UniComp: Rethinking Video Compression Through Informational Uniqueness

Supplementary Material

1. Derivation of the Reconstruction Error Bound

In this section, we provide the algebraic derivation supporting the upper bound relation

$$\mathcal{E}(\mathcal{S}) = \sum_{j\in\mathcal{X}} \|x_j - \hat{x}_j\|^2 \leq 2 \sum_{j\in\mathcal{X}} \min_{i\in\mathcal{S}} u_{ij} \tag{1}$$

where $\hat{x}_j$ is the reconstruction of x_j using tokens in $\mathcal{S}$. where $\hat{x}_j = \sum_{i\in\mathcal{S}} w_{ij} x_i$, $\sum_i w_{ij} = 1$, $w_{ij} \geq 0$, and $u_{ij} = 1 - s_{ij}$, and s_{ij} denotes the cosine similarity between normalized features. The weight w_{ij} reflects the relative similarity between x_i and x_j:

$$w_{ij} = \frac{\exp(s_{ij})}{\sum_{i\in\mathcal{S}} \exp(s_{ij})} \tag{2}$$

Step 1. Expansion of the reconstruction error. Given normalized feature vectors $\|x_i\| = \|x_j\| = 1$, we can expand the reconstruction error term as

$$\|x_j - \hat{x}_j\|^2 = \|x_j - \sum_i w_{ij} x_i\|^2 \tag{3}$$

$$= \|x_j\|^2 - 2\sum_i w_{ij} x_j^\top x_i + \sum_{i,k} w_{ij} w_{kj} x_i^\top x_k \tag{4}$$

$$= 1 - 2\sum_i w_{ij} s_{ij} + \sum_{i,k} w_{ij} w_{kj} s_{ik} \tag{5}$$

Step 2. Bounding the cross-term. Since all pairwise similarities satisfy $s_{ik} \leq 1$, the last term in Eq. (5) can be upper-bounded by

$$\sum_{i,k} w_{ij} w_{kj} s_{ik} \leq \sum_{i,k} w_{ij} w_{kj} = \Big(\sum_i w_{ij}\Big)^2 = 1 \tag{6}$$

Substituting back gives a coarse upper bound:

$$\|x_j - \hat{x}_j\|^2 \leq 2(1 - \sum_i w_{ij} s_{ij}) \tag{7}$$

Step 3. Approximating with the most similar token. Since all the discarded tokens has a very similar token has been selected, so, one selected token x_{i^*} could dominate the reconstruction (i.e., $w_{i^*j} \approx 1$ and $s_{i^*j} = \max_i s_{ij}$), then

$$\sum_i w_{ij} s_{ij} \approx s_{i^*j} \tag{8}$$

and Eq. (7) becomes

$$\|x_j - \hat{x}_j\|^2 \leq 2(1 - s_{i^*j}) \tag{9}$$

Thus, we obtain the final form:

$$\sum_{j\in\mathcal{X}} \|x_j - \hat{x}_j\|^2 \leq 2 \sum_{j\in\mathcal{X}} \min_{i\in\mathcal{S}} (1 - s_{ij}) = 2 \sum_{j\in\mathcal{X}} \min_{i\in\mathcal{S}} u_{ij} \tag{10}$$

Step 4. Interpretation. This result implies that the reconstruction error of any discarded token x_j is upper-bounded by the angular distance to its most similar selected token in the normalized feature space. The bound holds under the convex-combination constraint ($w_{ij} \geq 0, \sum_i w_{ij} = 1$), which ensures that $\hat{x}_j$ lies inside the convex hull of the selected features. Thus, a higher pairwise similarity s_{ij} directly corresponds to a smaller reconstruction uncertainty, providing a principled link between similarity-based redundancy reduction and information preservation.

2. Auto Compression Analysis

More details of main experiment settings. In the main experiments, Frame Group Fusion (FGF) may directly compress video lower than a given retained ratio due to temporal correlations and redundancies, so that Spatial Dynamic Compression (SDC) may be bypassed when the resulting token count is below the target. Without a preset ratio, the model performs automatic compression based on both temporal and spatial redundancies.

