InstAP: Instance-Aware Vision-Language Pre-Train for Spatial-Temporal Understanding

Supplementary Material

7. Experiment Details

Hardware and Infrastructure. All training experiments are executed on a dedicated high-performance compute cluster. The infrastructure allocation is divided by training stage to optimize for computational demands:

- **Self-Supervised Masked Video Modeling:** Executed on 320 NVIDIA H100 GPUs (80 GB VRAM per GPU).
- **Instance-aware Alignment Learning:** Executed on 200 NVIDIA B200 GPUs (180 GB VRAM per GPU), allowing for efficient processing of larger frame samples.

We adopt PyTorch Distributed Data Parallel for multi-node training, as the entire model fits in a single GPU and thus data-parallel scaling is sufficient.

Training Configurations. For the initial self-supervised masked video modeling, we train for 800 epochs using an 80% masking ratio on 8-frame clips. The detailed hyperparameters for this stage are listed in Table 7. For the subsequent instance-aware alignment, we initialize from the best previous stage checkpoint and train for 15 epochs. In this stage, we increase temporal resolution to sample 16 frames per video. The joint optimization of global and instance-level alignment losses uses the hyperparameters detailed in Table 8.

Baseline Configurations. We evaluate state-of-the-art video-text models using their official checkpoints, inference code, and hyperparameters. Coca [64], ViCLIP [54], OpenCLIP [11], and UMT-L [29] use ViT-L backbones, while VideoPrism [70], CLIP4Clip [38], CLIP-ViP [60], SigLIP (ViT-B) [68], and MCQ [17] use ViT-B backbones. For UMT-L, we use the officially released *L* variant trained on the authors' 25-million image/video-text corpus, and train `UMT-L (InstVL; g)` and `UMT-L (InstVL; g+i)` on the same corpus as InstAP.

Dataset Curation. We use the following diverse specialized prompts to facilitate instance/global caption generation for images and videos:

image_caption_prompt: "You are a powerful image captioner. Create a detailed caption describing the image. Include the object types and color, object actions, precise object locations. Do not describe the contents in list form. Minimize aesthetic descriptions as much as possible. Do not repeat the same information in the caption. You must only describe the image."

image_instance_caption_prompt: "You are a powerful bounding box captioner. Create detailed caption describing the contents of the red colored bounding box in the image. Include the object types and color, object actions, precise object locations. If there is person in bounding box then include the person actions, expressions, precise person and object locations, etc. Instead of describing the imaginary content, only describe the content one can determine confidently from the bounding box. Do not describe the contents in list form. Minimize aesthetic descriptions as much as possible. Do not repeat same information in the caption. You must only describe about the content of bounding box, nothing else."

video_frame_caption_prompt: "You are an excellent video frame analyst. Utilizing your incredible attention to detail, you provide clear and detailed descriptions of a video frame. You excel in identifying and conveying actions, behaviors, environment, states and attributes of all the objects in image. ## Skills #### Skill 1: Describing Object Actions and Behaviors - Describe the action or behavior of objects within the image. #### Skill 2: Describing Environment and Background Variations - Elaborate on environment #### Skill 3: Describing Object Appearances - Describe the appearance of objects within the current frame. Do not repeat the information already included in previously. Do not describe the contents by itemizing them in list form. State facts objectively without using any rhetorical devices such as metaphors or personification. Descriptions should be fluent and precise, avoiding analyzing and waxing lyrical. Descriptions need to be concise, describing only the information that can be determined, without analysis or speculation. Do not mention the frame number and timestamp of the current frame. You must generate the caption in one paragraph. Here is the frame:"

video_frame_caption_merge_prompt: "You are an expert in generating a single detailed summary. ## Skills: Summarize sequentially, maintaining coherence between frames and the integrity of the timeline. Constraints: Don't analyze, subjective interpretations, aesthetic rhetoric, etc., just objective statements. - Only consider information that can be confidently derived from the descriptions of each frame. - Do not extrapolate or imagine, remove uncertain information. - No mention of specific frames index or timestamps. You are given a chronological sequence of captions for a sequence of frames. Generate a single detailed chronological caption for the video with the help the captions provided. The following are the captions for each frame in a video:"

