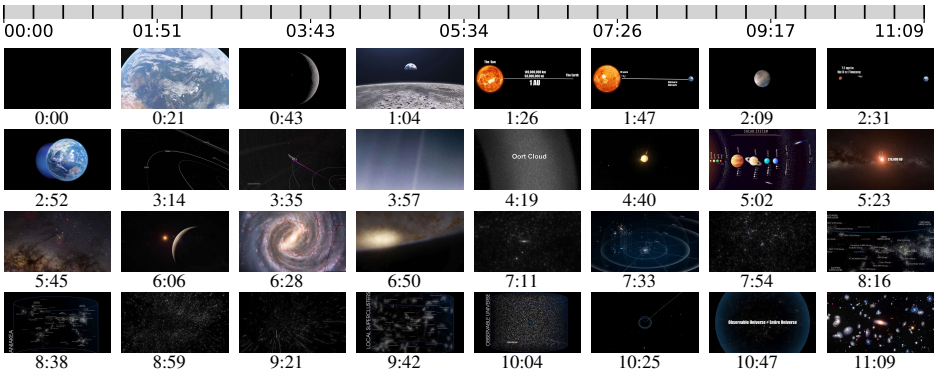


Figure 11: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

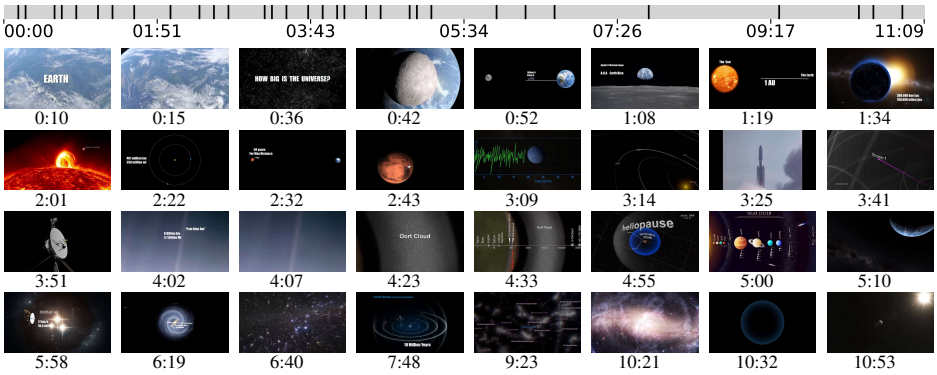
Question:
The video shows how long it takes to drive from the Earth to the Moon?

Options:
- A. 160 days.
- B. 50 days.
- C. 180 days.
- D. 19 days.

Uniform **Ans: B. 50 days.**



MDP^3 **Ans: B. 50 days.**



ReFoCUS **Ans: A. 160 days**

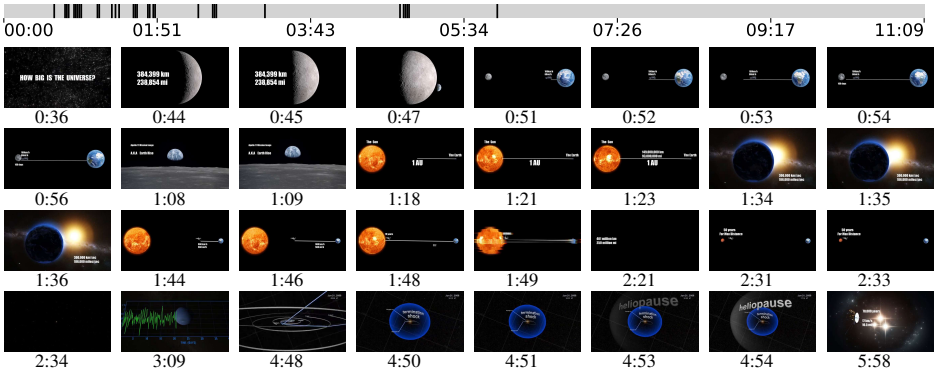


Figure 12: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

Question:
In which part of the video is a man wearing a red jersey interviewed?

Options:
- A. There is no interview with a man in red clothing.
- B. End of the video.
- C. Beginning of the video.
- D. Middle of the video.