Fully automatically compression without retention limitation. Since ***UniComp*** supports fully automatic video compression (results shown in Table 1), we visualize the retention tokens under the automatic setting across four benchmarks in Figure 1. The visualization also shows the information redundancy of videos in different benchmarks, which means a lot of videos can be represented with only a few tokens, and LongVideoBench exhibits the highest redundancy.

Moreover, in Figure 2, we show the average automatically compressed tokens on VideoMME, which has equal numbers of short, medium, and long videos (100 each). It shows the automatically compressed tokens, the hyper-parameter U_c, and their performance on VideoMME.

3. More Experiments

We further evaluate *UniComp* on sensitive sub-tasks and long-video scenarios. As shown in Table 2, *UniComp*

Table 1. *UniComp* fully automatically compress results on LLava-OneVision-7B. It may be slightly lower than the main experiments since this experiment will not limit the retention ratio and it could compress much lower (e.g. 2% as shown in Figure 1).

Method	Retention Ratio	LVB	Ego	MLVU	VMME
Baseline	100%	56.3	60.4	64.7	58.4
UniComp (auto)	26.3%	57.3	61.5	64.0	58.8

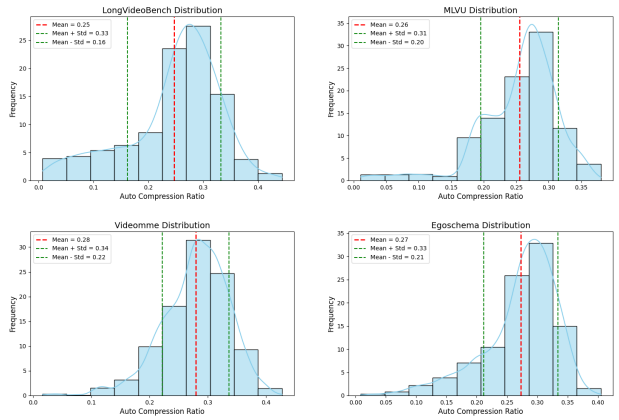


Figure 1. Auto compression on different benchmarks of LLava-OneVision-7B. It demonstrates the information redundancy of videos in different benchmarks.

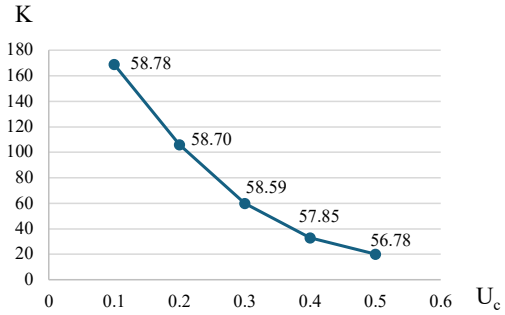


Figure 2. Each frame average retained token number K under auto compression settings with U_c on VideoMME with LLava-OneVision-7B, which shows the best retained tokens K of different U_c, given a detailed understanding of hyper-parameter U_c.

achieves the best performance across all four sub-tasks (TP, OR, AR, CP). Notably, it preserves temporal and object information as well as the full-frame baseline while offering clear gains on action recognition and counting, indicating its stronger ability in retaining critical visual cues during compression.

For long-video evaluation on Eagle-2.5 (Table 3), which is naturally designed for hour-long videos understanding, *UniComp* consistently outperforms VisionZip and HoliTom under all compression settings, from 128f to 512f inputs

Table 2. Comparison on sub-tasks that may be heavily affected by token compression on VideoMME. Temporal Perception (TP), Object Recognition (OR), Action Recognition (AR), and Counting Problem (CP).

Method	TP	OR	AR	CP
Base (32f)	63.6	65.3	54.3	37.3
VisionZip	61.8	64.4	58.1	35.4
FastVid	61.8	63.8	54.0	35.4
HoliTom	60.0	64.1	55.3	37.3
UniComp	**63.6**	**65.5**	**58.5**	**39.6**

Table 3. Performance comparison on VideoMME under long video settings on **Eagle-2.5 which is designed for hour long video** with **hundreds of frames input** to show the real video understanding ability. Each experiment compresses to 64 frames' tokens (64*256). HoliTom* means without the inner-LLM mode (w/o M), as this mode is very difficult to implement in other models.