video_instance_caption_prompt: "You are a powerful bounding box captioner. Create detailed caption describing the contents of the red colored bounding box in the image. Include the object types and color, object actions, precise object locations. If there is person in bounding box then include the person actions, expressions, precise person and object locations, etc. Instead of describing the imaginary content, only describe the content one can determine confidently from the bounding box. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible. Do not repeat same information in the caption. You must only describe about the content of bounding box."

To synthesize comprehensive video descriptions, we first utilize *video_frame_caption_prompt* to generate a caption for each sampled frame. Subsequently, these frame level descriptions are aggregated into a unified whole video caption using *video_frame_caption_merge_prompt*.

Dataset Filtering. For each instance, we compute CLIP similarity between its bbox crop (*crop*) and the first caption sentence (*sent1*); keep the instance iff $\max(s_{\text{crop}}, s_{\text{sent1}}) \geq$

τ ($\tau = 0.15$; minimum size of bounding boxes must be $\geq 50 \times 50$). For a video to be included in the final dataset, it must have at least one instance with CLIP score $\geq \tau$.

Qualitative Error Analysis. In Table 6, we report human-in-the-loop audit results for caption quality. We evaluate 1K video captions and 1K image captions following [10] on three aspects: **Omission (O)**, **Fabrication (F)**, and **Distortion (D)**. **O/F/D** $\in [0, 1]$ are normalized scores, where higher is better, and **S** $\in [0, 3]$ denotes the aggregate score.

Table 6. Caption quality audit (S=O+F+D; higher is better)

	Global				Instance			
Split	**O**	**F**	**D**	**S**	**O**	**F**	**D**	**S**
Video	0.93	0.47	1.00	2.40	0.87	0.33	1.00	2.20
Image	1.00	0.53	1.00	2.53	1.00	0.67	1.00	2.67

Table 7. Hyperparameters used for self-supervised masked video modeling, as presented in Sec. 4.1.

Hyperparameter	**Value**
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1.5×10^{-4}
Batch size	64
Total epochs	800
Warmup epochs	40
Input frames	8

Table 8. Hyperparameters used for global and instance alignment learning, as presented in Sec. 4.2.

Hyperparameter	**Value**
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1×10^{-4}
Batch size	64
Total epochs	15
Warmup epochs	1
Drop-path rate	0.2
Input frames	16
Image size (Student/Teacher)	224/196
Patch size (Student/Teacher)	16/14

8. Extended Evaluation on InstVL (img-zero)

In Table 9, we provide a comprehensive evaluation including R@1, R@5, and R@10 metrics across both 1K and 10K candidate splits for the `InstVL(img-zero)` benchmark. To ensure clarity and focus on the most competitive approaches, we omit lower-performing baselines such as Coca [64] and CLIP-ViP [60] from this detailed comparison.

The extended metrics on the *instance* split reveal that InstAP consistently outperforms prior methods across the retrieval ranking. On the challenging instance split with 1K candidates, InstAP achieves a V2T R@10 of 78.19%, surpassing the strongest baseline, `UMT-L (InstVL; g+i)`, by nearly 7.4 percentage points. This indicates that even when the exact ground truth is not the top prediction, InstAP places it within the immediate semantic neighborhood far more reliably than global-only models.

When the candidate pool increases to 10K, the difficulty of the retrieval task scales significantly. Notably, our method demonstrates higher resilience to this expanded search space. In the 10K instance split, InstAP outperforms `UMT-L (InstVL; g+i)` by 9.05 percentage points in V2T R@1 (31.87% vs. 22.82%). This suggests that our instance-aware alignment creates a more discriminative feature space that resists degradation better than global averaging when faced with a larger number of semantic distractors.

