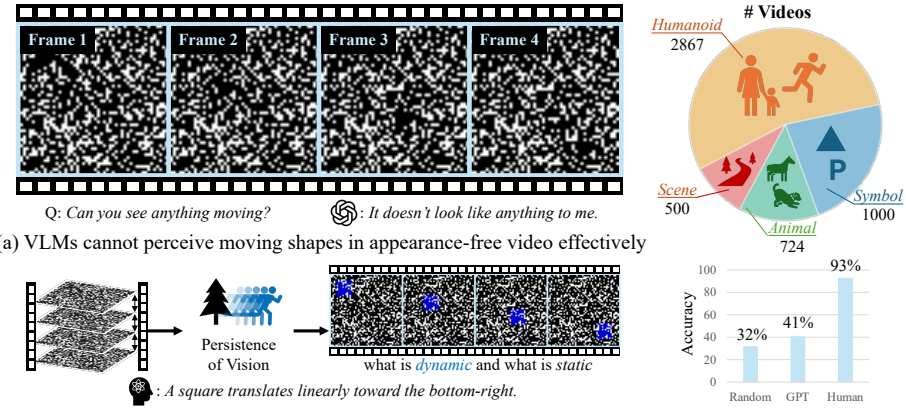
APPEARANCE-FREE VIDEO GENERATION FOR MEASURING A PERCEPTION GAP IN MOTION-DEFINED FORMS

Anonymous authors
Paper under double-blind review



(a) VLMs cannot perceive moving shapes in appearance-free video effectively

(b) Humans can see moving shapes in appearance-free video

(c) Our AppearFreeVQA

Figure 1: **(a)** Current VLMs fail to detect motion in videos with minimal appearance cues, perceiving only noise. **(b)** Humans leverage persistence of vision to perceive motion and identify the underlying shape (blue highlight is for visualization only). **(c)** We create *AppearFreeVQA*, a benchmark of appearance-free motion videos across four categories: symbol, humanoid, animal, and scene. Our analysis shows a significant perception gap between frontier VLMs and human performance. GPT-5.1 (OpenAI, 2025a) achieves near-chance-level performance, while humans consistently demonstrate the ability to answer these questions. Please find video examples in the supplementary materials.

ABSTRACT

VLMs have achieved remarkable progress on video understanding benchmarks, yet it remains unclear whether they are truly as capable as humans. A key aspect of human perception is *persistence of vision*. We can identify motion patterns in videos whose individual frames appear as pure noise (without any appearance or recognizable content). To quantify this capability for VLMs, we propose an algorithm to generate *appearance-free videos* from any given source of motion. Each frame is indistinguishable from noise, and only motion contains the information. We then introduce AppearFreeVQA, a benchmark with ≈5,100 appearance-free videos spanning four motion categories and three difficulty levels. We demonstrate that humans achieve approximately 93% accuracy, while frontier VLMs (GPT-5.1) reach only 41.7%, slightly above random chance. Oracle experiments show that supplying target optical flow improves Qwen3.5-9B from 14.3% to 66.8%, indicating that motion extraction is a major bottleneck. Finally, we show that training on AppearFreeVQA can improve VLMs' capabilities on camouflaged animal detection, a real-world application where appearance cues are limited.

1 INTRODUCTION

The human visual system is known to be remarkably capable at perceiving structure from motion cues (Ullman, 1979). Humans exploit temporal patterns of movement alongside appearance to understand the world (Palmer, 1999; Gibson, 1954; 2014; Burr, 2014). This is demonstrated by classic experiments with the Point-Light Walker (PLW) displays (Johansson, 1973; Pollick et al., 2005), in which moving dots are shown, and people can immediately recognize a walking person. Another example is the Random Dot Kinematogram (RDK) experiments (Braddick, 1973; Van Doorn

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

& Koenderink, 1982; Julesz, 1971) where global motion is reliably detected even when every individual frame is pure noise. More broadly, motion-defined forms, *i.e.*, visual structures that emerge exclusively from temporal pixel displacements, can be reliably perceived by humans even in the absence of appearance or color cues (Regan, 1989; Regan et al., 1992; Blake & Lee, 2005).

Recently, VLMs have achieved impressive performance across a broad range of visual understanding tasks (Zhang et al., 2024; Sun et al., 2019; Hong et al., 2025; Carreira et al., 2019). Frontier models, *e.g.*, GPT (OpenAI, 2025a) and Gemini (Google, 2025b), alongside open-weight VLMs, *e.g.*, QwenVL (Bai et al., 2025b;a) and InternVL (Wang et al., 2025), now demonstrate near-human or even "superhuman" performance on several video-understanding benchmarks (Li et al., 2024; Fu et al., 2025; Liu et al., 2024). This raises a natural question: *are VLMs truly as capable as humans in all aspects of perception?*

Mainstream VLM benchmarks evaluate motion understanding through natural videos, where appearance cues, *e.g.*, textures, scene context, and color, are readily available. Hence, it remains unclear whether these models are capable of perceiving motion in a fashion that is similar to humans. A VLM might correctly answer "*Is the dog running?*" by recognizing the dog's legs lifting in the first frame rather than identifying motion across the whole video. In Marr's terms (Marr, 2010), such models are heuristic programs: they produce correct answers without actually perceiving motion. To identify such behavior, we take inspiration from the random-dot stereogram example from Marr (2010). We synthesize videos where each frame is seemingly pure noise, *i.e.*, *appearance-free*, but across frames, humans can identify moving objects. As illustrated in Fig. 1(a), VLMs fail to recognize the moving square, whereas humans can perceive it moving towards the lower-right (see Fig. 1(b)). **Please see the supplement for better illustrations of these appearance-free videos.**

To systematically study this human-VLM gap, we introduce AppearFreeVQA for *appearance-free* motion reasoning. We propose a generation pipeline that turns any reference motion, whether real or synthetic, into an appearance-free video. Each frame is indistinguishable from background noise, yet encodes motion cues that are obvious to the human eye. We cover four categories (symbols, humanoids, animals, and driving scenes) to provide a testbed across different contexts; see Fig. 1(c).

On AppearFreeVQA, human annotators achieve approximately 93% accuracy across all difficulty levels, while the best-performing VLM, GPT-5.1, reaches only 41.7%. The gap is more prominent at the *fine* level, where models must answer specific motion details. Oracle experiments further show that providing target optical flow improves Qwen3.5-9B from 14.3% to 66.8%, pointing to motion extraction as a major bottleneck (Sec. 4.2). Beyond the benchmark, we demonstrate that training on AppearFreeVQA improves performance on camouflaged animal detection (Lamdouar et al., 2020), where appearance cues alone make the animals difficult to detect.

Our main contributions are as follows:

- We identify and quantify a gap between humans and VLMs in motion perception.
- We propose a two-step video generation pipeline (`ForwardWarp` + `MotionOptim`) that produces appearance-free videos with controlled motion. At the same time, each frame is statistically indistinguishable from noise.
- We introduce AppearFreeVQA, a benchmark designed to isolate motion perception from appearance reasoning, comprising approximately 5,100 videos and 17,600 questions across four motion categories and three difficulty levels.

2 Related Work

Human vision and motion perception. Humans perceive structure and motion from visual stimuli through mechanisms studied across decades of psychophysical and computational research (Palmer, 1999; Marr, 2010; Burr, 2014). Visual changes are interpreted into coherent percepts, *i.e.*, patterns (Johansson, 1973), structures (Petersik, 1987), and causal events (Michotte, 2017), without rich appearance cues. Point-Light Displays (Johansson, 1973; Pollick et al., 2005) demonstrate that biological motion is perceivable from as few as twelve moving dots. Random Dot Kinematograms (RDKs) (Braddick, 1973; Van Doorn & Koenderink, 1982; Scase et al., 1996) show that the human visual system extracts global motion from individually uninformative signals through cortical spatiotemporal integration (Britten et al., 1992). Motion-defined forms (Regan, 1989) are objects whose

boundaries emerge exclusively from differential motion rather than luminance or texture. They are perceived by humans even when every frame is uninformative (Regan et al., 1992). In stereopsis, depth-defined forms are similarly perceived from binocular disparity alone (Julesz, 1971; Howard & Rogers, 2012). Spatiotemporal energy models (Adelson & Bergen, 1985) and structure-from-motion analyses (Ullman, 1979) formalized these mechanisms computationally. In summary, motion provides useful cues independent of appearance (Gibson, 2014; Nakayama, 1985; Marr, 2010).

In terms of computer vision, early two-stream architectures (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2016) explicitly separated appearance (RGB) and motion (optical flow) branches. SlowFast (Feichtenhofer et al., 2019) further decomposed these into a slow appearance pathway and a fast motion pathway. Despite this decomposition, modern VLMs mainly rely on one vision branch for appearance cues, with motion-energy processing (Tangemann et al., 2024) as an exception. It enables zero-shot generalization to random-dot stimuli by modeling motion energy instead of appearance.

Vision-language models. Proprietary VLMs, *e.g.*, GPT (OpenAI, 2025a) and Gemini (Google, 2025b), and open-source models, *e.g.*, Qwen (Bai et al., 2025b;a), InternVL (Wang et al., 2025), and GLM (Team et al., 2025), achieve near/superhuman performance on a broad range of benchmarks (Carreira et al., 2019; Xu et al., 2016; Li et al., 2024; Liu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Meng et al., 2026; Huang et al., 2026). A recent line of work probes where VLMs diverge from human perception: visual illusions and anomalies (Hou et al., 2026), color-blindness stimuli (Ling et al., 2025), and Dalmatian-style camouflage (Wiedemer et al., 2025). Our work identifies a more fundamental gap, namely that VLMs fail to perceive objects defined purely by motion when appearance cues are absent, even though such motion is immediately obvious to humans.

Motion-aware benchmarks. MotionBench (Hong et al., 2025) evaluates fine-grained motion understanding in natural videos. HuPerFlow (Yang et al., 2025) benchmarks optical flow estimation for humans *vs*. machines. Two recent benchmarks target motion-defined stimuli more directly. SpookyBench (Upadhyay et al., 2026) uses noise-based stimuli to expose "time blindness" in VLMs, but restricts generation to opposing-motion patterns for simple text, shapes, or depth maps, limiting category diversity and realism. The ActPLD benchmark (Kadambi et al., 2025) evaluates VLMs on point-light biological motion, showing gaps in sparse human-motion understanding. Our proposed AppearFreeVQA differs from the above: it has (a) diverse motion categories (symbols, humanoids, animals, and driving scenes), (b) a controllable two-step pipeline that converts *any* reference motion **(not just hand-crafted rule-based motion)** into an appearance-free video, and (c) three difficulty levels (coarse/medium/fine) to provide fine-grained understanding of the human-VLM gap.

3 APPEARFREEVQA

We systematically evaluate the gap between human vision and VLM capabilities in appearance-free motion reasoning. While humans excel at perceiving motion patterns from minimal appearance cues, it remains unclear whether VLMs possess similar abilities. We design AppearFreeVQA to answer: *Can VLMs understand motion when there are limited appearance cues?*

Task definition. Given an input video sequence $\boldsymbol{V} = [\boldsymbol{X}^{(n)}]_{n=1}^{N}$, where each frame is denoted $\boldsymbol{X}^{(n)} \in \mathbb{R}^{C\times H\times W}$, a generic VQA benchmark evaluates the model's ability to answer questions about the content of $\boldsymbol{V}$. In appearance-free VQA, we say $\boldsymbol{V}$ is *appearance-free* if each frame $\boldsymbol{X}^{(n)}$ contains minimal visual information. That is, it is challenging for human eyes to distinguish between $\boldsymbol{X}^{(n)}$ and a sample drawn from a noise distribution $\mathcal{N}$, *e.g.*, salt-and-pepper or Gaussian noise. By observing the changes between frames in $\boldsymbol{V}$, a model with strong motion perception capability should perceive *motion-defined forms* (Regan, 1989; Regan et al., 1992), visual structures that emerge from motion information rather than static appearance cues. These include geometric transformations of shapes or biological motion patterns from articulated movement. In other words, a model would be able to answer questions regarding these motion-defined forms, such as *"Is there any motion in this video?"*, *"What object is moving?"*, or *"How is the object moving?"*.

While underlying motion-defined forms can be specified in multiple ways, *e.g.*, trajectories or text descriptions, we aim to control motion precisely and thus adopt an optical flow representation. Our generation pipeline takes *any* input video, computes optical flow, and turns it into an appearance-free video. In the following, we denote the target forms via an optical flow sequence of the moving object $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$, where $\boldsymbol{F}^{(n)} \in \mathbb{R}^{2\times H\times W}$ specifies pixel displacement from frame $n-1$ to n.