Method	Frames/Retain	All	Short	Medium	Long
Vanilla	64f / 100%	68.9	80.6	67.1	59.1
VisionZip	128f / 50%	67.8	78.9	65.1	59.4
HoliTom*	128f / 50%	68.4	81.0	65.4	58.9
UniComp	128f / 50%	**70.1**	**81.4**	**69.0**	**59.8**
VisionZip	256f / 25%	66.3	76.0	64.0	58.9
HoliTom*	256f / 25%	67.9	78.7	65.4	59.4
UniComp	256f / 25%	**70.4**	**80.8**	**69.7**	**60.8**
VisionZip	512f / 12.5%	65.0	72.9	62.9	59.2
HoliTom*	512f / 12.5%	67.4	77.9	65.4	58.9
UniComp	512f / 12.5%	**70.7**	**80.3**	**69.3**	**62.4**

compressed to 64-frame tokens.

4. ViT keys versus other representations in SDC

Using ViT Keys/Values or Last layer feats as representations to calculate uniqueness. See Table 4.

Table 4. Comparison of ViT keys versus other representations

	Longvideo	Egoschema	MLVU	Videomme
Attn Values	57.4	58.8	63.6	57.4
Last layer feats	55.8	**61.1**	**64.7**	58.4
Attn Keys (ours)	**57.7**	**61.1**	64.4	**58.7**

5. Breakdown TTFT latency

Tab. 5 breaks down TTFT latency, showing that SDC contributes most to the overhead, while the "Other" proportion increases with the number of frames. Tab. 6 compares methods under the 320-frame setting. Our algorithm introduces less overhead than others.

Table 5. TTFT breakdown (ms) under the 320-frame setting with 10% retained ratio..

Frame	Compress	FGF	TA	SDC	Other
32	111	1.60	1.08	61.78	46.54
128	395	6.21	0.93	206.20	181.66
256	754	13.48	1.41	367.82	371.30
320	927	17.73	1.73	447.38	460.15

Table 6. Comparison of different methods under the 320-frame setting with 10% retained ratio.

Method	ViT	Compress	LLM	Avg. Score
VisionZip	3111	474	948	59.53
HoliTom	3111	2020	993	61.43
UniComp	3111	927	846	**62.45**
Full Tokens	3111	455	16716	57.55

6. Comparisons on temporal grouping strategies

Tab.7 shows comparison with temporal grouping strategies. Unlike standard attention methods that may overlook subtle changes, UniComp maximizes information retention. We preserve motion cues via weighted fusion (FGF), specifically anchored on the first frame to maintain trajectory boundaries (avoiding ambiguity in directional motion), and adaptive token allocation (TA) for unique movements.

Table 7. Comparison on different temporal grouping strategies

Method	LVB	Ego	MLVU	VMME
Dyseg (FastVid)	51.6	56.5	58.8	53.7
Redundancy-Aware (HoliTom)	57.0	60.1	61.6	55.6
FGF (Ours)	**57.7**	**61.1**	**64.4**	**58.7**

7. Implementation Details

Our method is implemented on the LLaVA-OneVision-7B, LLaVA-Video-7B and Eagle2.5-7B models. We conduct evaluation on NVIDIA H800 (80 GB) GPUs, while inference is tested on NVIDIA H20-3e (141 GB) GPUs to better reflect practical deployment scenarios. All benchmark evaluations are performed using LMMs-Eval.

8. Qualitative Examples

8.1. Real example visualizations and workflow

To further illustrate how *UniComp*'s real behaves under different temporal dynamics, we present two real examples that visualize both the retained tokens and the detailed fusion patterns generated by our pipeline.