Finally, while our primary objective is fine-grained instance alignment, the results on the global split confirm that this does not come at the cost of general scene understanding. InstAP achieves comparable or superior performance to the strongest baselines on the global metrics, with only marginal gaps on a few R@10 scores, confirming that our joint training objective successfully injects local granularity without compromising the model's holistic representation capabilities.

9. Analysis of Baseline Limitations

To further understand the impact of our instance-aware alignment strategies, we present more qualitative analysis of failure cases exhibited by the global-only baseline [29]. We examine retrieval results on both fine-grained instance retrieval (`InstVL-1K (img-zero)`) and global video retrieval (MSR-VTT [58]).

Instance Retrieval. In the instance retrieval task, the model must align a cropped/instance-aware image query with a specific caption. As shown in Fig. 6, the baseline frequently suffers from semantic confusion. Because the baseline relies solely on global scene statistics, it retrieves captions that match the general "vibe" or texture of the image but fail to align with the specific object in the crop. For example, when querying a specific stack of books, the baseline retrieves a caption describing a "wooden chest" in a similar room setting. Similarly, for a query of a specific person, the baseline retrieves a generic description of a woman with similar hair color ("dark hair"), ignoring the specific identity or context. In contrast, InstAP correctly distinguishes these fine-grained details.

Table 9. Results on `InstVL(img-zero)`. We report T2V/V2T R@1, R@5, and R@10 on the *instance* and *global* splits for 1K and 10K candidates. Best results in each column are highlighted in bold.

Method	Split	InstVL(img-zero)											
		1K						10K					
		T2V			V2T			T2V			V2T		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VideoPrism [70]	Instance	21.32	50.23	58.37	27.39	56.65	63.89	13.85	35.22	42.65	20.04	42.00	49.21
	Global	85.70	89.70	**90.60**	85.80	90.20	**90.80**	73.05	85.32	87.34	75.11	86.18	88.01
CLIP4Clip [38]	Instance	17.82	42.57	49.84	25.10	53.74	61.44	9.11	24.96	31.98	16.30	35.77	42.79
	Global	78.20	87.60	89.10	81.70	89.30	90.00	56.95	76.25	80.96	63.96	80.25	84.20
ViCLIP [54]	Instance	18.25	37.28	43.81	20.93	40.97	47.51	9.57	21.84	27.19	11.21	23.65	29.20
	Global	77.80	88.30	89.30	77.60	87.80	89.10	58.51	76.08	80.58	58.21	76.02	80.56
OpenCLIP [11]	Instance	26.73	48.13	55.25	36.19	61.40	66.73	17.28	34.53	40.29	25.57	46.39	52.05
	Global	83.40	88.70	89.20	86.90	90.00	90.20	70.75	82.17	84.62	78.13	86.47	87.84
SigLIP [68]	Instance	28.25	50.08	57.35	35.56	60.93	65.72	16.98	34.06	39.91	25.19	45.14	50.83
	Global	83.90	88.90	89.60	86.50	90.10	90.40	68.64	81.90	84.87	75.66	86.11	87.96
MCQ [17]	Instance	17.08	37.74	46.69	19.61	42.22	50.39	7.04	17.81	23.72	8.55	21.07	27.54
	Global	58.90	79.30	83.60	62.70	81.70	85.80	34.13	55.88	64.24	38.26	59.78	68.05
UMT-L (InstVL; g) [29]	Instance	25.97	51.82	59.47	31.97	61.36	68.22	13.33	29.93	36.05	19.21	38.47	44.58
	Global	85.30	89.80	90.00	86.40	**90.40**	90.40	72.50	82.98	84.67	74.18	84.78	86.25
UMT-L (InstVL; g+i) [29]	Instance	34.68	63.99	69.28	34.99	64.97	70.81	21.13	42.28	48.05	22.82	43.03	48.42
	Global	82.40	89.50	90.10	84.30	89.90	90.50	68.16	80.47	83.02	69.76	82.84	85.25
InstAP (Ours)	Instance	**41.94**	**71.40**	**76.19**	**42.53**	**73.40**	**78.19**	**28.25**	**51.80**	**56.56**	**31.87**	**54.14**	**58.15**
	Global	**88.70**	**90.40**	90.50	**88.30**	**90.40**	90.40	**83.33**	**87.37**	**87.95**	**82.21**	**88.13**	**88.68**