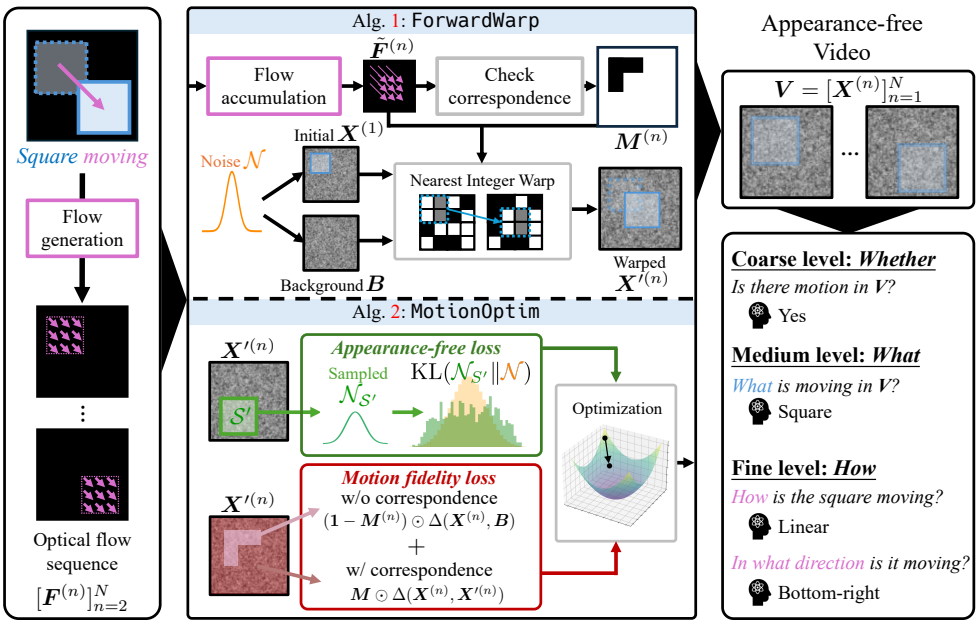


Figure 2: **AppearFreeVQA generation pipeline.** (a) We start with a motion-defined form, *e.g.*, a moving square, and compute its optical flow $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$ (Sec. 3.1). (b-c) We generate appearance-free videos $\boldsymbol{V}$ through forward warping from noise-initialized $\boldsymbol{X}^{(1)}$ and $\boldsymbol{B}$, followed by motion optimization that balances two objectives (Sec. 3.2). Appearance-free loss enforces local noise-like statistics on randomly sampled $\mathcal{S}'$, while motion fidelity loss ensures consistency between $\boldsymbol{V}$ and $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$. (d) We then generate questions to evaluate motion perception (Sec. 3.3).

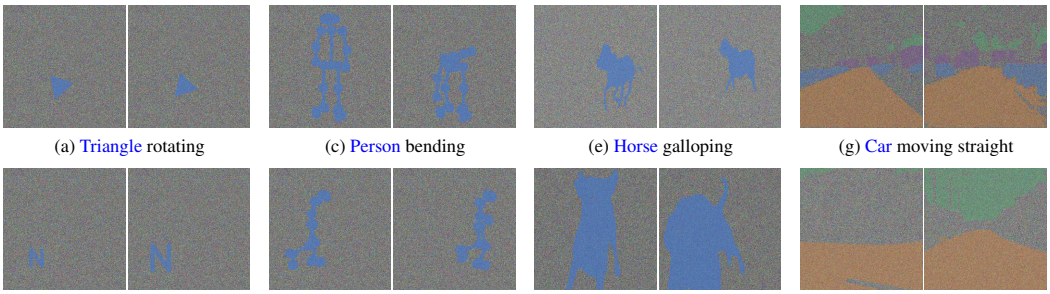
Generation overview. The overall method for generating appearance-free video is shown in Fig. 2 and contains the following three steps: (a) **Motion input** (Sec. 3.1): We identify a target motion-defined form and represent it using an optical flow sequence $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$. (b) **Video generation** (Sec. 3.2): We generate an appearance-free video $\boldsymbol{V}$ that matches $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$ while maintaining noise-like statistics $\mathcal{N}$. (c) **Question generation** (Sec. 3.3): We generate multi-level questions following the *whether-what-how* hierarchy: the coarse level detects motion presence, the medium level identifies the moving object, and the fine level analyzes detailed motion characteristics.

3.1 Motion-Defined Form

To evaluate VLMs' capabilities on appearance-free videos, we design AppearFreeVQA to include videos spanning a wide range of motion and object types. We choose four categories with different levels of *object realism* and *motion realism*: (a) **Symbol** uses fully synthetic objects (geometric primitives) with analytically defined, rule-based motion. This includes synthetic objects with synthetic motion. (b) **Humanoid** uses synthetic skeletal representations (point-light displays) driven by real human motion-capture data. As in (a), these are also synthetic objects with real motion. (c) **Animal** uses real objects with real motion, estimated from natural video, representing the hardest setting. (d) **Scene** uses synthetic driving video with realistic ego-motion, representing scene-level motion. This progression lets us study VLM performance across levels of motion complexity and realism. We now briefly describe and show the categories.

Symbol (Fig. 3(a-b)) consists of motion from simple primitives, *e.g.*, shapes, digits, and letters, providing a controlled setting for fundamental geometric transformations. We synthesize target forms by defining parametric transformations and computing the corresponding optical flow. We generate motion sequences including translation, rotation, and scaling. The optical flow $[\boldsymbol{F}^{(n)}]$ is then analytically derived from these transformations.

Humanoid (Fig. 3(c-d)) consists of motion of human actions, providing a natural setting for realistic motion perception. We utilize the MPOSE dataset (Mazzia et al., 2022), which provides motion



(a) Triangle rotating (c) Person bending (e) Horse galloping (g) Car moving straight

(b) Character N coming closer (d) Person running (f) Dog jumping (h) Car turning left

Figure 3: **Appearance-free video V examples.** (a-b) **Symbol** category with digits, shapes, and characters. (c-d) **Humanoid** category with humanoid skeletal keypoint sequences. (e-f) **Animal** category with selected animals. (g-h) **Scene** category with driving video. For (a-f), the blue overlay denotes the single moving object's mask. For (g-h), we overlay the mask of road, building, sky, and other cars. All overlays are shown for visualization only and are not present in the actual data.

capture as point-light display keypoint sequences. The dataset contains various human actions, *e.g.*, running, jumping, and other dynamic movements. We extract 2D keypoint sequences and compute $[\boldsymbol{F}^{(n)}]$ from frame-to-frame keypoint displacements.

Animal (Fig. 3(e-f)) consists of motion from animal locomotion, providing a challenging setting for complex biological motion perception. We utilize the TigDog dataset (Del Pero et al., 2015; 2016), which includes a collection of animal video clips of dogs, tigers, and horses. They are annotated with action labels, landmarks, and segmentation masks. Each video involves actions such as walking, opening the mouth, or turning the head.

Scene (Fig. 3(g-h)) consists of motion from driving footage, testing whether VLMs can perceive scene-level appearance-free ego-motion. We utilize Virtual KITTI 2 (Cabon et al., 2020), a synthetic rendering of real KITTI (Geiger et al., 2013) driving sequences with dense, per-pixel target $[\boldsymbol{F}^{(n)}]$. We derive 4 actions, *i.e.*, straight, left-turn, right-turn, and reverse, from their camera-pose trajectories.

3.2 Appearance-free Video Generation

Given a motion-defined form represented by an optical flow sequence $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$, we generate appearance-free videos where individual frames resemble noise $\mathcal{N}$ while consecutive frames exhibit the specified form. We introduce a two-step approach (see Fig. 2(b)). For each frame, we first generate a warped frame via `ForwardWarp`, which is effective for simple motion. To handle complex motion, we take the result from `ForwardWarp` and further refine it with `MotionOptim`, an optimization-based approach that balances per-frame noise statistics and motion.

Forward warping. Starting from a noise-initialized first frame $\boldsymbol{X}^{(1)} \sim \mathcal{N}$, `ForwardWarp` (Alg. 1) generates each subsequent frame while handling occlusions. $\text{Warp}(\boldsymbol{X}, \boldsymbol{F})$ splats each source pixel (u, v) to its destination $(u', v') = \text{round}((u, v) + \boldsymbol{F}(u, v))$. To avoid error accumulation, we warp directly from $\boldsymbol{X}^{(1)}$ using the accumulated flow $\tilde{\boldsymbol{F}}^{(n)}$, yielding $\boldsymbol{X}'^{(n)} = \text{Warp}(\boldsymbol{X}^{(1)}, \tilde{\boldsymbol{F}}^{(n)})$. Next, a correspondence mask $\boldsymbol{M}^{(n)} \in \{0, 1\}^{H \times W}$ is computed to track which destination pixels receive a splatted value; pixels without correspondence ($\boldsymbol{M}^{(n)} = 0$) are filled with background noise $\boldsymbol{B} \sim \mathcal{N}$, preserving the noise across all frames.

Motion optimization. Motion-defined forms, *e.g.*, running horses, can involve large displacements and occlusions, which lead to undesirable artifacts in the warped frames from `ForwardWarp`. We address this through the second step, `MotionOptim` (Alg. 2), which further performs gradient-based refinement. After obtaining an initial warped frame $\boldsymbol{X}^{(n)} \leftarrow \boldsymbol{X}'^{(n)}$, we optimize $\boldsymbol{X}^{(n)}$ as follows:

$$\min_{\boldsymbol{V}} \sum_{n} \mathcal{L}(\boldsymbol{X}^{(n)}) = \min_{\boldsymbol{V}} \sum_{n} (1 - \lambda)\, \mathcal{L}_{\text{appear}}(\boldsymbol{X}^{(n)}) + \lambda\, \mathcal{L}_{\text{motion}}(\boldsymbol{X}^{(n)}), \tag{1}$$

where $\mathcal{L}_{\text{appear}}$ enforces noise-like statistics and $\mathcal{L}_{\text{motion}}$ ensures motion fidelity.

Appearance-free loss. To preserve a noise-like appearance, we minimize the KL divergence on the local patches of the frames. A KL divergence over all pixels would force the entire frame to

Algorithm 1 `ForwardWarp`

Input: Flow $[\boldsymbol{F}^{(n)}]_{n=2}^{N}$, noise $\mathcal{N}$
Output: Video sequence $\boldsymbol{V}'$ and masks $[\boldsymbol{M}^{(n)}]_{n=2}^{N}$

Initialize first frame $\boldsymbol{X}^{(1)} \sim \mathcal{N}$,
noisy background $\boldsymbol{B} \sim \mathcal{N}$, and accumulated flow $\tilde{\boldsymbol{F}}^{(1)} \leftarrow \mathbf{0}$
for $n = 2$ to N **do**
  $\tilde{\boldsymbol{F}}^{(n)} \leftarrow \text{Warp}(\tilde{\boldsymbol{F}}^{(n-1)}, \boldsymbol{F}^{(n)}) + \boldsymbol{F}^{(n)}$
  Initialize $\boldsymbol{X}'^{(n)} \leftarrow \mathbf{0}$ and mask $\boldsymbol{M}^{(n)} \leftarrow \mathbf{0}$
  for $(u, v) \in \mathcal{S} = \{0, \cdots, H\} \times \{0, \cdots, W\}$ **do**
    $(u', v') \leftarrow \text{round}\left((u, v) + \tilde{\boldsymbol{F}}^{(n)}(u, v)\right)$
    # Correspondence check
    if $(u', v') \in \mathcal{S}$ **then**
      $\boldsymbol{X}'^{(n)}(u', v') \leftarrow \boldsymbol{X}^{(1)}(u, v)$
      $\boldsymbol{M}^{(n)}(u', v') \leftarrow 1$
  for $(u, v) \in \mathcal{S}$ **do**
    # If no correspondence, then paste background pixels
    if $\boldsymbol{M}^{(n)}(u, v) = 0$ **then**
      $\boldsymbol{X}'^{(n)}(u, v) \leftarrow \boldsymbol{B}(u, v)$
return $\boldsymbol{V}' = [\boldsymbol{X}'^{(n)}]_{n=1}^{N}$ and $[\boldsymbol{M}^{(n)}]_{n=2}^{N}$

Algorithm 2 `MotionOptim`

Input: $\boldsymbol{V}' = [\boldsymbol{X}'^{(n)}]_{n=1}^{N}$, $[\boldsymbol{M}^{(n)}]_{n=2}^{N}$ from Alg. 1, λ, η
Output: Video sequence $\boldsymbol{V}$

for $n = 2$ to N **do**
  Initialize $\boldsymbol{X}^{(n)} \leftarrow \boldsymbol{X}'^{(n)}$
  for $t = 1$ to T **do**
    # See $\mathcal{S}$ from Alg. 1
    Sample K pixels as $\mathcal{S}' \subset \mathcal{S}$
    Fit distribution $\mathcal{N}_{\mathcal{S}'}$ from $\{\boldsymbol{X}^{(n)}(u, v)\}_{(u,v)\in\mathcal{S}'}$
    # Appearance-free loss
    Compute $\mathcal{L}_{\text{appear}}(\boldsymbol{X}^{(n)})$ from Eq. 2
    # Motion fidelity loss
    Compute $\mathcal{L}_{\text{motion}}(\boldsymbol{X}^{(n)})$ from Eq. 3
    $\mathcal{L}(\boldsymbol{X}^{(n)}) \leftarrow (1-\lambda)\mathcal{L}_{\text{appear}}(\boldsymbol{X}^{(n)})$
      $+\lambda\mathcal{L}_{\text{motion}}(\boldsymbol{X}^{(n)})$
    # Gradient update
    $\boldsymbol{X}^{(n)} \leftarrow \boldsymbol{X}^{(n)} - \eta \cdot \nabla_{\boldsymbol{X}^{(n)}} \mathcal{L}(\boldsymbol{X}^{(n)})$
return $\boldsymbol{V}$

match $\mathcal{N}$ exactly, which tends to eliminate the motion-defined forms. Instead, for each frame, we randomly sample a subset $\mathcal{S}' \subset \mathcal{S}$ of K pixels, fit a distribution to the sampled pixels, and penalize their deviation from the target noise distribution $\mathcal{N}$ as follows:

$$\mathcal{L}_{\text{appear}}(\boldsymbol{X}^{(n)}) = \mathbb{E}_{\mathcal{S}'\sim\mathcal{S}}\left[\text{KL}\left(\mathcal{N}_{\mathcal{S}'} \| \mathcal{N}\right)\right], \tag{2}$$

where $\mathcal{N}_{\mathcal{S}'}$ is the empirical distribution fitted to pixels in $\mathcal{S}'$. The locality ensures that small patches maintain noise-like statistics, while allowing the overall frame to encode motion-defined forms.