Figure 3 shows a case with **frequent scene switches**. In such highly dynamic scenarios, consecutive frames exhibit strong semantic differences. *UniComp* adaptively identifies these transitions through Frame Group Fusion (FGF), generating finer-grained groups. Token Allocation (TA) assigns more tokens to these semantically distinctive segments to preserve critical visual cues. During Spatial Dynamic Compression (SDC), tokens are selected based on their local uniqueness, with the selection order indicated by white labels. Tokens appearing earlier in the list exhibit higher uniqueness. In the fusion maps, tokens sharing the same color are fused into the representative token marked by a red rectangle. Despite significant scene changes, *UniComp* consistently preserves the most informative regions, producing high-fidelity visual reconstructions that remain clear even when zoomed in.

Figure 4 presents a scenario where the scene remains **nearly unchanged**. Here, the global uniqueness across frames is low, leading FGF to merge many consecutive frames into a single stable semantic segment. Consequently, TA allocates fewer tokens to this region. SDC further identifies large areas of spatial redundancy, allowing many tokens to be merged into a small set of representative ones. Even under aggressive compression, *UniComp* maintains coherent visual content and produces high-quality images.

Together, these examples demonstrate the adaptive nature of *UniComp*: it allocates richer capacity to segments with high semantic variation while aggressively compressing redundant regions, achieving efficient yet information-preserving visual token compression.

8.2. Performance under different compression ratio

As shown in Figure 5. We visualize two examples with open-end video question-answer under different compression ratio. It could be seen that *UniComp* keep performing well.

As shown in Figure 6. We visualize an example with video caption generation task under 10% compression ratio. *UniComp* can still achieve strong understanding compared to other SOTA methods.

8.3. Performance under different benchmarks

As shown in Figure 7 and Figure 8. We visualize examples on four benchmarks compared to SOTA compression methods.

Figure 3. **A real example that scenes switch frequently** of *UniComp* on LLaVA-OneVision-7B with 32 frames input under 10% retention ratio. The images are **high-fidelity** and can be **zoomed in to view details**. After Spatial Dynamic Compression, it shows the retained tokens with the selection order labeled white. Each line of detailed fusion map shows 10 tokens, same color token will be fused to the token with red rectangle. The order from top to bottom is the selection order, which means higher one is more unique.