Global Retrieval. While the baseline establishes a robust standard for global video retrieval, it still exhibits limitations in fine-grained semantic grounding, as shown in Fig. 7. We observe two primary failure modes. First, regarding entity hallucination, the baseline occasionally retrieves captions containing objects absent from the video; for instance, in the last row, it misidentifies rocks in a water filtration clip as "two snakes," likely misled by textural similarities. Second, regarding identity confusion, the baseline struggles to distinguish specific public figures in news or interview settings (Rows 1 and 2), misidentifying "Donald Trump" or "Bill Murray" as "Chris Christie," potentially due to overfitting on common newsroom contexts during pre-training. InstAP, leveraging explicit object-level supervision, demonstrates significantly higher fidelity to the actual video content.

10. Attention Visualization

To interpret the spatial-temporal reasoning of our model, we visualize the regions driving the cross-modal alignment decisions by analyzing the gradients flowing into the global visual token sequence $\mathbf{V} \in \mathbb{R}^{L_v \times d}$ (defined in §3.3).

Let $s \in \mathbb{R}^2$ denote the logits produced by the matching head h. We identify the features contributing to a successful alignment by backpropagating from the positive match score $y = s_1$.

We calculate the importance weight $\alpha_k \in \mathbb{R}$ for the k-th feature channel by global average pooling the gradients over the spatial-temporal sequence:

$$\alpha_k = \frac{1}{L_v} \sum_{l=1}^{L_v} \frac{\partial y}{\partial \mathbf{V}_{l,k}}. \tag{14}$$

These weights quantify the sensitivity of the matching score to specific feature dimensions. The final localization map $\mathcal{M} \in \mathbb{R}^{L_v}$ is obtained by a weighted combination of the forward activation maps, followed by ReLU rectification to isolate regions with a positive influence:

$$\mathcal{M}_l = \mathrm{ReLU}\Big(\sum_{k=1}^{d} \alpha_k \mathbf{V}_{l,k}\Big). \tag{15}$$

The resulting map $\mathcal{M}$ is reshaped from the flattened sequence to the spatial-temporal grid (T, N) and upsampled to the input resolution of sampled frames and spatial patches responsible for the prediction. We present additional visualization results in Fig. 8.

11. InstVL Dataset Gallery

We present a gallery from InstVL in Fig. 9 and Fig. 10. Each example combines a global scene caption with fine-grained captions attached to individual instances, using consistent IDs and color-coded bounding boxes for videos and object-level bounding boxes for images.

12. Ethical Considerations and Data Filtering

InstVL is constructed on top of large-scale web datasets (e.g., LAION [45], COYO [5], HD-VILA [59]), which may

Figure 6. Across a variety of query instances in the `InstVL-1K (img-zero)`, our model InstAP consistently retrieves the correct, fine-grained description, whereas the global-only baseline [29] is easily confounded by semantic distractors, mismatching the query with a different caption.