Uniform **Ans: D. Middle of the video.**

MDP^3 **Ans: D. Middle of the video.**

ReFoCUS **Ans: C. Beggining of the video**

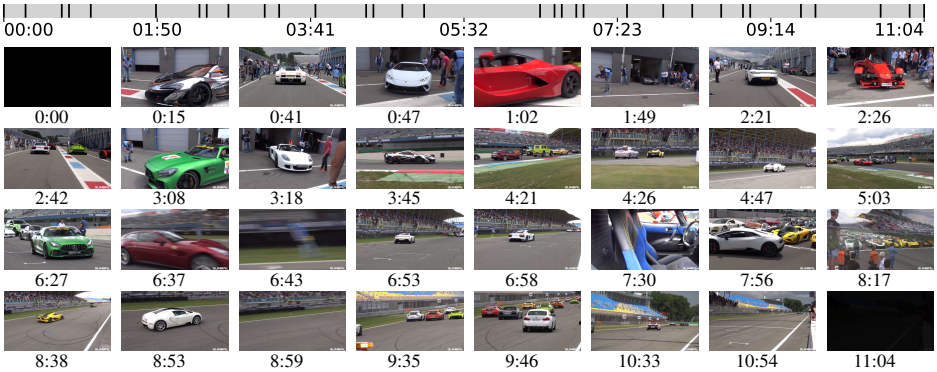
Figure 13: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

Question:
In which order are the first four cars driven out of the garage in this video? (a) The yellow car. (b) The black car. (c) The silver car. (d) The white car.

Options:
- A. (a)(b)(c)(d).
- B. (a)(c)(b)(d).
- C. (b)(c)(d)(a).
- D. (b)(d)(a)(c).

Uniform **Ans: C. (b)(c)(d)(a).**

MDP^3 **Ans: D. (b)(d)(a)(c).**



ReFoCUS **Ans: B. (a)(c)(b)(d).**

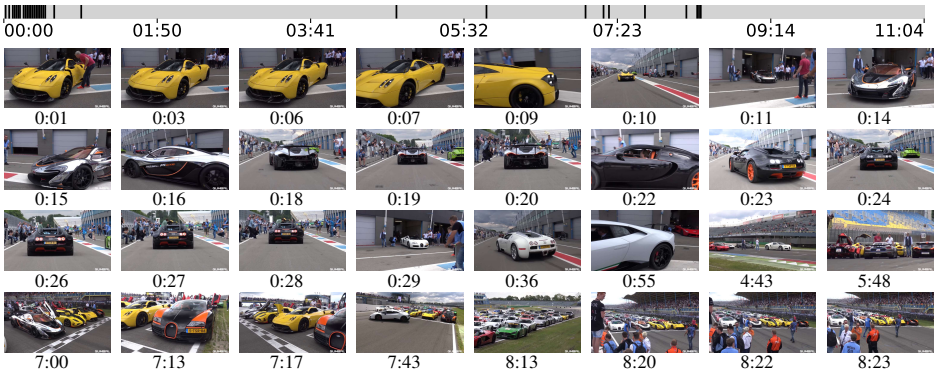


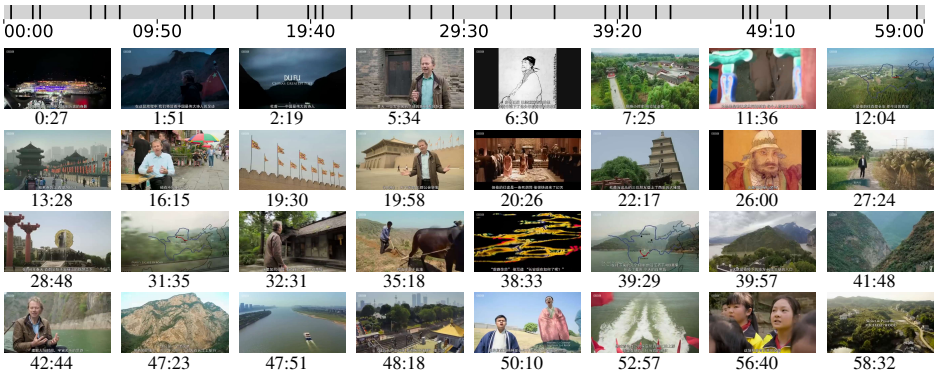
Figure 14: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

Question:
What do the expanding red lines on the map in the first few minutes of the video stand for?

Options:
- A. The Yellow River.
- B. The Silk Road.
- C. Du Fu's route to Xi'an.
- D. The Yangtze River.