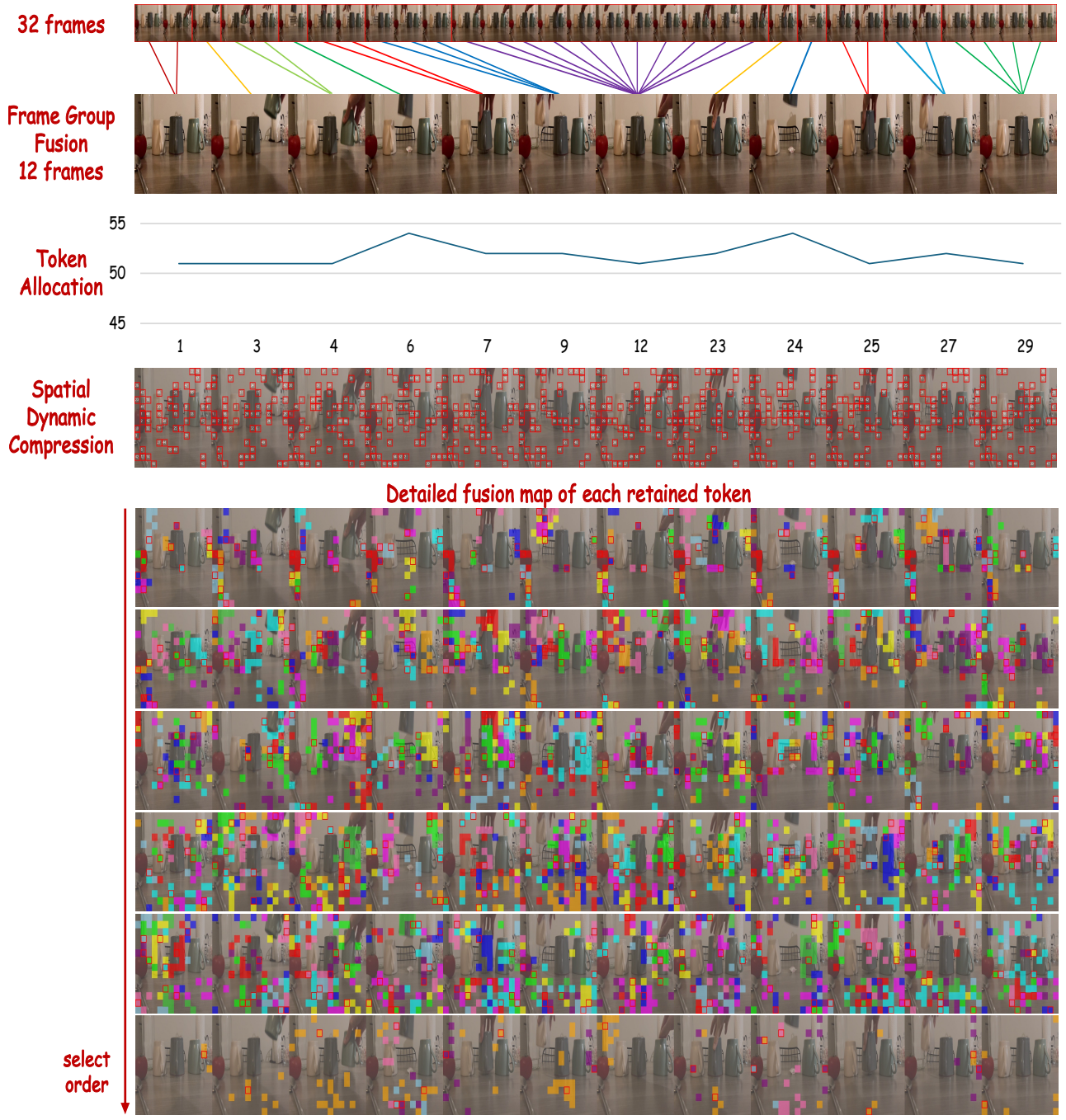
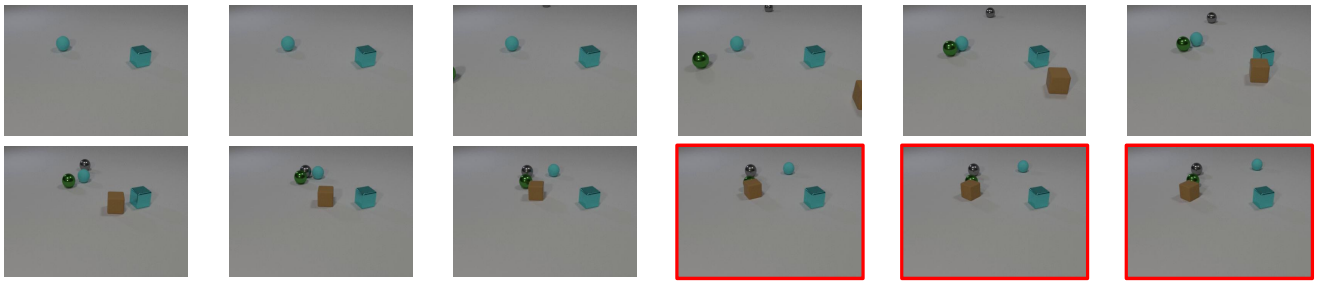


Figure 4. **A real example that the scene remains nearly unchanged** of *UniComp* on LLaVA-onvision-7B with 32 frames input under 10% retention ratio. The images are **high-fidelity** and can be **zoomed in to view details**. After Spatial Dynamic Compression, it shows the retained tokens with the selection order labeled white. Each line of detailed fusion map shows 10 tokens, same color token will be fused to the token with red rectangle. The order from top to bottom is the selection order, which means higher one is more unique.



Q: At the end, what object does the green ball collide with?

	Retention 20%	Retention 10%	Retention 5%
FastVid	The green ball collides with the brown cube.	The green ball collides with the brown cube.	The green ball collides with the cyan sphere.
HoliTom	The green ball collides with the cyan sphere.	The green ball collides with the cyan sphere.	The green ball collides with the cyan sphere.
UniComp	The green ball collides with the brown cube.	The green ball collides with the brown cube.	The green ball collides with the brown cube.