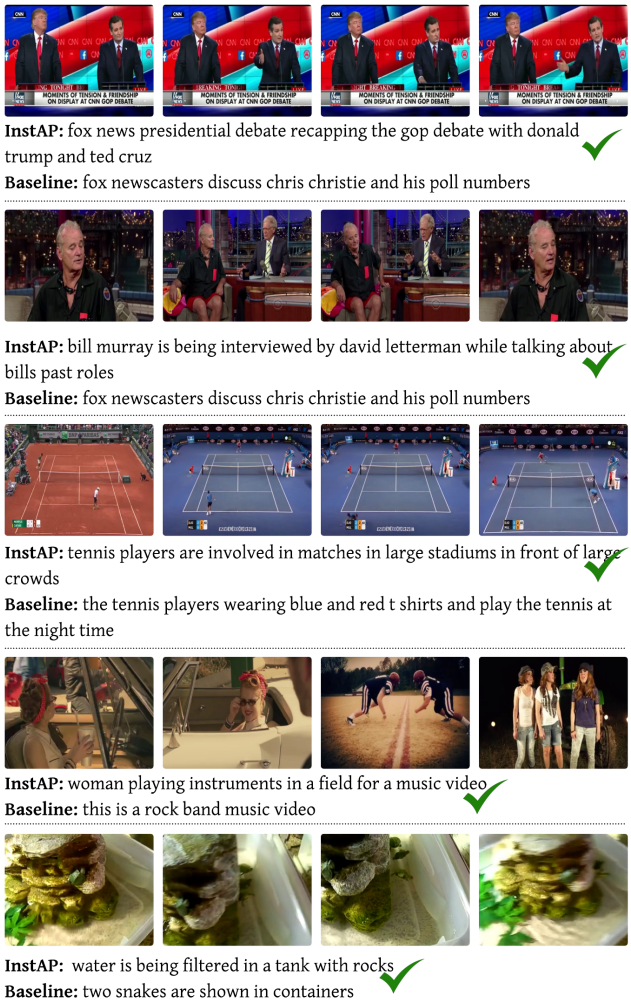


Figure 7. Qualitative comparison on the MSR-VTT [58] video retrieval task. The baseline model frequently exhibits hallucinations (e.g., perceiving 'snakes' in a rock tank) or identity confusion (e.g., misidentifying the speakers in news clips). In contrast, InstAP accurately grounds the specific actions, settings, and entities present in the video.

contain low-quality, NSFW, or otherwise inappropriate content. To mitigate this, we use a data-centric pipeline with automatic filtering for basic image/video quality, NSFW detection, text quality (length, special characters, repetition, flagged terms), and CLIP-based visual-text alignment, discarding samples that fail any of these checks. For re-captioned subsets, we retain only captions with high CLIP alignment to the visual content and remove samples with ambiguous crops or other sensitive entities.

Human Verification and Limitations. In early iterations, a team of annotation experts manually inspected random samples and edge cases (e.g., visually complex scenes, ambiguous instance crops), and their feedback was used both to refine thresholds and filtering rules and to improve the annotation pipeline (e.g., captioning and bounding-box quality). While these safeguards substantially improve data quality, automatic filters and spot checks remain imperfect. InstVL is therefore intended for research use only, and not for applications such as surveillance, biometric identification, or other high-stakes decision-making.

Data Release. To respect copyright and platform terms of service, we do *not* redistribute raw images or videos. We release only InstVL annotations (bounding boxes, instance IDs, captions) together with URLs and metadata (e.g., timestamps, frame indices) needed to retrieve the corresponding segments from the original platforms.