Motion fidelity loss enforces motion consistency through reconstruction. As in step one, we use $\boldsymbol{M}^{(n)}$ to keep track of which pixels have correspondences. Intuitively, pixels with correspondence, *i.e.*, $\boldsymbol{M}^{(n)} = 1$, should faithfully match the warped frame from motion $\boldsymbol{X}'^{(n)}$; otherwise ($\boldsymbol{M}^{(n)} = 0$), they should match the background noise $\boldsymbol{B}$. We capture such intuition in the following loss:

$$\mathcal{L}_{\text{motion}}(\boldsymbol{X}^{(n)}) = \text{mean}\left[\boldsymbol{M}^{(n)} \odot \Delta(\boldsymbol{X}^{(n)}, \boldsymbol{X}'^{(n)}) + (1 - \boldsymbol{M}^{(n)}) \odot \Delta(\boldsymbol{X}^{(n)}, \boldsymbol{B})\right], \tag{3}$$

where $\Delta(\cdot, \cdot)$ is the ℓ_2-loss, $\odot$ is elementwise multiplication, and $\text{mean}[\cdot]$ is the pixel average.

3.3 Fine-grained Question-Answer Generation

We consider the benchmark in the form of visual question answering (VQA). We design multi-level question-answer pairs that probe different aspects of motion reasoning. Our three-level difficulty hierarchy follows a natural progression: *whether* motion exists (coarse), *what* is moving (medium), and *how* it moves (fine). The **coarse** level asks *whether* motion is present; since appearance-free videos offer no static cues, models must rely entirely on inter-frame changes. The **medium** level asks *what* is moving, requiring models to identify the moving entity(ies) without appearance cues for boundary or texture. The **fine** level asks *how* objects move, probing detailed attributes such as bounce counts, rotation/driving direction, or action type. Full question templates are in Sec. A2.5.

4 Experiments

We evaluate VLMs on appearance-free motion reasoning to answer the following questions: (a) How far behind are VLMs compared to humans (Sec. 4.1)? (b) Is the bottleneck in motion perception or motion reasoning (Sec. 4.2)? (c) Does a higher frame rate help VLMs perceive motion (Sec. A1.3)? (d) How much appearance information do VLMs need (Sec. 4.3)? Our AppearFreeVQA has approximately 5,100 videos and 17,600 questions across four categories: **Symbol** (1,000 videos / 5,334 questions), **Humanoid** (2,867 / 8,601), **Animal** (724 / 2,147), and **Scene** (500 / 1,500). Additionally, we provide a **Noise** control (100 videos, coarse-only), which contains pure Gaussian noise with no motion at all, testing whether models default to predicting motion regardless of input. Full per-category and per-type breakdowns are provided in the Appendix.

Table 1: **Multiple-choice accuracy (%) on AppearFreeVQA** with Noise control and under three difficulty levels: **C**oarse, **M**edium, **F**ine. Values in red fall below that column's Random-row baseline.

Agent	Noise	All	Symbol				Humanoid				Animal				Scene			
			Avg	C	M	F	Avg	C	M	F	Avg	C	M	F	Avg	C	M	F
Baselines																		
Random	50.0	32.4	30.8	50.0	26.7	26.3	33.3	50.0	25.0	25.0	33.4	50.0	25.0	25.0	33.3	50.0	25.0	25.0
Human	**100.0**	**93.3**	**95.7**	**100.0**	**93.3**	**95.4**	**95.7**	**100.0**	**98.9**	**88.3**	**78.2**	**100.0**	**60.0**	**74.1**	**94.6**	**100.0**	96.7	**87.1**
Proprietary Models																		
GPT 5.1	3.0	41.7	31.8	72.8	29.2	16.9	48.3	**100.0**	12.8	32.0	45.2	**100.0**	7.0	27.9	45.3	**100.0**	4.0	32.0
GPT-5.6 Sol	99.0	30.2	26.8	19.9	26.2	30.3	27.3	33.2	16.8	32.1	45.1	85.5	14.5	35.0	42.7	90.0	10.0	28.0
Claude Sonnet 5	**100.0**	20.0	20.9	9.5	26.0	21.3	17.7	0.0	21.3	31.9	26.7	0.0	17.9	63.6	21.3	0.0	40.0	24.0
Gemini 3 Flash Preview	**100.0**	20.7	31.2	10.5	48.0	25.7	13.8	0.3	4.4	36.6	15.8	11.7	4.8	31.4	38.0	86.0	2.0	26.0
Open-Source Models																		
VideoLLaMA 3	**100.0**	23.8	21.0	8.8	22.6	24.9	29.0	42.1	13.6	31.4	16.0	2.2	13.7	32.8	14.9	4.2	15.4	25.2
InternVL 3.5 8B	**100.0**	16.3	19.3	8.8	25.1	18.9	15.7	0.0	11.9	35.2	14.9	0.0	1.9	43.6	11.2	0.0	8.2	25.4
Cosmos-Reason2-2B	**100.0**	15.6	17.7	8.8	25.3	15.0	14.9	0.0	21.1	23.6	14.1	0.0	12.8	30.0	14.7	0.0	19.2	25.0
Qwen3-VL-8B-Instruct	**100.0**	17.5	14.5	8.8	21.8	10.7	20.8	0.0	29.4	33.1	12.4	0.0	13.7	23.9	16.0	0.0	22.0	26.0
Qwen3.5-9B	**100.0**	14.3	18.1	9.5	26.4	14.5	13.1	2.5	8.3	28.3	12.4	0.0	3.9	34.0	10.9	1.4	6.4	25.0
NVILA-Lite-8B	85.0	20.2	26.9	8.8	36.1	26.7	16.1	0.0	22.3	26.2	7.1	0.0	1.4	20.3	37.9	0.0	88.4	25.4
DeepSeek-VL2	**100.0**	28.4	24.6	8.8	21.6	33.9	28.4	0.0	53.9	31.1	28.4	0.0	34.5	51.5	41.7	0.0	**100.0**	25.0
Kimi-VL-A3B-Instruct	99.0	17.4	23.1	8.8	27.4	25.6	11.1	0.0	1.4	32.0	11.3	0.0	4.1	30.5	41.7	0.0	**100.0**	25.0

4.1 Human-VLM Gap

Experimental setup. We evaluate proprietary models, *e.g.*, GPT (OpenAI, 2025a; 2026), o4-mini (OpenAI, 2025b), Claude (Anthropic, 2026;?), Gemini (Google, 2025a;b), and open-source models, *e.g.*, DeepSeek (Wu et al., 2024), VideoLLaMA 3 (Zhang et al., 2025), InternVL (Wang et al., 2025), Cosmos-Reason2 (NVIDIA, 2025), QwenVL (Wang et al., 2024; Bai et al., 2025b;a; Qwen Team, 2026), NVILA (Liu et al., 2025), and Kimi-VL (Kimi Team, 2025). All models are evaluated in a zero-shot setting with a multiple-choice question format. The human baseline is established by surveying 12 human annotators. Humans answer each question quickly and effortlessly: mean response time is 5.2s ($\sigma = 3.5$s, median 3.8s). More results are in Appendix Sec. A1.

Results. Tab. 1 reports accuracy on AppearFreeVQA. Human performance is consistently high (93.3% overall), achieving 100% at the coarse level, confirming that persistence of vision alone suffices for basic motion understanding. Performance declines with difficulty: coarse-level (*Whether*) is easiest and fine-level (*How*) is hardest. In contrast, VLMs struggle on AppearFreeVQA. Only GPT-5.1 surpasses random chance (32.4%) among the models shown here; most models fall below random chance. Notably, newer models do not necessarily improve performance. Qwen3-VL-8B (Bai et al., 2025a) scores 17.5%, while the newer Qwen3.5-9B (Qwen Team, 2026) drops to 14.3%.

Models also exhibit opposite failure modes at the coarse level: GPT-5.1 (OpenAI, 2025a) is over-confident, predicting motion in every appearance-free video, resulting in 100% on Humanoid coarse-level questions, while InternVL3.5-8B (Wang et al., 2025) and NVILA-Lite-8B (Liu et al., 2025) are under-confident, scoring 0% on those same categories by consistently predicting no motion. The Noise column in Tab. 1 provides more clues to this phenomenon: GPT-5.1 scores only 3.0% on Noise, answering "yes, there is motion" even for pure static noise, so its high coarse accuracy reflects a fixed "yes" bias that happens to align with the ground truth rather than genuine detection; several other models instead score near 100% on Noise, by defaulting to "no motion" regardless of input. A related failure mode appears on Scene medium, where several models are biased to predict the moving object as the car for most questions, yet fail to answer which direction the car is moving at the fine level.

Takeaway 1: VLMs do not really understand appearance-free videos. Even frontier models fail to surpass random chance on most categories, far behind human performance.

4.2 Optical Flow as Oracle Control

Experimental setup. The low VLM performance raises a natural question: *Is the bottleneck in motion perception or motion reasoning?* We replace the appearance-free video $\boldsymbol{V}$ with an optical flow visualization video as an oracle (see Appendix for details) and evaluate Qwen3.5-9B (Qwen Team, 2026) on six such conditions. The first two directly test this question: (a) **Estimated optical flow $\hat{\boldsymbol{F}}$**: we use an off-the-shelf flow estimator, *i.e.*, FlowSeek (Poggi & Tosi, 2025), to predict $\hat{\boldsymbol{F}}$ from appearance-free videos $\boldsymbol{V}$. (b) **Target optical flow $\boldsymbol{F}$**: for Symbol and Humanoid, we acquire $\boldsymbol{F}$ from the known warping parameters in our video generation process; for Animal, $\boldsymbol{F}$ is instead

Table 2: **Oracle controls on AppearFreeVQA. Property** marks whether each control preserves **T**emporal alignment, **Mo**tion/source alignment, **S**patial alignment, and **C**olor convention. We report accuracy (%) of Qwen3.5-9B (Qwen Team, 2026) following the format in Tab. 1.

Condition	Property				All	Symbol				Humanoid				Animal				Scene			
	T	Mo	S	C		Avg	C	M	F	Avg	C	M	F	Avg	C	M	F	Avg	C	M	F
Random	–	–	–	–	32.4	30.8	50.0	26.7	26.3	33.3	50.0	25.0	25.0	33.4	50.0	25.0	25.0	33.3	50.0	25.0	25.0
Human	–	–	–	–	**93.3**	**95.7**	**100.0**	**93.3**	**95.4**	**95.7**	**100.0**	**98.9**	**88.3**	**78.2**	**100.0**	**60.0**	**74.1**	**94.6**	**100.0**	96.7	**87.1**
Appearance-free	–	–	–	–	14.3	18.0	9.2	26.3	14.6	13.1	2.5	8.3	28.3	12.4	0.0	3.9	34.0	10.9	1.4	6.4	25.0
Estimated $\hat{\boldsymbol{F}}$	✓	✓	✓	✓	49.4	45.7	90.6	38.5	32.6	50.4	**100.0**	8.9	42.3	44.4	93.4	3.0	36.5	63.6	66.6	99.2	25.0
Target $\boldsymbol{F}$	✓	✓	✓	✓	66.8	74.6	93.4	88.4	54.6	63.0	98.7	43.6	46.6	62.6	95.0	33.7	58.8	67.2	76.8	99.8	25.0
Temporally misaligned	✗	✓	✓	✓	65.4	73.6	92.6	90.2	51.2	63.2	99.6	43.6	46.6	61.5	92.3	33.7	58.5	54.2	37.8	99.8	25.0
Motion misaligned	✓	✗	✓	✓	53.7	43.5	78.6	46.1	26.3	58.7	98.8	44.0	33.3	53.0	95.2	19.9	43.6	62.7	63.2	**100.0**	25.0
Spatially misaligned	✓	✓	✗	✓	20.0	22.3	14.6	27.1	21.4	16.0	11.9	3.5	32.8	16.4	10.9	3.6	35.3	40.3	14.0	82.0	25.0
Unconventional color	✓	✓	✓	✗	67.1	75.2	91.9	89.4	55.8	63.6	98.9	45.4	46.4	63.3	95.4	33.8	60.4	64.5	69.6	99.0	25.0

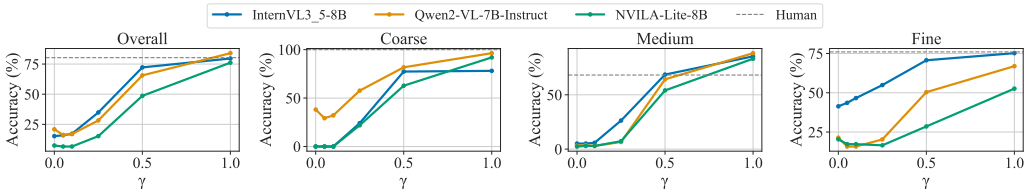


Figure 4: **Effect of appearance level** γ on the Animal category.

FlowSeek prediction on the raw video; for Scene, $\boldsymbol{F}$ is provided from VKITTI2. However, $\boldsymbol{F}$'s gain could instead come from the VLM's pretraining exposure to optical-flow-style visualizations, *i.e.*, recognizing a familiar flow pattern, rather than genuinely reasoning about motion-defined form. To rule this out, we conduct four leave-one-out controls that each remove exactly one property of $\boldsymbol{F}$ while holding the other three fixed: (c) **Temporally misaligned flow** removes temporal alignment: frame order is scrambled. (d) **Motion misaligned flow** removes source alignment: $\boldsymbol{F}$ is replaced with the target flow of a different, randomly sampled video. (e) **Spatially misaligned flow** removes spatial alignment: pixel locations of $\boldsymbol{F}$ are randomly permuted. (f) **Unconventional color** removes the conventional Middlebury color encoding, replacing it with an unfamiliar hue-rotated one.