Q: The women shows three bottles in front of the camera with her hands. List them in detail.

	Retention 20%	Retention 10%	Retention 5%
FastVid	The bottles are labeled 'CRONDOX', 'IMUNE', and 'MAGNESIUM'.	The three bottles are labeled 'CRONoLITE', 'IMMUNE', and 'Malato Magnésico'.	The women shows three bottles in front of the camera: 'CRONoVIT', 'IMMUNE', and 'Ora-100'.
HoliTom	The bottles are labeled 'CRONoLIVE', 'IMMUNE', and 'Ora pro nobis'.	The bottles are labeled 'CRONDO', 'VITAMINA C', and 'IMMUNO'.	The bottles are labeled 'MAGNIFICA', 'MAGNIFICA', and 'MAGNIFICA'.
UniComp	The women shows three bottles in front of the camera with her hands. They are labeled "CronoVit," "Orna," and "MAGNÉSIO."	The three bottles shown by the women are labeled "CRONÓXIA," "IMMUNE," and "MALAGOS DE MAGNÉSIO."	The woman is holding three bottles in front of the camera. The first bottle is labeled "CRONÓXIA" and has a white cap. The second bottle is labeled "IMMUNE" and also has a white cap. The third bottle is labeled "MALAGOS" and has a blue cap.

Figure 5. Two examples of open-end question answer compare with FastVid and HoliTom.

Q: Briefly describe this video.

Baseline (no compression)

The video begins with a title screen displaying the words 'COCHINEAL' and 'ALUM' on a beige background, followed by an illustration of red and white substances being mixed with water. It then transitions to a blackboard with the word 'PIGMENT' written in white chalk. The scene develops through a series of frames showing colorful letters spelling out 'PIGMENT' on the blackboard, with each letter in a different color. Below the letters, there are seven test tubes, each containing a different colored liquid, arranged in a rainbow order from red to purple. The video continues with a close-up of the test tubes, now filled with colorful powders instead of liquids, also arranged in a rainbow order. The scene transitions to a blackboard with the word 'PIGMENT' written in white chalk again, followed by a series of frames showing the same colorful powders arranged in a rainbow order below the letters. The video concludes with a world map drawn on the blackboard, with lines connecting various locations across the continents.

FastVid

The video begins with a close-up of a red pigment being mixed with water, followed by the addition of white pigment. The mixture is then poured into a beaker, creating a colorful liquid. The scene transitions to a world map with green dots representing different locations. The map is zoomed in on South America, highlighting specific countries with more green dots. The focus shifts to a blackboard where the word 'Pigment' is written in colorful chalk, and a series of colorful pigments are displayed in a rainbow order. The video wraps up with a close-up of the colorful pigments on the blackboard, emphasizing their vibrant colors.

HoliTom

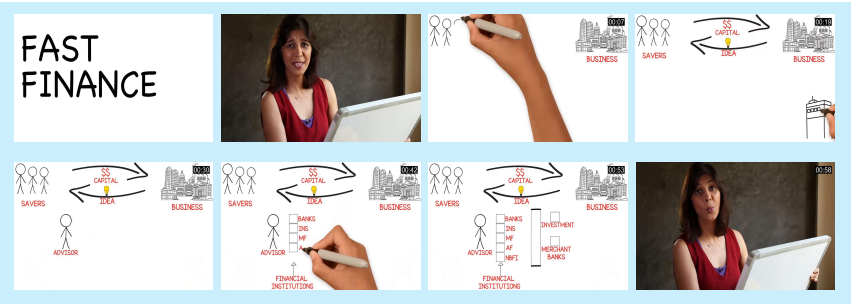
The video begins with a title 'COCHLEAN + ALUM' and an arrow pointing downwards, leading to a red background. It then transitions to a world map with green dots, followed by the word 'PIGMENT' on a blackboard with chalk drawings of paintbrushes in various colors. The video wraps up with a world map showing lines connecting different locations, symbolizing the journey of pigments around the world.