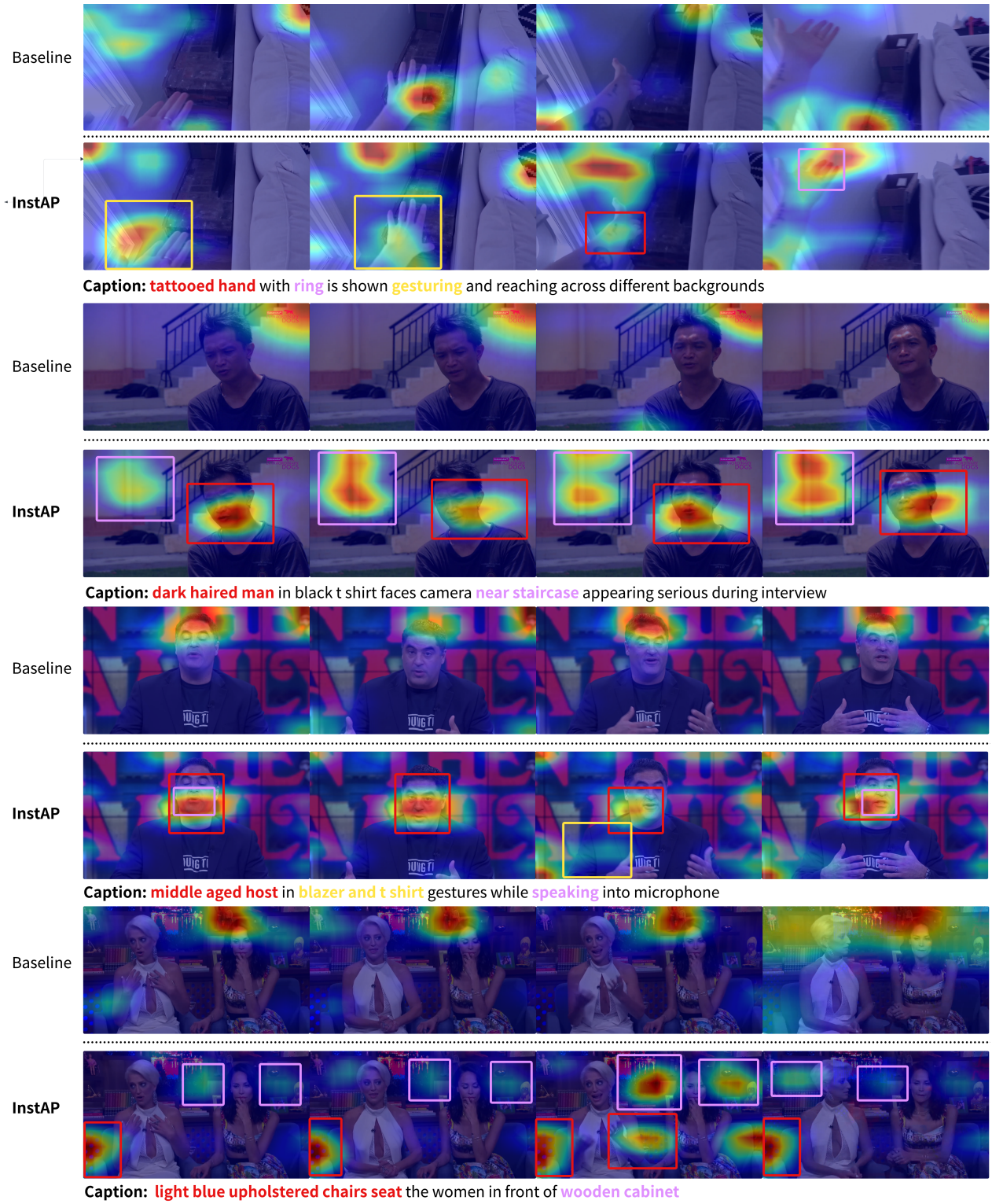
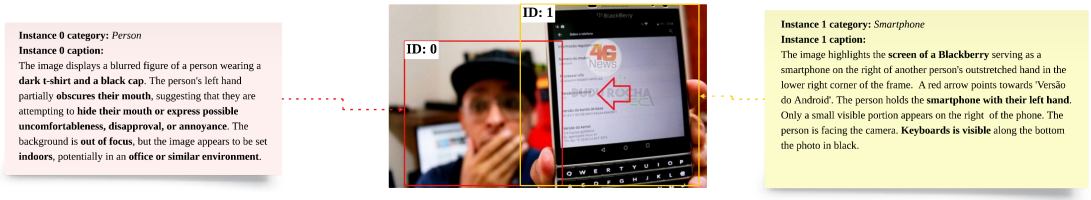


Figure 8. InstAP tends to attend more closely to caption-relevant regions, effectively isolating small, dynamic details (e.g., 'tattooed hand'), capturing environmental context (e.g., 'staircase'), and correctly targeting non-salient background objects (e.g., 'upholstered chairs,' 'wooden cabinet') even in the presence of dominant human faces. In contrast, the global-only baseline [29] exhibits diffuse and often misaligned attention, frequently drifting towards background clutter or generic salient regions rather than the specific entities described in the text.