Results. Tab. 2 reports all conditions. When providing the estimated optical flow $\hat{\boldsymbol{F}}$, Qwen3.5-9B yields a moderate gain, from 14.3% to 49.4%. The gain is modest because appearance-free videos are out-of-distribution for off-the-shelf flow estimators (Poggi & Tosi, 2025), and $\hat{\boldsymbol{F}}$ is too noisy to serve as a true oracle. Target flow $\boldsymbol{F}$ leads to a much larger gain: Qwen3.5-9B jumps from 14.3% to 66.8%. On Animal, where $\boldsymbol{F}$ is a FlowSeek prediction from the real RGB video rather than literal ground truth, this gap (44.4% to 62.6%) is smaller than the improvement in other categories. Among the leave-one-out controls, removing spatial arrangement drops performance back close to the appearance-free baseline (20.0% vs. 14.3%), while removing source alignment (53.7%) or the conventional color encoding (67.1%) leaves performance much closer to full target-flow level. Interestingly, removing temporal progression via frame shuffling (Symbol 73.6%, Animal 61.5%) barely hurts at all, *i.e.*, the target-flow gain depends mainly on spatially faithful motion structure.

Takeaway 2: Motion extraction is a major bottleneck, and the oracle's gain requires genuine motion structure.

4.3 Analysis on Appearance Level

Experimental setup. We further investigate the effect of appearance cues on motion perception: *How much appearance information do VLMs need to perceive motion?* We introduce $\gamma \geq 0$ to control appearance level, producing *semi-appearance-free* videos that fall between fully appearance-free ($\gamma = 0$) and natural RGB video ($\gamma \to \infty$). Let $\boldsymbol{X}^{(n)}$ denote the noise-warped frame from Sec. 3.2 and $\boldsymbol{I}^{(n)}$ the corresponding original RGB frame. The semi-appearance-free frame $\hat{\boldsymbol{X}}^{(n)}$ is obtained by optimizing a least-squares problem that pulls $\hat{\boldsymbol{X}}^{(n)}$ toward the original RGB frame $\boldsymbol{I}^{(n)}$ with strength γ (see Eq. A6 in the Appendix).

Results. Fig. 4 reports accuracy across γ values on the Animal category and Fig. 5 shows example semi-appearance-free frames. Performance mostly rises monotonically with γ, confirming that VLMs



(a) $\gamma = 0.1$ (b) $\gamma = 0.2$ (c) $\gamma = 1.0$ (d) $\gamma \to \infty$

Figure 5: **Appearance Level γ.** Visual examples from low appearance-level (γ=0.1) to natural RGB ($\gamma \to \infty$).

Table 3: **Accuracy on MoCA eval. split.** **M** = MoCA-train; **S** = SpookyBench; **A** = AppearFreeVQA.

Training data	**MoCA Avg** (%)
Zero-shot	$76.6_{\pm 1.3}$
M	$80.7_{\pm 2.6}$
M+S	$80.7_{\pm 2.3}$
M+A (Ours)	$\mathbf{84.3}_{\pm 1.7}$

rely on appearance cues to perceive motion. Specifically, $\gamma = 0.1$ (Fig. 5(a)) provides a small boost to InternVL3.5-8B, improving accuracy from 14.2% to 17.2%; $\gamma = 0.2$ (Fig. 5(b)) further improves it to 34.9%; $\gamma = 1.0$ (Fig. 5(c)) finally yields a dramatic boost to 79.6%, approaching the 78.2% human performance. Overall, for coarse-level, medium-level, and fine-level questions, the expected γ values for VLMs to achieve human-level performance are approximately $\gamma = 1.0$, $\gamma = 0.5$, and $\gamma = 1.0$, respectively, though the threshold varies across individual models.

Takeaway 3: VLMs are exploiting appearance cues. VLM accuracy rises with γ and approaches near-human performance.

4.4 Application

A camouflaged animal is challenging to detect because its appearance blends seamlessly with the background (Fan et al., 2020; Gao et al., 2025). When appearance cues are misleading, motion becomes the primary signal (Lamdouar et al., 2020; Feng et al., 2025). This mirrors the challenge in our benchmark of perceiving motion under minimal appearance. Hence, we ask: *Do motion perception skills learned on appearance-free videos transfer to detecting camouflaged animals?*

Experimental setup. We evaluate on MoCA (Moving Camouflaged Animals) (Lamdouar et al., 2020), a video dataset of naturally camouflaged animals. We cast each video as a 5-way multiple-choice question ("What animal is in the video?"). We split the dataset into a training split, denoted as "MoCA-train", and an evaluation split. We compare the performance of NVILA-Lite-2B (Liu et al., 2025) under zero-shot and different fine-tuning settings: (1) fine-tuning on MoCA-train, (2) fine-tuning on MoCA-train and SpookyBench (Upadhyay et al., 2026), and (3) fine-tuning on MoCA-train and AppearFreeVQA. Each is repeated 5 times. See Appendix (Sec. A1.4) for details.

Results. Tab. 3 reports accuracy on MoCA. Fine-tuning on MoCA-train alone (M) achieves an accuracy of $80.7 \pm 2.6\%$. Adding SpookyBench (M+S) yields the same $80.7 \pm 2.3\%$, suggesting that SpookyBench offers minimal benefit to camouflage detection. In contrast, adding AppearFreeVQA (M+A) improves accuracy to $84.3 \pm 1.7\%$. This is initial evidence that appearance-free motion reasoning transfers to camouflage detection in the wild. Note that we do not claim this result establishes AppearFreeVQA as the unique or exclusive source of the gain for camouflage detection.

Limitation: The gain in camouflage detection comes with a trade-off. On MVBench (Li et al., 2024), a general video-understanding benchmark, accuracy drops from 50.5% (zero-shot) to 42.5% (M) to 38.0% (M+A), indicating that fine-tuning for this specialized task may degrade general video understanding. Studying prior preservation techniques is an interesting direction for future work.

5 Conclusion

We present AppearFreeVQA, a VQA benchmark designed to isolate motion perception from appearance reasoning. The proposed generation pipeline creates appearance-free videos where each frame appears as noise, yet the motion contains information that humans can easily perceive. Experiments show a human-VLM performance gap on this benchmark. Oracle experiments identify motion extraction, rather than reasoning, as a major source of this gap. Despite recent progress, VLMs may still fall short of human vision in motion perception. Finally, we demonstrate that fine-tuning on AppearFreeVQA improves a model's performance on a real-world application of camouflaged animal detection, where appearance cues are limited.

Reproducibility Statement

We will release the full AppearFreeVQA benchmark and code, *i.e.*, the appearance-free video generation pipeline, the per-category question generators, and the evaluation harness, upon acceptance. To support review in the meantime, we attach a small subset of AppearFreeVQA together with the data-generation code for the Symbol category as supplementary material (`code/` and `data/` in the attached `supp.zip`), sufficient to reproduce the `ForwardWarp` and `MotionOptim` pipeline end-to-end on a representative category.

AI Use Statement

In this work, we used generative AI coding assistants (Claude Code) to assist with software engineering on the code repository, with LaTeX formatting of tables, and with minor editing and polishing of paper prose (wording and structure suggestions on text the authors wrote). AI tools were not used to develop the method or theoretical framing, design experiments, generate synthetic data, or interpret results/formulate claims. All research contributions and conclusions are the authors' own. Ultimately, the authors take full responsibility for all content in this paper, including any AI-assisted text or code.

References

Edward H Adelson and James R Bergen. Spatiotemporal energy models for the perception of motion. *JOSA A*, 1985. 3

Anthropic. Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, 2026. 7

Anthropic. Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, 2026. 7, 1

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 2, 3, 7

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. 2, 3, 7, 1

Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011. 17

Randolph Blake and Sang-Hun Lee. The role of temporal structure in human vision. *Behavioral and cognitive neuroscience reviews*, 2005. 2

Oliver Braddick. The masking of apparent motion in random-dot patterns. *Vis. Res.*, 1973. 1, 2

Kenneth H Britten, Michael N Shadlen, William T Newsome, and J Anthony Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 1992. 2

David Burr. Motion perception: Human psychophysics. In *The New Visual Neurosciences*, chapter 54. MIT Press, 2014. 1, 2

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. 5, 13, 17

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *Proc. ICLR*, 2026. 11

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2, 3

L. Del Pero, S. Ricco, R. Sukthankar, and V. Ferrari. Articulated motion discovery using pairs of trajectories. In *Proc. CVPR*, 2015. 5, 11

L. Del Pero, S. Ricco, R. Sukthankar, and V. Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *IJCV*, 2016. 5, 11

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proc. CVPR*, 2020. 9

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. CVPR*, 2016. 3

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. ICCV*, 2019. 3

Yu'ang Feng, Shuyong Gao, Fuzhen Yan, Yicheng Song, Lingyi Hong, Junjie Hu, and Wenqiang Zhang. Scoring, remember, and reference: Catching camouflaged objects in videos. In *Proc. ICCV*, 2025. 9

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proc. CVPR*, 2025. 2, 3

Shuyong Gao, Yu'ang Feng, Qishan Wang, Lingyi Hong, Xinyu Zhou, Liu Fei, Yan Wang, and Wenqiang Zhang. MSVCOD: a large-scale multi-scene dataset for video camouflage object detection. *arXiv preprint arXiv:2502.13859*, 2025. 9

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 2013. 5

James J Gibson. The visual perception of objective motion and subjective movement. *Psychological review*, 1954. 1

James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 1, 3

Google. Gemini 2.5 pro. https://ai.google.dev/, 2025a. Accessed via Gemini API / Google AI Studio, 2026. 7, 1

Google. Gemini 3 flash preview. https://ai.google.dev/, 2025b. Accessed via Gemini API, 2026. 2, 3, 7

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proc. CVPR*, 2025. 2, 3

Wenjin Hou, Wei Liu, Han Hu, Xiaoxiao Sun, Serena Yeung-Levy, and Hehe Fan. Seeing is believing? a benchmark for multimodal large language models on visual illusions and anomalies. *arXiv preprint arXiv:2602.01816*, 2026. 3

Ian P Howard and Brian J Rogers. *Perceiving in depth, volume 2: Stereoscopic vision*. Oxford University Press, 2012. 3

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, et al. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world. In *Proc. CVPR*, 2026. 3

Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 1973. 1, 2

Bela Julesz. *Foundations of cyclopean perception.* U. Chicago Press, 1971. 2, 3

Akila Kadambi, Marco Iacoboni, Lisa Aziz-Zadeh, and Srini Narayanan. Evaluating point-light biological motion in multimodal large language models. *arXiv preprint arXiv:2509.23517*, 2025. 3

Kimi Team. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025. 7, 2

Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proc. ACCV*, 2020. 2, 9, 3

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proc. CVPR*, 2024. 2, 3, 9

Zijian Ling, Han Zhang, Yazhuo Zhou, and Jiahao Cui. ColorBlindnessEval: Can vision-language models pass color blindness tests? *arXiv preprint arXiv:2509.19070*, 2025. 3

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Proc. ACL Findings*, 2024. 2, 3

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. NVILA: Efficient frontier visual language models. In *Proc. CVPR*, 2025. 7, 9, 1, 2, 3

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2, 3

Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *PR*, 2022. 4, 9

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. VideoZeroBench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*, 2026. 3

Albert Michotte. *The perception of causality*. Routledge, 2017. 2

Ken Nakayama. Biological image motion processing: a review. *Vis. Res.*, 1985. 3

NVIDIA. Cosmos-Reason2: Physical AI reasoning vision-language model. https://huggingface.co/nvidia/Cosmos-Reason2-2B, 2025. No dedicated technical report as of this writing; builds on Cosmos-Reason1 (arXiv:2503.15558). 7, 2

OpenAI. Gpt-5.1. https://platform.openai.com/, 2025a. Large language model. 1, 2, 3, 7

OpenAI. o4-mini. https://platform.openai.com/, 2025b. Accessed via OpenAI API, 2026. 7, 1

OpenAI. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6, 2026. 7

Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999. 1, 2

J Timothy Petersik. Recovery of structure from motion: Implications for a performance theory based on the structure-from-motion theorem. *Perception & Psychophysics*, 1987. 2

Matteo Poggi and Fabio Tosi. FlowSeek: optical flow made easier with depth foundation models and motion bases. In *Proc. ICCV*, 2025. 7, 8, 11, 17

Frank E Pollick, Jim W Kay, Katrin Heim, and Rebecca Stringer. Gender recognition from point-light walkers. *Journal of Experimental Psychology: Human Perception and Performance*, 2005. 1, 2

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. No dedicated non-Omni technical report as of this writing. 7, 8, 5

D Regan. Orientation discrimination for objects defined by relative motion and objects defined by luminance contrast. *Vis. Res.*, 1989. 2, 3

D Regan, D Giaschi, JA Sharpe, and XH Hong. Visual processing of motion-defined form: selective failure in patients with parietotemporal lesions. *Journal of Neuroscience*, 1992. 2, 3

Mark O Scase, Oliver J Braddick, and Jane E Raymond. What is noise for the motion system? *Vis. Res.*, 1996. 2

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. NeurIPS*, 2014. 3

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proc. ICCV*, 2019. 2

Matthias Tangemann, Matthias Kümmerer, and Matthias Bethge. Object segmentation from common fate: Motion energy processing enables human-like zero-shot generalization to random dot stimuli. In *Proc. NeurIPS*, 2024. 3, 5, 6

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. 3

Shimon Ullman. *The interpretation of visual motion.* Massachusetts Inst of Technology Pr, 1979. 1, 3

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? In *Proc. CVPR*, 2026. 3, 9

AJ Van Doorn and JJ Koenderink. Temporal properties of the visual detectability of moving spatial white noise. *Experimental Brain Research*, 1982. 1, 2

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7, 1, 2, 3

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 3, 7, 1

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 3

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 7

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. CVPR*, 2016. 3

Yung-Hao Yang, Zitang Sun, Taiki Fukiage, and Shin'ya Nishida. Huperflow: A comprehensive benchmark for human vs. machine motion estimation comparison. In *Proc. CVPR*, 2025. 3

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. CREMA: Generalizable and efficient video-language reasoning via multimodal modular fusion. In *Proc. ICLR*, 2025. 6

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 7, 2

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *TPAMI*, 2024. 2

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards spatiotemporal awareness in vision language models. In *Proc. ICCV*, 2025. 3

Interactive Demo. We provide an interactive HTML demo of AppearFreeVQA as part of the supplementary material. Reviewers are warmly invited to try the benchmark themselves. Please unzip the attached `supp.zip` and open `index.html` in any modern browser (no server or internet access needed) to play 80 real AppearFreeVQA entries (20 each of Symbol/Humanoid/Animal/Scene) and check yourself against each question's coarse/medium/fine answer options.