UniComp

The video begins with a title card that reads "COCHINELAL + ALUM + WATER," indicating the ingredients needed for the experiment. It then shows a series of steps where cochineal (a red dye) is mixed with alum and water, resulting in a red liquid. The mixture is then poured into a container, and the liquid turns a vibrant red color. The video then transitions to a world map highlighting countries where cochineal is used as a natural dye. The word "PIGMENT" is written on a chalkboard, followed by a sequence showing different colored pigments being applied to a surface, creating a rainbow effect. The video concludes with a blackboard displaying a colorful pattern created by the pigments, symbolizing the use of cochineal as a natural dye in various cultures around the world.

Figure 6. An caption examples compare with FastVid and HoliTom.

Source: Videomme (Gh_GtYtQoVI.mp4)



What is the function of the stopwatch in the upper right corner of the video?

A. To indicate the total duration of this video.
B. The video author made the annotations in order to explain clearly what finance is within one minute.
C. To display the current time when the narrator is sharing financial knowledge.
D. It is a countdown timer.

VisionZip	C
FastVid	C
HoliTom	C
UniComp	B

Source: Videomme (fo-mVfOsC-E.mp4)

How many people can be seen holding cameras in the video?

A. 2.
B. 3.
C. 4.
D. 1.

VisionZip	B
FastVid	B
HoliTom	D
UniComp	A

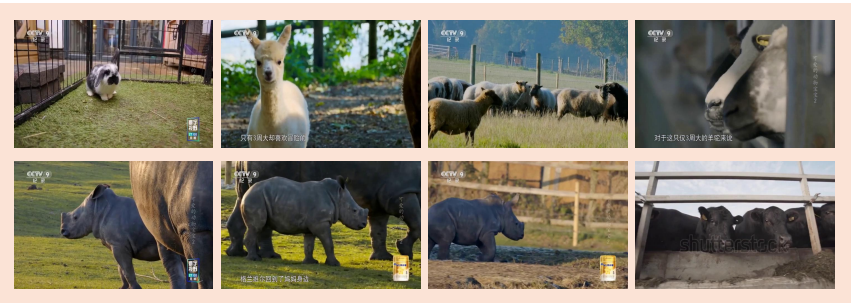
Source: MLVU (xiaoliyu_2.mp4)

What does the lobster use to hang the frog upside down?

A. Hand
B. Seaweed
C. Rope
D. Antennae

VisionZip	C
FastVid	C
HoliTom	C
UniComp	D

Source: MLVU (needle_26.mp4)



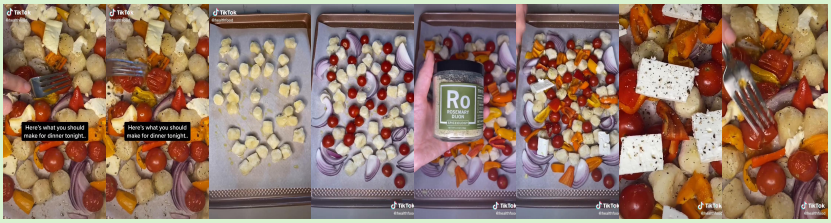
Where are the black cows with yellow tags on their ears eating hay on a sunny day?

A. In the barn
B. At the farm
C. In the forest
D. In the field

VisionZip	D
FastVid	D
HoliTom	D
UniComp	B

Figure 7. Examples on VideoMME and MLVU.

Source: LongVideoBench (@healthfood-6999715804623293702.mp4)

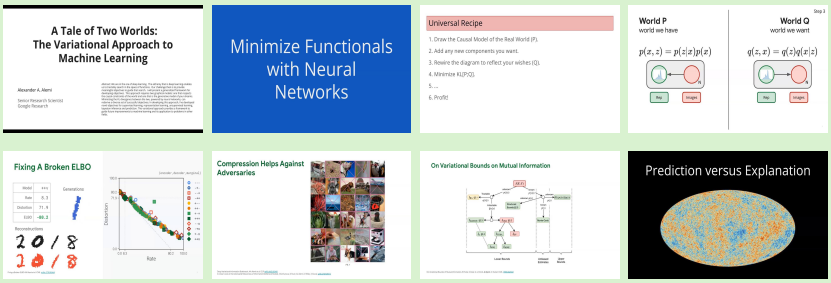


The screen shows a pile of colorful foods, with a white paper towel underneath, a silver plate below the paper towel, and the plate has floral patterns. What is not present in the screen?