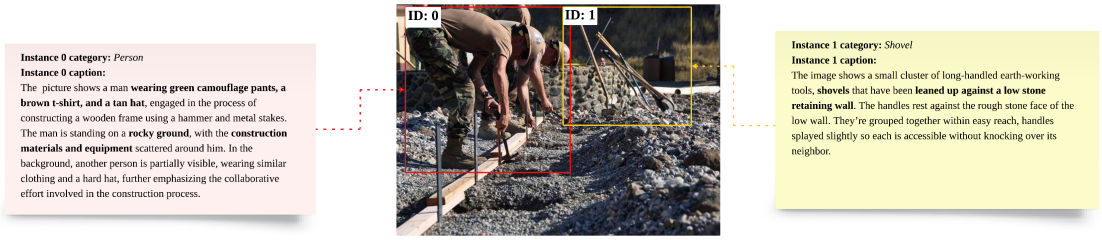
Figure 9. Illustration of annotation for videos in the *InstVL* dataset. For a single video segment, we display sampled frames, each overlaid with color-coded bounding boxes and unique instance IDs that remain consistent across time. The bottom text shows the global caption describing the whole scene together with the fine-grained captions automatically paired to each tracked instance shown on top.

Global Caption:
The image depicts a **person riding a bicycle** down an urban street. The bicycle is **black with yellow accents**. The person is wearing **blue jeans** and a **blue long-sleeved shirt** underneath a white jacket. They are wearing a **white helmet** and have a **backpack** strapped to their back. The person is **leaning forward**, holding onto the **handlebars as they pedal**. The backpack is black and has several pockets and compartments. The background of the image is a **city street**. The street is made of concrete and has several buildings lining it. The **buildings are made of brick** and have **graffiti** on them. The overall atmosphere of the image is one of adventure and exploration. The person appears to be enjoying their ride and taking in the sights around them. Overall, the image captures a moment of freedom and joy as the person rides their bicycle through the city streets.

Global Caption:
The image depicts a hand grasping a **Sony camera lens** against a textured background. The lens is held by **a left hand**, with the thumb on top and fingers wrapped around the bottom. It features a dark grey color scheme and is labeled with white text, indicating that it is a 'ULTRA WIDE CONVERTER' lens for a '**FE 2/28**' camera. The lens is positioned against a background that appears to have a **woven pattern of black and grey threads**.



Global Caption:
The image shows a person **holding a smartphone with an app** on the screen. The purpose of the image is to **demonstrate the use of an app**. The man is wearing a **black hat and has his hand covering his mouth**. He is holding a phone in his other hand. The app is showing a list of information on the screen. The image suggests that the person is using an app to access news or **some information on their smartphone**. The app appears to provide a range of categories and topics for users to explore.



Global Caption:
The image depicts **two men engaged in construction work on a sunny day**. The man on the left, with his head cropped out of the frame, wears a **camouflage outfit and a brown t-shirt**, and is crouched down, with a long **wooden plank nearby and a metal hook in his hands**. The man in the foreground, **wearing a tan hat, brown t-shirt, and military fatigues, leans over the wood plank, grasping the hook between his hands**. In the background, **a stone wall and a brown shed** are visible, with a pile of wooden posts and pipes leaning against the wall. A person's head is partially visible to the right of the shed, while a **grassy hill rises in the distance**. The ground is covered in **gravel and dirt**, with the men's shadows cast on the ground. A small portion of the **building is visible** on the left side. The image captures a moment of **construction activity** on a sunny day.

Figure 10. Illustration of annotations in the images of the *InstVL* dataset. The bottom text shows the global caption describing the whole scene together with the fine-grained captions paired to each bounding box on the sides.