The appendix is organized as follows:

- In Sec. A1, we provide additional experimental results and analysis.
- In Sec. A2, we provide the detailed specifications of AppearFreeVQA, including all shapes, motion types, directions, question templates, and answer options for each category.
- In Sec. A3, we provide additional details and discussion regarding our video generation algorithm.

A1 Additional Results

A1.1 Additional Results for Human-VLM Gap

Tab. A1 extends the human-VLM gap analysis in Tab. 1 to additional proprietary models (o4-mini (OpenAI, 2025b), Claude Opus 4.7 (Anthropic, 2026), Gemini 2.5 Pro (Google, 2025a)) and open-source model scales (Qwen2-VL-7B (Wang et al., 2024), Qwen2.5-VL-7B/3B (Bai et al., 2025b), NVILA-15B/NVILA-Lite-15B (Liu et al., 2025), InternVL3.5-1B (Wang et al., 2025)), broadening the evidence for the human-VLM gap across model families and parameter counts. Claude Opus 4.7 (34.0%) and Gemini 2.5 Pro (38.6%) exceed random chance (32.4%), joining GPT-5.1 (Tab. 1) as the only models across both tables to do so; every open-source model here falls below random (12.8%–31.7%), and humans maintain a substantial advantage at 93.3%. Larger models do not consistently outperform smaller ones: Qwen2.5-VL-7B (12.8%) scores below the smaller Qwen2.5-VL-3B (19.0%), and the older Qwen2-VL-7B (16.6%) outperforms the newer Qwen2.5-VL-7B despite being the previous generation. This confirms that the failure on AppearFreeVQA is not primarily a capacity or recency bottleneck.

Table A1: **Extended results on AppearFreeVQA.** Multiple-choice accuracy (%) across four motion categories and three difficulty levels: **C**oarse, **M**edium, **F**ine, extending Tab. 1 with additional proprietary and open-source models. Values in red fall below random performance.

Agent	Noise	All	Symbol				Humanoid				Animal				Scene			
			Avg	C	M	F	Avg	C	M	F	Avg	C	M	F	Avg	C	M	F
Baselines																		
Random	50.0	32.4	30.8	50.0	26.7	26.3	33.3	50.0	25.0	25.0	33.4	50.0	25.0	25.0	33.3	50.0	25.0	25.0
Human	**100.0**	**93.3**	**95.7**	**100.0**	**93.3**	**95.4**	**95.7**	**100.0**	**98.9**	**88.3**	**78.2**	**100.0**	**60.0**	**74.1**	**94.3**	**100.0**	96.7	**87.1**
Proprietary Models																		
o4-mini	98.0	19.5	21.3	16.8	23.2	21.6	18.5	8.2	19.9	27.4	18.1	2.8	11.3	41.2	14.0	0.0	18.0	24.0
Claude Opus 4.7	26.0	34.0	28.6	27.2	27.2	30.2	36.3	36.3	36.3	36.3	41.9	88.7	11.3	25.0	26.0	52.0	2.0	24.0
Gemini 2.5 Pro	82.0	38.6	30.2	52.8	24.8	25.2	44.5	57.3	45.2	31.0	40.0	73.2	22.5	23.5	48.0	**100.0**	22.0	22.0
Open-Source Models																		
Qwen2-VL-7B-Instruct	**100.0**	16.6	22.6	12.5	28.6	21.8	11.3	2.2	12.5	19.1	8.1	0.6	2.6	21.6	38.3	2.4	87.4	25.0
Qwen2.5-VL-7B-Instruct	**100.0**	12.8	11.5	0.0	15.9	12.6	12.0	0.0	10.9	25.0	12.1	0.0	17.3	19.3	22.5	0.0	42.4	25.0
Qwen2.5-VL-3B	**100.0**	19.0	25.7	8.8	33.5	26.3	14.8	0.0	11.9	32.4	12.9	0.0	7.6	31.6	28.1	0.0	59.2	25.0
InternVL3.5-1B	**100.0**	14.0	17.5	8.8	30.0	10.5	11.2	0.0	10.2	23.3	8.6	0.0	0.7	25.8	25.7	0.0	52.0	25.0
NVILA-15B	**100.0**	22.5	28.4	8.8	39.4	27.4	19.0	0.0	32.4	24.8	8.2	1.5	5.4	18.0	41.1	0.0	98.2	25.0
NVILA-Lite-15B	**100.0**	31.7	37.3	10.6	43.4	43.3	28.6	3.1	57.6	25.2	23.4	49.0	0.8	20.2	41.7	0.0	**99.8**	25.2

A1.2 Oracle Results Across Additional Models

Tab. A2 extends the model coverage of Tab. 2's three matching conditions, appearance-free video ($\boldsymbol{V}$), estimated optical flow ($\hat{\boldsymbol{F}}$), and ground-truth optical flow ($\boldsymbol{F}$), to additional VLM backbones beyond Qwen3.5-9B. (Sec. A1.6 instead extends the same baseline to further alternative input *encodings* on a single model.) Providing ground-truth flow raises accuracy from below random to well above it for every model tested. For example, Qwen2-VL-7B-Instruct improves from 16.6% ($\boldsymbol{V}$) to 62.2% ($\boldsymbol{F}$); InternVL3.5-8B from 16.3% to 65.4%.

Table A2: **Oracle results across additional VLMs.** Multiple-choice accuracy (%) when each model is given appearance-free video ($\boldsymbol{V}$), estimated optical flow ($\hat{\boldsymbol{F}}$), or ground-truth optical flow ($\boldsymbol{F}$), extending the matching three conditions in Tab. 2 to additional model backbones.

Agent	All	Symbol				Humanoid				Animal				Scene			
		Avg	C	M	F	Avg	C	M	F	Avg	C	M	F	Avg	C	M	F
Random	32.4	30.8	50.0	26.7	26.3	33.3	50.0	25.0	25.0	33.4	50.0	25.0	25.0	33.3	50.0	25.0	25.0
Human	**93.3**	**95.7**	**100.0**	**93.3**	**95.4**	**95.7**	**100.0**	**98.9**	**88.3**	**78.2**	**100.0**	**60.0**	**74.1**	**94.3**	**100.0**	96.7	**87.1**
NVILA-Lite-8B (Liu et al., 2025)																	
With $\boldsymbol{V}$	24.9	32.0	25.3	40.6	27.6	22.5	29.2	9.9	28.5	7.5	0.0	2.4	20.4	37.9	0.0	88.4	25.4
With $\hat{\boldsymbol{F}}$	49.0	46.2	72.8	48.4	33.2	49.5	**100.0**	19.6	28.8	38.6	**100.0**	2.8	11.8	70.7	97.2	91.8	23.2
With $\boldsymbol{F}$	61.2	58.8	62.0	77.9	41.6	61.6	**100.0**	47.0	37.7	56.8	**100.0**	29.9	39.8	72.6	93.0	94.2	30.6
NVILA-15B (Liu et al., 2025)																	
With $\boldsymbol{V}$	22.5	28.4	8.8	39.4	27.4	19.0	0.0	32.4	24.8	8.2	1.5	5.4	18.0	41.1	0.0	98.2	25.0
With $\hat{\boldsymbol{F}}$	35.5	41.3	47.7	45.2	35.6	34.6	48.8	27.6	27.3	20.6	38.6	4.6	18.6	41.1	0.0	98.2	25.0
With $\boldsymbol{F}$	54.7	55.8	50.6	73.5	43.3	59.2	77.0	60.0	40.6	42.7	62.5	26.3	39.0	42.1	5.6	95.8	25.0
NVILA-Lite-15B (Liu et al., 2025)																	
With $\boldsymbol{V}$	31.7	37.3	10.6	43.4	43.3	28.6	3.1	57.6	25.2	23.4	49.0	0.8	20.2	41.7	0.0	99.8	25.2
With $\hat{\boldsymbol{F}}$	39.4	43.6	40.5	45.3	43.4	40.5	51.0	41.9	28.6	23.3	48.7	0.8	20.2	41.0	0.0	98.2	24.8
With $\boldsymbol{F}$	53.2	54.8	30.5	73.3	49.6	55.9	70.0	56.6	40.9	44.2	56.3	31.7	44.7	44.5	12.6	97.0	23.8
Qwen2-VL-7B-Instruct (Wang et al., 2024)																	
With $\boldsymbol{V}$	16.6	22.6	12.5	28.6	21.8	11.3	2.2	12.5	19.1	8.1	0.6	2.6	21.6	38.3	2.4	87.4	25.0
With $\hat{\boldsymbol{F}}$	49.1	45.9	91.2	45.0	27.2	47.8	**100.0**	7.0	36.4	43.8	99.4	6.1	25.2	76.1	**100.0**	96.8	31.4
With $\boldsymbol{F}$	62.2	70.3	99.4	87.0	43.5	54.5	**100.0**	16.7	46.6	64.2	99.9	38.4	53.9	74.6	**100.0**	97.4	26.4
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)																	
With $\boldsymbol{V}$	19.0	25.7	8.8	33.5	26.3	14.8	0.0	11.9	32.4	12.9	0.0	7.6	31.6	28.1	0.0	59.2	25.0
With $\hat{\boldsymbol{F}}$	22.3	35.4	9.2	48.6	35.2	16.8	1.7	12.0	36.8	12.0	3.3	3.9	29.5	22.0	0.0	41.0	25.0
With $\boldsymbol{F}$	33.6	50.4	12.4	83.3	38.5	22.4	14.2	8.9	44.0	40.8	34.9	39.1	48.5	27.7	1.4	56.6	25.0
InternVL3.5-1B (Wang et al., 2025)																	
With $\boldsymbol{V}$	14.0	17.5	8.8	30.0	10.5	11.2	0.0	10.2	23.3	8.6	0.0	0.7	25.8	25.7	0.0	52.0	25.0
With $\hat{\boldsymbol{F}}$	38.1	35.3	79.8	31.7	19.3	43.2	98.3	2.0	29.2	27.3	55.5	1.4	24.9	34.4	1.8	76.4	25.0
With $\boldsymbol{F}$	57.3	60.2	68.7	87.4	33.2	54.3	99.3	18.7	45.0	64.3	95.9	42.3	54.4	54.2	53.2	84.4	25.0
Kimi-VL-A3B-Instruct (Kimi Team, 2025)																	
With $\boldsymbol{V}$	17.4	23.1	8.8	27.4	25.6	11.1	0.0	1.4	32.0	11.3	0.0	4.1	30.5	41.7	0.0	**100.0**	25.0
With $\hat{\boldsymbol{F}}$	32.3	36.4	43.2	30.9	38.1	31.9	59.1	5.2	31.6	14.9	18.4	4.0	22.6	44.7	5.6	99.6	28.8
With $\boldsymbol{F}$	51.6	56.0	38.8	78.1	44.4	48.6	75.7	15.1	55.0	57.4	78.0	42.1	51.9	44.9	15.4	99.8	19.4
InternVL3.5-8B (Wang et al., 2025)																	
With $\boldsymbol{V}$	16.3	19.3	8.8	25.1	18.9	15.7	0.0	11.9	35.2	14.9	0.0	1.9	43.6	11.2	0.0	8.2	25.4
With $\hat{\boldsymbol{F}}$	40.2	32.0	57.2	30.1	22.9	48.7	98.9	7.4	39.7	31.7	60.4	1.9	32.9	32.5	11.2	72.2	14.0
With $\boldsymbol{F}$	65.4	72.4	88.2	91.1	49.6	60.4	99.9	35.9	45.5	63.0	97.0	32.7	59.1	72.9	95.8	90.6	32.2
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)																	
With $\boldsymbol{V}$	14.9	18.7	8.8	23.3	18.9	12.0	0.0	10.9	25.0	12.1	0.0	17.3	19.3	22.5	0.0	42.4	25.0
With $\hat{\boldsymbol{F}}$	37.2	35.0	60.0	28.2	30.0	43.9	88.0	8.5	35.3	23.4	32.7	14.8	22.7	26.7	0.4	54.4	25.2
With $\boldsymbol{F}$	57.0	66.2	71.1	86.2	47.0	53.4	98.3	16.3	45.8	58.6	88.4	40.6	46.4	42.5	16.8	83.0	27.8
VideoLLaMA3-7B (Zhang et al., 2025)																	
With $\boldsymbol{V}$	23.8	21.0	8.8	22.6	24.9	29.0	42.1	13.6	31.4	16.0	2.2	13.7	32.8	14.9	4.2	15.4	25.2
With $\hat{\boldsymbol{F}}$	34.7	34.6	69.6	27.8	25.4	33.0	48.6	24.2	26.2	39.4	77.2	13.0	27.5	38.6	58.2	30.4	27.2
With $\boldsymbol{F}$	44.0	56.5	79.0	75.5	30.6	33.6	49.1	25.3	26.4	51.9	86.3	28.2	40.8	48.3	74.4	40.0	30.6
Cosmos-Reason2-2B (NVIDIA, 2025)																	
With $\boldsymbol{V}$	15.6	17.7	8.8	25.3	15.0	14.9	0.0	21.1	23.6	14.1	0.0	12.8	30.0	14.7	0.0	19.2	25.0
With $\hat{\boldsymbol{F}}$	20.8	23.2	19.1	28.5	20.3	17.8	10.8	16.0	26.4	24.5	31.2	9.3	33.2	24.5	6.6	41.8	25.0
With $\boldsymbol{F}$	40.2	48.9	44.0	59.4	42.0	34.4	30.9	35.3	37.1	52.3	60.9	40.1	55.9	25.6	9.8	42.0	25.0
Qwen3-VL-8B-Instruct (Bai et al., 2025a)																	
With $\boldsymbol{V}$	17.5	14.5	8.8	21.8	10.7	20.8	0.0	29.4	33.1	12.4	0.0	13.7	23.9	16.0	0.0	22.0	26.0
With $\hat{\boldsymbol{F}}$	44.4	35.9	91.1	28.8	18.3	50.5	99.9	14.4	37.4	39.4	85.5	7.9	24.2	47.0	46.6	69.4	25.0
With $\boldsymbol{F}$	61.9	69.0	92.7	87.2	43.3	56.8	98.5	27.8	44.2	67.7	96.5	47.8	58.4	57.3	68.4	77.0	26.4