A. Yellow bell peppers
B. Cherry tomatoes
C. Cauliflower florets
D. Carrot
E. White blocks sprinkled with seasoning

VisionZip	A
FastVid	A
HoliTom	A
UniComp	D

Source: LongVideoBench (OAHsR02dUc0.mp4)

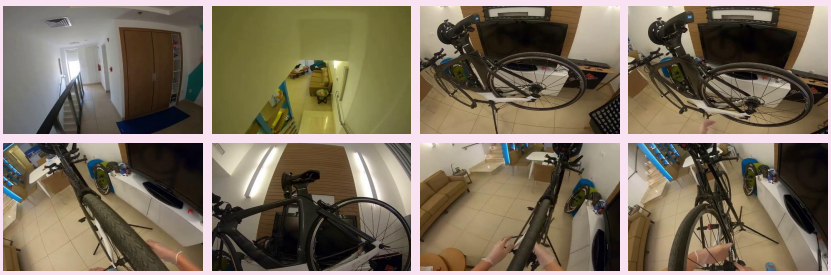


On a black screen, there are ten different colored particle-like objects sprayed on the screen, and at the bottom, there are circles of various colors from 0 to 9. After this, what happened to these different colored objects?

A. These differently colored objects became blurry and rectangular in shape
B. These differently colored objects became blurry and strip-like in shape
C. These differently colored objects became blurry and circular in shape
D. These differently colored objects were assembled together

VisionZip	B	HoliTom	D
FastVid	A	***UniComp***	C

Source: Egoschema (3a94a8d4-9b0f-49d9-9479-80ccf3a9ac3a.mp4)

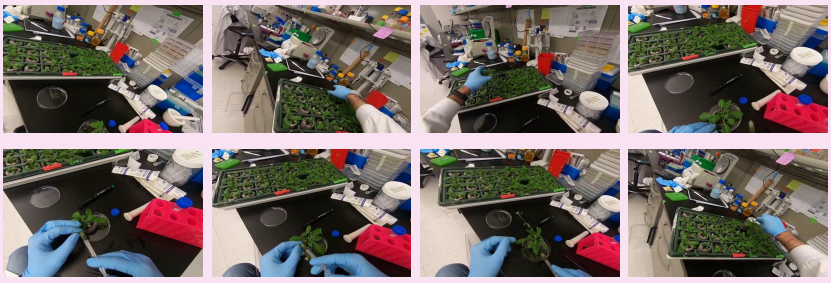


From the sequence of actions, identify a turning point or moment where c's focus shifts to a different task. explain why you believe this is the most significant part of the video.

A. The turning point is when c unfastens the hub axle.
B. The crucial turning point occurs when character c picks up the screwdriver from the table.
C. The pivotal turning point occurs when character c decides to put on the gloves.
D. The turning point is when c removes the tire.
E. The critical turning point occurs when character c successfully patches the hole, fixing it.

VisionZip	B	HoliTom	B
FastVid	B	***UniComp***	A

Source: Egoschema (3c63a6f6-842c-4d66-aa7f-2e36e1c73110.mp4)



What are the main objectives and activities that c is performing throughout the video? keep your answer concise and avoid listing all the actions.

A. C is watering the plants.
B. Currently, c is actively engaged in fertilizing the plants diligently.
C. Currently, c is carefully pruning and caring for the plants diligently.
D. C is performing an experiment on plants.
E. Currently, c is carefully repotting the various plants indoors.

VisionZip	C	HoliTom	E
FastVid	C	***UniComp***	D

Figure 8. Examples on LongVideoBench and Egoschema.