Uniform **Ans: C. Du Fu's route to Xi'an.**

MDP^3 **Ans: C. Du Fu's route to Xi'an.**



ReFoCUS **Ans: B. The Silk Road.**

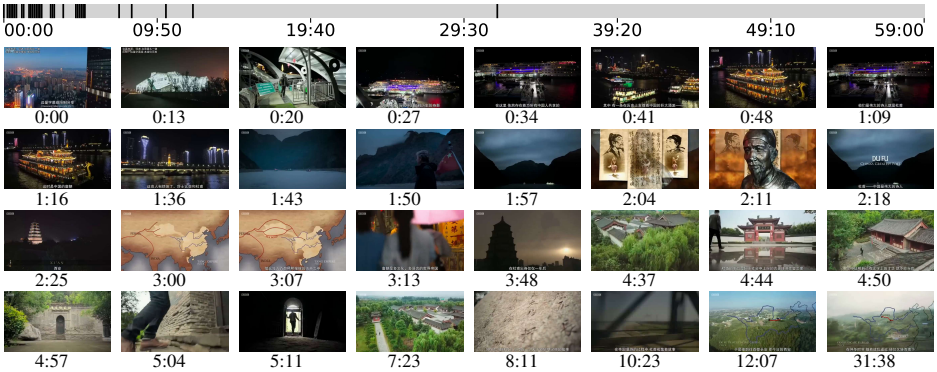
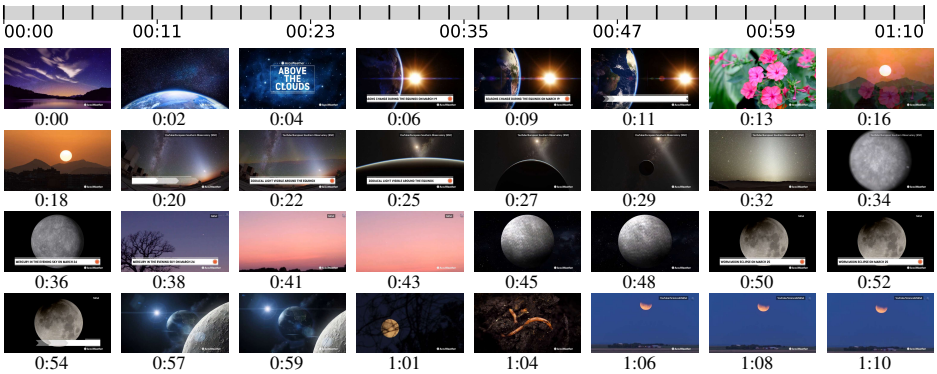


Figure 15: Qualitative comparison of temporal localization across uniform sampling, MDP^3, and *ReFoCUS*.

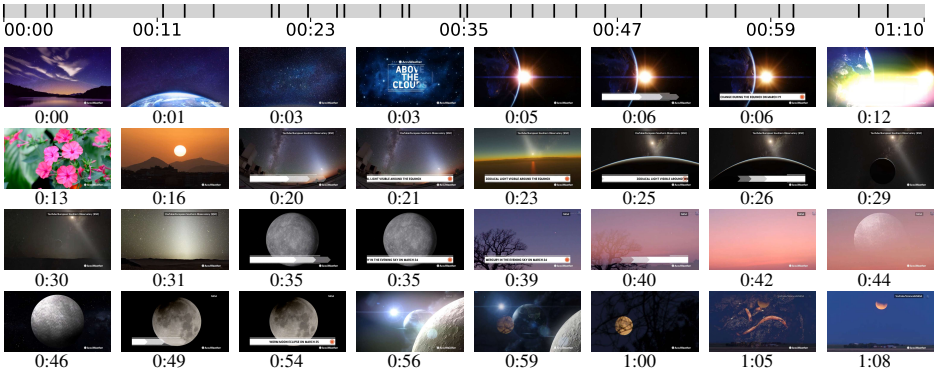
Question:
How many times is the sun visible in the video?

Options:
- A. 2.
- B. 4.
- C. 3.
- D. 1.

Uniform **Ans: C. 3.**



MDP^3 **Ans: A. 2.**



ReFoCUS **Ans: C. 3.**

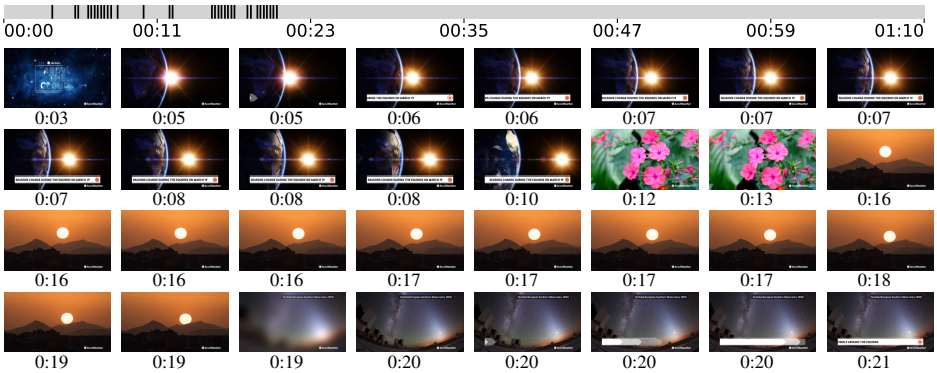


Figure 16: Comparison of representative failure cases across uniform sampling, MDP^3, and *ReFoCUS*.

Question:
How many magic shows are included in this video?

Options:
- A. 10.
- B. 9.
- C. 8.
- D. 7.

Uniform **Ans: C. 8.**

MDP^3 **Ans: B. 9.**

ReFoCUS **Ans: D. 7.**

Figure 17: Comparison of representative failure cases across uniform sampling, MDP^3, and *ReFoCUS*.