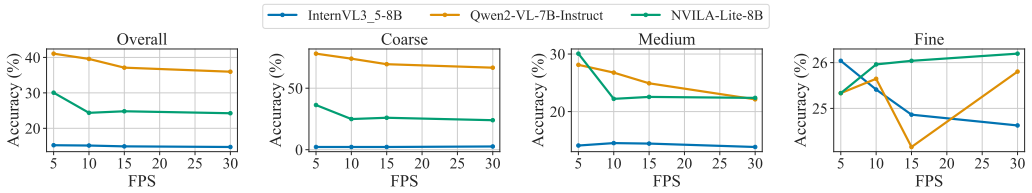


Figure A1: **Effect of frame rate.** We report accuracy across different FPS for InternVL3.5-8B (Wang et al., 2025), Qwen2-VL-7B-Instruct (Wang et al., 2024), and NVILA-Lite-8B (Liu et al., 2025) on different question difficulty levels.

A1.3 Analysis on Frame Rate

Experimental setup. Human motion perception relies on *persistence of vision*: the visual system integrates information across successive frames, so a higher frame rate provides a denser temporal signal that helps in perceiving smooth, coherent motion. If VLMs similarly exploit temporal integration, providing more frames per second (FPS) should improve their performance on appearance-free motion reasoning. We ask: *Does a higher frame rate help VLMs perceive motion?* We test this by evaluating VLMs on AppearFreeVQA at various frame rates.

Results. Fig. A1 reports accuracy across FPS settings on selected VLMs (Wang et al., 2024; 2025; Liu et al., 2025) and question difficulty levels. Contrary to this expectation, increasing the frame rate yields only a marginal difference for VLMs, with some models even showing slight performance drops at higher FPS. Specifically, Qwen2-VL-7B-Instruct drops from 41.1% to 36.0% when FPS increases from 5 to 30, indicating that current models do not meaningfully integrate temporal information across frames and instead treat each frame largely in isolation. **Higher frame rate does not improve VLM performance.** Current VLMs do not utilize the additional temporal signal from higher frame rates as effectively as humans.

A1.4 Camouflage Detection

MoCA (Lamdouar et al., 2020) contains 141 image sequences of naturally camouflaged animals, *e.g.*, arctic foxes blending into snowy backgrounds or leaf insects mimicking foliage. We convert each image sequence to video at 15 fps and split the dataset into training (20%) and test (80%) sets. Animal species are divided into five ecological groups: marine, terrestrial, birds, reptiles, and arthropods. Tab. A3 lists the sub-categories and sequence counts per group.

Table A3: **MoCA group composition.** We categorized MoCA videos into 5 groups based on the animal species.

Group	Species
Marine	Fish (flounder, scorpionfish, flatfish), cephalopods (octopus, cuttlefish), crustaceans (crab, shrimp), other (seal, seahorse, turtle)
Terrestrial	Cats (snow leopard, lion), canids (arctic fox, wolf), bears (polar bear), ungulates (elephant, goat), small mammals (hedgehog, jerboa)
Birds	Owls (screech owl, snowy owl, horned owl), nightjars (potoo, nightjar), grouse
Reptiles	Snakes (horned viper, copperhead), lizards (chameleon, gecko, iguana)
Arthropods	Insects (grasshopper, mantis, stick insect, katydid), arachnids (crab spider)

Question formulation. For each video, we pose the multiple-choice question: "What animal is in the video?" with 5 answer options. The options consist of the correct animal name plus 4 distractors: 2 randomly sampled from the same taxonomic group and 2 from other groups. This 5-way classification task requires the model to both detect the camouflaged animal via motion cues and recognize its

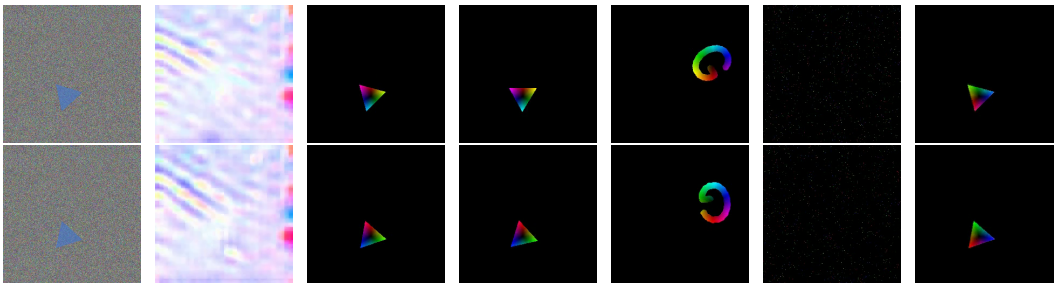


Figure A2: **Oracle-control example frames (Symbol).** We show the sample from Fig. 3(a) under each of the 7 non-baseline conditions in Tab. 2. The Appearance-free panel's blue overlay denotes the single moving object's mask, shown for visualization only and not present in the actual data.

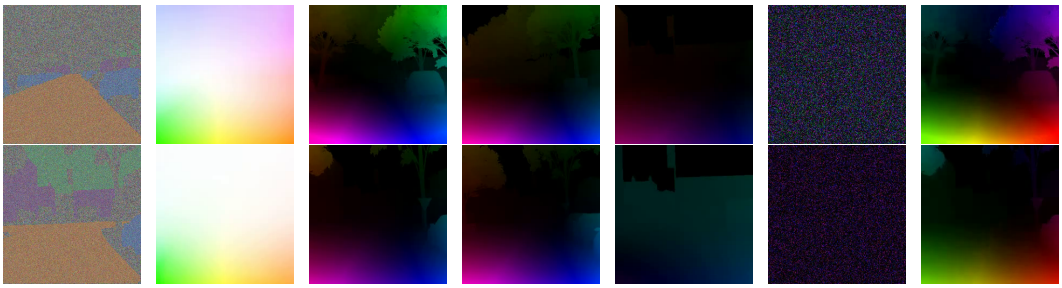


Figure A3: **Oracle-control example frames (Scene).** The first and last frame of the same Scene clip as Fig. 3(g) (car moving straight) under the same 7 conditions as Fig. A2. As in Fig. 3, the Appearance-free panel is overlaid with the 4-class SAM3 semantic mask (road, building, sky, car), shown for visualization only and not present in the actual data.

species. We generate 5 question variants per video with different distractor combinations. All variants share the same video but differ only in the distractor set.

Results. Tab. A4 reports accuracy broken down by the five taxonomic groups: **Arth**ropods, **Bird**s, **Mar**ine, **Rep**tiles, and **Terr**estrial mammals. M+S matches M on average (80.7%), improving on Arth and Mar but regressing on Rep and Terr. SpookyBench provides little benefit to camouflage detection due to its limited scope. M+A improves over M on every group except Birds. The largest average gain is in Terrestrial (+4.2%).

Table A4: **MoCA per-group accuracy (%).** Same training conditions as Tab. 3. Values are mean ± std over 5 distractor sets. Best result per column in **bold**.

Training data	Avg	Arth	Bird	Mar	Rep	Terr
Zero-shot	$76.6_{\pm 1.3}$	$78.2_{\pm 10.4}$	$86.7_{\pm 13.9}$	$77.0_{\pm 6.5}$	$72.7_{\pm 9.1}$	$75.3_{\pm 6.6}$
M	$80.7_{\pm 2.6}$	$80.0_{\pm 11.9}$	$\mathbf{93.3}_{\pm 9.1}$	$80.0_{\pm 7.0}$	$78.2_{\pm 8.1}$	$80.5_{\pm 4.0}$
M+S	$80.7_{\pm 2.3}$	$\mathbf{83.6}_{\pm 7.6}$	$\mathbf{93.3}_{\pm 9.1}$	$81.7_{\pm 7.8}$	$72.7_{\pm 9.1}$	$79.0_{\pm 8.1}$
M+A	$\mathbf{84.3}_{\pm 1.7}$	$81.8_{\pm 11.1}$	$\mathbf{93.3}_{\pm 9.1}$	$\mathbf{83.8}_{\pm 5.9}$	$\mathbf{81.8}_{\pm 6.4}$	$\mathbf{84.7}_{\pm 5.7}$

A1.5 Oracle-control Example Frames

Fig. A2 shows the same frame from one representative Symbol clip under each of the 7 non-baseline conditions in the main-paper leave-one-out table (Tab. 2): appearance-free video, estimated flow, target flow, and the four leave-one-out controls (temporally misaligned, motion misaligned, spatially misaligned, unconventional color). Spatially misaligned breaks the triangle's shape into scattered pixels despite preserving the same per-pixel flow values; motion misaligned instead shows a coherent but unrelated flow field (a different source clip's genuine motion); unconventional color preserves the exact same spatial pattern as target flow under a different color-to-direction mapping. Fig. A3 shows the same 7 conditions on a representative Scene clip, at its first and last frame rather than interior indices, since the relevant motion (ego-camera translation) is clearest between the two endpoints of the clip.

A1.6 Additional Oracle-Control Encodings

To further probe what makes an oracle encoding sufficient (Sec. 4.2), we test five additional input encodings beyond the seven-condition leave-one-out grid in Tab. 2: (1) **Silhouette image**: the first frame with a non-empty binary object mask, repeated for every frame of the clip, so the video carries a static shape cue but no motion. (2) **Silhouette video**: the binary object mask rendered at each frame's own timestep, so shape and temporal alignment are both present but the mask carries no directional flow information. (3) **Moving random mask**: a mask patch unrelated to the source object follows a smooth, source-independent trajectory. (4) **Motion text tokens**: a fixed-length natural-language description of the direction and per-frame-normalized magnitude of the target flow $\boldsymbol{F}$, with no visual input at all. (5) **Motion numerical tokens**: the same target-flow information as a fixed grid of (x, y, dx, dy) tokens instead of natural language.

Tab. A5 reports the results. A static silhouette carries little signal: Silhouette image scores 31.0%, close to random (32.4%) and only slightly above the appearance-free baseline (14.3%), confirming that shape alone, without any temporal variation, is nearly uninformative. Silhouette video jumps to 71.8%, indicating that a purely binary, colorless shape sequence lets the VLM recover most motion perception capabilities. Moving random mask, a shape unrelated to the true object that merely moves smoothly, still reaches 52.8%, well above the appearance-free baseline despite carrying no information about the object's true identity or trajectory; this is consistent with the coarse-level yes-bias documented for several models in Sec. 4.1 and suggests part of the oracle gain reflects detecting *any* moving shape rather than the correct one. Motion text tokens (53.7%) and motion numerical tokens (54.9%) recover comparable accuracy with no visual input at all, showing that Qwen3.5-9B can exploit a structured motion description through language alone, a distinct capability from perceiving motion in pixels.

Table A5: **Additional oracle-control encodings.** We report accuracy (%) of Qwen3.5-9B (Qwen Team, 2026), extending Tab. 2's conditions with five further input encodings on the same model, following the format in Tab. 1.

Condition	All	Symbol				Humanoid				Animal				Scene			
		Avg	C	M	F	Avg	C	M	F	Avg	C	M	F	Avg	C	M	F
Random	32.4	30.8	50.0	26.7	26.3	33.3	50.0	25.0	25.0	33.4	50.0	25.0	25.0	33.3	50.0	25.0	25.0
Human	**93.3**	**95.7**	**100.0**	**93.3**	**95.4**	**95.7**	**100.0**	**98.9**	**88.3**	**78.2**	**100.0**	**60.0**	**74.1**	**94.3**	**100.0**	**96.7**	**87.1**
Appearance-free	14.3	18.0	9.2	26.3	14.6	13.1	2.5	8.3	28.3	12.4	0.0	3.9	34.0	10.9	1.4	6.4	25.0
Silhouette image	31.0	30.0	8.8	56.1	16.7	26.7	0.0	52.3	27.8	13.3	0.0	12.7	27.9	39.3	0.0	92.8	25.0
Silhouette video	71.8	74.5	89.0	91.1	54.2	68.7	81.1	76.9	48.0	44.3	85.9	9.9	36.8	67.9	83.2	95.6	25.0
Moving random mask	52.8	50.9	70.4	63.0	32.3	48.8	69.8	43.8	32.7	38.6	80.3	8.5	26.5	63.5	75.2	90.6	24.8
Motion text tokens	53.7	49.6	93.0	55.7	25.9	50.3	91.8	28.1	31.0	55.2	91.6	2.8	72.1	69.7	**100.0**	73.2	36.0
Motion numerical tokens	54.9	50.2	93.0	54.6	28.1	52.1	91.8	32.4	32.0	54.8	91.6	9.9	63.2	73.5	**100.0**	95.4	25.0

A1.7 Additional Baselines

Motion-energy preprocessing. Following Tangemann et al. (Tangemann et al., 2024), we run their motion-energy model on the appearance-free video and binarize its output at the official logit threshold of 0 (probability 0.5), producing a single motion-segmentation frame that is supplied to the VLM in place of the raw video.

Tab. A6 reports Qwen3.5-9B accuracy on this motion-segmentation input. Motion-energy preprocessing raises accuracy well above both random chance and the raw appearance-free baseline in every category (Symbol: 18.1% → 62.5%; Humanoid: 13.1% → 51.7%; Animal: 12.4% → 47.9%; Scene: 10.9% → 55.0%). This corroborates the flow-oracle result in Sec. 4.2: an explicit, appearance-independent motion representation recovers most of the missing performance, even though motion-energy segmentation does not itself require optical flow estimation. A gap to human performance (78.2%–95.7%) nonetheless remains in every category, so motion-energy preprocessing narrows but does not close the human-VLM gap.

Table A6: **Static noise and motion-energy baselines on AppearFreeVQA.** Qwen3.5-9B accuracy (%) per category. **Bold** marks the best value in each column among the 4 conditions (Random/Human excluded as reference baselines); values in red fall below that column's Random baseline.

Input to Qwen3.5-9B	Symbol	Humanoid	Animal	Scene
Random	30.8	33.3	33.4	33.3
Human	95.7	95.7	78.2	94.3
Appearance-free video	18.1	13.1	12.4	10.9
Motion-energy preprocessing (Tangemann et al., 2024)	**62.5**	**51.7**	**47.9**	**55.0**

CREMA VDFN. We evaluate CREMA's (Yu et al., 2025) released `nextqa_vdfn.pth` checkpoint which is trained to take in additional modalities like optical flow. In Tab. A7, we test three flow conditions: no flow, predicted flow $\hat{\boldsymbol{F}}$, and target flow $\boldsymbol{F}$. Flow input monotonically improves CREMA's accuracy (20.8% → 26.4% → 32.2%), but even ground-truth flow leaves it below the 32.4% random baseline, unlike the flow-oracle gain from directly supplying a general-purpose VLM with a flow visualization (Sec. 4.2), suggesting CREMA's dedicated flow-fusion branch does not translate ground-truth motion into correct answers as effectively as prompting a strong zero-shot VLM.

Table A7: **Flow-aware VLM (CREMA VDFN (Yu et al., 2025)) zero-shot on AppearFreeVQA.** **Bold** marks the best value among the 3 flow conditions (Random/Human excluded as reference baselines); values in red fall below the Random baseline.

Condition	All
Random	32.4
Human	**93.3**
RGB (no flow)	20.8
RGB + predicted flow	26.4
RGB + ground-truth flow	**32.2**

A1.8 Limitations of `ForwardWarp` and Motivation for `MotionOptim`

`ForwardWarp` generates appearance-free frames by splatting each source pixel to its destination according to the accumulated flow $\tilde{\boldsymbol{F}}^{(n)}$. This works well for motions with small displacements, where the foreground region remains largely self-covered across frames. However, for more complex motion-defined forms with large displacements or intricate spatial structure (*e.g.*, running horses in the **Animal** category), the foreground object shifts substantially between frames, leaving large disoccluded regions ($\bar{\boldsymbol{M}}^{(n)}$). `ForwardWarp` fills these regions with fresh background noise $\boldsymbol{B} \sim \mathcal{N}$, but when the disoccluded area is large, a visible boundary forms between the splatted foreground and the freshly sampled background, producing structured artifacts in the output. These artifacts break the noise-like appearance constraint and can inadvertently introduce appearance-based cues.

`MotionOptim` addresses this by casting video generation as a gradient-based optimization (Eq. 1). Starting from the `ForwardWarp` output as initialization, it jointly minimizes $\mathcal{L}_{\text{appear}}$ and $\mathcal{L}_{\text{motion}}$, smoothly reconciling the boundary between filled and disoccluded regions while preserving the underlying motion-defined form. This refinement step is most beneficial for the **Animal** category, where large articulated motions produce the most severe disocclusion.

A1.9 GENERATOR ABLATION

We compare `ForwardWarp` vs. `MotionOptim` with two measures (Tab. A8).

Object-mask AUROC. We train a regularized linear probe to predict, for each pixel, whether it lies inside the ground-truth object mask using only that pixel's appearance in a single frame. Evaluation uses video-grouped 5-fold cross-validation, so frames from the same video never appear in both the train and test folds; we report the resulting AUROC, where 0.5 indicates the object's location cannot be recovered from single-frame appearance.

Local-variance mismatch. We compute the local pixel variance in a 15-pixel window and take the absolute log-ratio between the mean local variance inside vs. outside the object mask; 0 indicates the two regions are statistically indistinguishable by local texture alone.

Tab. A8 shows that `MotionOptim` reduces appearance leakage relative to `ForwardWarp` on both measures. Object-mask AUROC drops from 0.6083 to 0.5295, closer to the uninformative 0.5, indicating that the frame-level artifacts `ForwardWarp` leaves at large-displacement disocclusions (Sec. A1.8) make the object's location partially recoverable from a single frame, and that `MotionOptim`'s joint optimization of $\mathcal{L}_{\text{appear}}$ and $\mathcal{L}_{\text{motion}}$ suppresses this. Local-variance mismatch falls from 0.0403 to 0.0263, corroborating the AUROC result with an independent, non-learned statistic.

Table A8: **Generator ablation:** `ForwardWarp` **vs.** `MotionOptim`**.**

	Object-mask AUROC	Local-variance mismatch
ForwardWarp	0.6083	0.0403
MotionOptim	**0.5295**	**0.0263**

A1.10 FOREGROUND/BACKGROUND APPEARANCE CONTROL

We extend the single global appearance-level parameter γ (Eq. A6, Sec. A3.2) to two independent strengths, γ_{fg} for pixels inside the moving object's mask $\boldsymbol{M}_{\text{obj}}^{(n)}$ (distinct from `ForwardWarp`'s correspondence mask $\boldsymbol{M}^{(n)}$, Sec. 3.2) and γ_{bg} for pixels outside it:

$$\mathcal{L}'_{\text{appear}}(\boldsymbol{X}) = \left\|\boldsymbol{X} - \boldsymbol{X}^{(n)}\right\|^2 + \gamma_{\text{fg}} \left\|\boldsymbol{X} - \boldsymbol{I}^{(n)}\right\|^2 \odot \boldsymbol{M}_{\text{obj}}^{(n)} + \gamma_{\text{bg}} \left\|\boldsymbol{X} - \boldsymbol{I}^{(n)}\right\|^2 \odot \left(1 - \boldsymbol{M}_{\text{obj}}^{(n)}\right). \tag{A4}$$

Table A9: **Foreground/background appearance-level grid.** Accuracy (%) as a function of independent foreground (γ_{fg}, rows) and background (γ_{bg}, columns) appearance strength.

$\gamma_{\text{fg}} \backslash \gamma_{\text{bg}}$	γ_{bg} 0.0	0.1	0.2	1.0
0.0	14.2	16.9	17.7	57.5
0.1	17.1	17.2	29.1	63.2
0.2	24.8	30.0	34.9	66.6
1.0	62.3	63.3	67.5	**79.6**

Holding $\gamma_{\text{bg}} = 0$ fixed, increasing foreground appearance from $\gamma_{\text{fg}} = 0$ to 1.0 raises accuracy from 14.2% to 62.3%; holding $\gamma_{\text{fg}} = 0$ fixed, increasing background appearance from $\gamma_{\text{bg}} = 0$ to 1.0 raises it from 14.2% to 57.5%. Both regions therefore contribute appearance cues that VLMs exploit, with a somewhat larger effect for foreground.

A1.11 Replay Sensitivity

Table A10: **Replay-sensitivity analysis.** GPT-5.1 accuracy (%) when shown the same video 1×, 3×, and 5×. Random/Human are the pooled (Symbol+Humanoid+Animal+Scene) reference values from Tab. A1 – flat across exposure count since neither was re-measured under replay.

Condition	Accuracy (%)
Random	32.4
Human	**93.3**
GPT-5.1 (1×)	41.5
GPT-5.1 (3×)	43.9
GPT-5.1 (5×)	42.8

Tab. A10 reports GPT-5.1 accuracy when the same video is shown 1×, 3×, and 5× before the model answers. Accuracy stays within a narrow 41.5–43.9% band across all three exposure counts, far below the 93.3% human baseline.

A1.12 VLM Evaluation Protocol

Video duration. Appearance-free videos $\boldsymbol{V}$ are rendered at 30 fps for Symbol/Humanoid/Animal and 15 fps for Scene (up to 75 frames per clip, Sec. A2). Across the frozen human-survey assignments, clip duration has a response-weighted median of 1.00 s, mean of 2.36 s, and range of 0.27–35.93 s. Regardless of native fps, we sample $N = 16$ frames uniformly spaced across the clip when querying a VLM by default, except for the frame-rate ablation (Sec. A1.3), which instead samples at an explicit target FPS.

Prompt template. Every question uses a fixed template: "Watch the video carefully and answer the following multiple-choice question. Question: {question}. Options: {options}. Reply with ONLY the option letter in parentheses, e.g. (A). Do not include any explanation or additional text.", with options listed as `(A) ...`, `(B) ...`, etc.

Answer extraction. We first try to match a leading option letter (*e.g.* "A", "(A)", "(A) answer"); failing that, we look for a leading option number, then for an exact or substring match against the option text; if none match, we fall back to a uniform random guess among the options. This fallback is rare in practice and only affects severely off-format responses.

Replay protocol. For the human study (Sec. A2.7), annotators could replay each video as many times as needed before answering; Sec. A1.11 reports a matched replay-sensitivity control for GPT-5.1.

Dataset source counts and category distribution. AppearFreeVQA contains approximately 5,100 videos and 17,600 questions across four categories: Symbol (1,000 videos / 5,334 questions, synthetic shapes/digits/characters with synthetic rigid motion), Humanoid (2,867 / 8,601, synthetic point-light figures driven by real MPOSE mocap), Animal (724 / 2,147, real TigDog video and motion), and Scene (500 / 1,500, real Virtual KITTI 2 driving video), plus a Noise control (100 videos, coarse-only). Full per-category, per-difficulty-level breakdowns are in Tab. A19.

Table A11: **Single- vs. multi-form Symbol accuracy (%), matched subset. Bold** marks the best value in each column (Random excluded as the reference baseline); values in red fall below random performance.

	Single-form	Multi-form
Random	30.8	30.8
Human	**95.3**	**96.8**
GPT-5.1	39.5	38.7

GPT-5.1 accuracy is similarly low on single-form (39.5%) and multi-form (38.7%) Symbol videos, while human accuracy stays high in both (95.3% and 96.8%, respectively); the 0.8-point gap is not

significant (95% video-clustered CI $[-20.6, +19.3]$ pp), so this shows the motion-understanding bottleneck persists with multiple forms rather than that multi-form videos are strictly harder.

A2 APPEARFREEVQA DETAILS

In this section, we provide details of AppearFreeVQA design across all four motion categories: Symbol, Humanoid, Animal , and Scene .

A2.1 SYMBOL CATEGORY

Setup. The Symbol category uses fully synthetic videos rendered on a 256×256 canvas. The symbol vocabulary contains ≈ 40 entries: 6 geometric shapes (circle, square, triangle, star, diamond, cross), 24 Latin letters ("A"–"Z", excluding "I" and "O" to avoid confusion with numerals), and 9 digits (1–9). Each video places 1 or 2 symbols sampled uniformly at random from this vocabulary.

Three motion types are rendered, each with a distinct optical flow signature.

1. **Translation**: each symbol moves linearly with a random 2-D velocity and reverses the corresponding velocity component upon wall collision; ground-truth flow equals the symbol's instantaneous velocity within its mask and zero in the background.
2. **Rotation**: each symbol rotates about its centroid at angular velocity $\omega \in [-\pi, \pi]$ rad/s; for a pixel at offset (dx, dy) from the centroid, the flow is $\omega(-dy,\ dx)$.
3. **Scaling**: each symbol expands or contracts uniformly, producing radially outward or inward flow from the centroid.

Distribution. Fine-level distractors are sampled from motion types outside the correct answer, plus options from the Humanoid and Animal categories. The full distribution is shown in Tab. A12.

Table A12: **Symbol category distribution.** C = Coarse, M = Medium, F = Fine.

Motion	# Videos	# Questions			
		C	M	F	Total
Translation	500	500	1,000	1,500	3,000
Rotation	500	500	1,000	1,000	2,500
Scaling	200	200	400	400	1,000
Total	**1,200**	**1,200**	**2,400**	**2,900**	**6,500**

A2.2 HUMANOID CATEGORY

Setup. The Humanoid category is built from MPOSE (Mazzia et al., 2022), which provides keypoint sequences for 20 human actions (Tab. A14). We render only the skeleton—discarding clothing, skin tone, and background—to remove all texture cues. Keypoints with confidence below 0.1 are discarded; the remainder are normalized to $[0.1, 0.9]$ while preserving the aspect ratio so the full skeleton fits within the frame. For RGB videos, 24 limb segments are drawn as green lines with red keypoint circles on a 256×256 black canvas at 30 fps.

The 20 actions are organized into five semantic groups (Tab. A13). Fine-level distractors are sampled from groups outside the correct answer, plus options from the Symbol and Animal categories.

Table A13: **Humanoid motion groups.** We group the 20 motion types of the Humanoid category by semantic category.

Group	Motions
Locomotion	walk, jog, run
Jumping	jump, pjump
Waving	wave1, wave2
Postural transitions	standing, sit-down, get-up
Upper body	check-watch, cross-arms, scratch-head, box, kick, point, pick-up, bend, hands-clap, turn

Each internal label is mapped to a human-readable display text in answer options. The full mapping is listed in Tab. A14.

Table A14: **Humanoid motion label mapping.**

Label	Display text	Label	Display text
standing	Standing	wave1	Waving
check-watch	Checking their watch	box	Boxing
cross-arms	Crossing their arms	kick	Kicking
scratch-head	Scratching their head	point	Pointing
sit-down	Sitting down	pick-up	Picking something up
get-up	Standing up	bend	Bending over
turn	Turning around	hands-clap	Clapping hands
walk	Walking	wave2	Waving
jog	Jogging	jump	Jumping
run	Running	pjump	Jumping

Distribution. The video and question distribution of Humanoid is shown in Tab. A15.

Table A15: **Humanoid category distribution.** Per-motion video and question counts of Humanoid, organized by semantic group (C = Coarse, M = Medium, F = Fine).

Motion	# Videos	# Questions			
		C	M	F	Total
Locomotion					
Walk	302	302	302	302	906
Jog	168	168	168	168	504
Run	162	162	162	162	486
Jumping					
Jump	73	73	73	73	219
Place-jump	78	78	78	78	234
Waving					
Wave (1)	252	252	252	252	756
Wave (2)	405	405	405	405	1,215
Postural transitions					
Stand	143	143	143	143	429
Sit down	69	69	69	69	207
Get up	65	65	65	65	195
Upper body					
Check watch	51	51	51	51	153
Cross arms	70	70	70	70	210
Scratch head	52	52	52	52	156
Box	388	388	388	388	1,164
Kick	45	45	45	45	135
Point	53	53	53	53	159
Pick up	51	51	51	51	153
Bend	45	45	45	45	135
Clap	304	304	304	304	912
Turn	91	91	91	91	273
Total	**2,867**	**2,867**	**2,867**	**2,867**	**8,601**

A2.3 Animal Category

The Animal category is derived from TigDog (Del Pero et al., 2015; 2016), which provides RGB clips of horses, tigers, and dogs with per-frame motion labels and segmentation masks. We refine masks with SAM3 (Carion et al., 2026) and estimate $[\boldsymbol{F}^{(n)}]$ using FlowSeek (Poggi & Tosi, 2025) within the masked foreground, extracting animal motion from background clutter.

The **horse** subset: locomotion gaits (walking, trotting, galloping), special gaits (walking backwards, jumping, turning), head movements (turning head, tilting head), postures (sitting, standing up, grazing), leg movements (lifting leg, piaffe), and activities (pulling, hurdles, rodeo, rolling). The **tiger** subset: locomotion (walking, walking backwards, running, swimming), head movements (turning head, tilting head, blinking), mouth movements (opening/closing mouth), body movements (turning, jumping, rolling, sliding leg), postures (sitting, standing up), paw movements (lifting paw, scratching, climbing), and feeding (drinking, eating, licking). The **dog** subset: locomotion (walking, running, jumping), head movements (turning head, tilting head), postures (standing up), and activities (pushing skateboard).

Motion types are organized into semantic groups per animal (Tab. A16). Fine-level distractors are sampled from groups outside the correct answer, plus options from the Symbol and Humanoid categories.

Table A16: **Animal motion groups.** Motion taxonomy for horse (15 motion types), tiger (17 motion types), and dog (7 motion types) from TigDog.

Animal	Group	Motions
Horse	Locomotion	walking, trotting, galloping, walking backwards, jumping, turning
	Head movements	turning head, tilting head
	Postures	sitting, standing up, grazing
	Leg actions	lifting leg, piaffe
	Others	pulling, hurdles, rodeo, rolling
Tiger	Locomotion	walking, walking backwards, running, swimming
	Head movements	turning head, tilting head, blinking
	Mouth movements	opening mouth, closing mouth
	Body movements	turning, jumping, rolling, sliding leg
	Postures	sitting, standing up
	Paw actions	lifting paw, scratching, climbing
	Others	drinking, eating, licking
Dog	Locomotion	walking, running, jumping
	Head movements	turning head, tilting head
	Postures	standing up
	Others	pushing skateboard

Each internal label is mapped to a human-readable display text in answer options. The full mapping is listed in Tab. A17.

Table A17: **Animal motion label mapping.**

Label	Display text	Label	Display text
walking	Walking	lifting_leg	Lifting a leg
trotting	Trotting	piaffe	Performing piaffe
galloping	Galloping	pulling	Pulling
running	Running	hurdles	Jumping over hurdles
swimming	Swimming	rodeo	Performing rodeo
walking_backwards	Walking backwards	rolling	Rolling on the ground
jumping	Jumping	lifting_paw	Lifting a paw
turning	Turning around	scratching	Scratching
turning_head	Turning its head	climbing	Climbing
tilting_head	Tilting its head	sliding_leg	Sliding a leg
blinking	Blinking	drinking	Drinking
opening_mouth	Opening its mouth	eating	Eating
closing_mouth	Closing its mouth	licking	Licking
sitting	Sitting	wagging_tail	Wagging its tail
standing_up	Standing up	pushing_skateboard	Pushing a skateboard
grazing	Grazing		

Distribution. The video and question distribution of Animal is shown in Tab. A18.

Table A18: **Animal category motion distribution.** Per-motion video and question counts aggregated across all three species (horse, tiger, dog). Each labeled video contributes one question per difficulty level (C = Coarse, M = Medium, F = Fine). Videos with unknown motion labels contribute only coarse and medium questions (F = 0).

Motion	# Videos	# Questions			
		C	M	F	Total
Blinking	12	12	12	12	36
Climbing	1	1	1	1	3
Closing mouth	2	2	2	2	6
Drinking	7	7	7	7	21
Eating	4	4	4	4	12
Galloping	27	27	27	27	81
Grazing	1	1	1	1	3
Hurdles	4	4	4	4	12
Jumping	18	18	18	18	54
Licking	8	8	8	8	24
Opening mouth	18	18	18	18	54
Piaffe	3	3	3	3	9
Pulling	7	7	7	7	21
Push skateboard	21	21	21	21	63
Rodeo	4	4	4	4	12
Rolling	5	5	5	5	15
Running	40	40	40	40	120
Scratching	1	1	1	1	3
Sitting	10	10	10	10	30
Slide leg	1	1	1	1	3
Stand up	9	9	9	9	27
Swimming	3	3	3	3	9
Tilt head	34	34	34	34	102
Trotting	17	17	17	17	51
Turn head	88	88	88	88	264
Turning	6	6	6	6	18
Wag tail	4	4	4	4	12
Walking	327	327	327	327	981
Walk backwards	5	5	5	5	15
Others	23	23	23	0	46
Total	**710**	**710**	**710**	**687**	**2,107**

A2.4 Scene Category

The Scene category is derived from Virtual KITTI 2 (Cabon et al., 2020), a synthetic re-rendering of real KITTI driving sequences that ships dense, per-pixel ground-truth optical flow. We use 5 scenes (Scene01, Scene02, Scene06, Scene18, Scene20). The remaining weather/lighting variants share the exact same camera trajectory and so contribute no additional motion content. Clips are cut to at most 5 seconds (75 frames at 15 fps) and labeled *straight*, *left*, or *right* from the real per-frame camera yaw rate, with trimming around the turn onset and a low-motion filter on mean camera speed. We additionally mirror clips left-to-right and reverse their frame order to balance left/right turns and to synthesize *reverse* segments from forward-driving footage, since $[\boldsymbol{F}^{(n)}]$ is the dataset's own ground-truth flow rather than an estimate. Unlike Symbol and Humanoid, Scene has no separate object undergoing motion, *i.e.*, the entire frame moves under ego-motion. Hence, it tests scene-level rather than subject-level motion perception.

A2.5 Question Design

Coarse level (motion detection). The coarse level asks *whether* any motion is present. In natural videos, static appearance provides strong cues: a car on a street or an athlete mid-stride implies

motion from a single frame. Appearance-free videos remove these cues entirely, forcing reliance on inter-frame changes alone.

Medium level (object recognition). The medium level asks *what* is moving. Without visible boundaries or texture, models must identify the moving entity from motion-defined form alone—analogous to spotting a camouflaged animal that reveals itself only through movement.

Fine level (motion characteristics). The fine level asks *how* objects move, the most demanding task. Models must track motion-defined forms over time and extract precise attributes: bounce counts, rotation direction, or action type.

Symbol questions. The **Symbol** category defines one question bank per motion type (5–6 questions each).

Rotation (5 questions):

1. *Coarse — "Is there any motion in the video?"*
2. *Medium — "What is(are) moving in the video?"*
3. *Medium — "How many objects are moving in the video?"*
4. *Fine — "What type of motion is occurring in the video?"*
5. *Fine — "In which direction is {shape} rotating?"*

Scaling (5 questions):

1. *Coarse — "Is there any scaling in the video?"*
2. *Medium — "What is(are) scaling in the video?"*
3. *Medium — "How many objects are scaling in the video?"*
4. *Fine — "What type of motion is occurring in the video?"*
5. *Fine — "Is {shape} moving toward or away from the camera?"*

Translation (6 questions):

1. *Coarse — "Is there any motion in the video?"*
2. *Medium — "What is(are) moving in the video?"*
3. *Medium — "How many objects are moving in the video?"*
4. *Fine — "What type of motion is occurring in the video?"*
5. *Fine — "How many times does {shape} bounce off the walls in total?"*
6. *Fine — "In which direction is {shape} moving at first?"*

Humanoid questions. The **Humanoid** category uses a uniform three-question set per video:

1. *Coarse — "Is there any motion in the video?"*
2. *Medium — "What is(are) moving in the video?"*
3. *Fine — "How is the person moving in the video?"*

Animal questions. The **Animal** category uses a uniform three-question set per video:

1. *Coarse — "Is there any motion in the video?"*
2. *Medium — "What is(are) moving in the video?"*
3. *Fine — "What is the {animal} doing in the video?"*

Scene questions. The **Scene** category uses a uniform three-question set per video:

1. *Coarse — "Is there any motion in the video?"*
2. *Medium — "What is(are) moving in the video?"*, correct answer "car camera (POV)" rather than a discrete object noun, since the entire frame moves under ego-motion.
3. *Fine — "How is the car moving?"*, with options drawn from turning left, turning right, going straight, and going in reverse.

A2.6 Overall Statistics

Tab. A19 summarizes the video and question counts across all four categories and difficulty levels.

Table A19: **AppearFreeVQA statistics.** Question counts broken down by category and difficulty level.

Category	Videos	# Questions			
		Coarse	Medium	Fine	Total
Symbol	1,000	1,000	2,000	2,334	5,334
Humanoid	2,867	2,867	2,867	2,867	8,601
Animal	724	724	724	699	2,147
Scene	500	500	500	500	1,500
Total	**5,091**	**5,091**	**6,091**	**6,400**	**17,582**

A2.7 Human Performance Evaluation

Procedure. We collected responses from 12 human annotators. Each question was presented as a self-paced video player embedded in a web interface. Each question page showed (1) the appearance-free video, (2) the multiple-choice question text, and (3) a number of labeled answer buttons. Annotators were allowed to replay the video as many times as needed before answering.

Annotators were shown the following instructions:

> *"1. Watch the appearance-free video carefully. 2. Read the question and pick the best matching option. 3. If you truly cannot determine the answer, select 'I cannot tell'. 4. Navigate with Prev/Next. Correct answers are revealed after you pick."*

No additional hints about optical flow encoding or motion types were given. Screenshots of the survey interface are shown in Fig. A4.

No time limit was imposed per question. Tab. A20 reports mean, standard deviation, and median response time broken down by category and difficulty level.

Table A20: **Human response time (seconds) by category and difficulty level.** Mean, standard deviation, and median are reported. Outlier responses were removed before computing statistics.

	Average			Coarse			Medium			Fine		
Category	**Mean**	**Std**	**Med**	**Mean**	**Std**	**Med**	**Mean**	**Std**	**Med**	**Mean**	**Std**	**Med**
Symbol	5.80	3.61	4.28	3.13	1.22	2.80	6.15	3.56	6.72	6.70	4.26	6.03
Humanoid	4.27	2.71	3.35	3.13	1.23	2.93	4.43	2.96	3.33	6.19	4.56	4.27
Animal	8.11	4.62	6.84	4.73	2.10	4.27	8.39	2.96	8.00	10.27	5.36	10.81
Scene	5.80	3.21	4.90	4.37	1.97	3.61	6.45	3.82	5.59	7.49	4.48	6.24
Overall	5.52	3.64	4.14	3.51	1.44	3.22	5.77	3.66	4.31	7.23	4.98	5.46

Analysis. The overall median response time was 3.82 seconds. We observe a consistent pattern across both category and difficulty. By *difficulty*, coarse-level questions are answered fastest (median 2.95 s), as detecting the presence of any motion requires only brief observation. Medium-level questions take moderately longer (median 4.04 s), while fine-level questions require substantially more time (median 4.94 s). By *category*, Animal questions take the longest overall (median 7.21 s), especially at the Fine level (median 10.98 s), reflecting the complexity of interpreting biological motion. Humanoid questions are fastest overall (median 3.25 s), likely because skeletal motion is the most familiar form of real-world motion. These trends confirm that question difficulty is well-calibrated: harder levels impose a greater perceptual burden on human annotators.

A3 Additional Method Details

A3.1 Implementation Details

The noise distribution is Gaussian, $\mathcal{N} = \mathcal{N}(0, 1)$, which admits smooth gradient-based optimization.

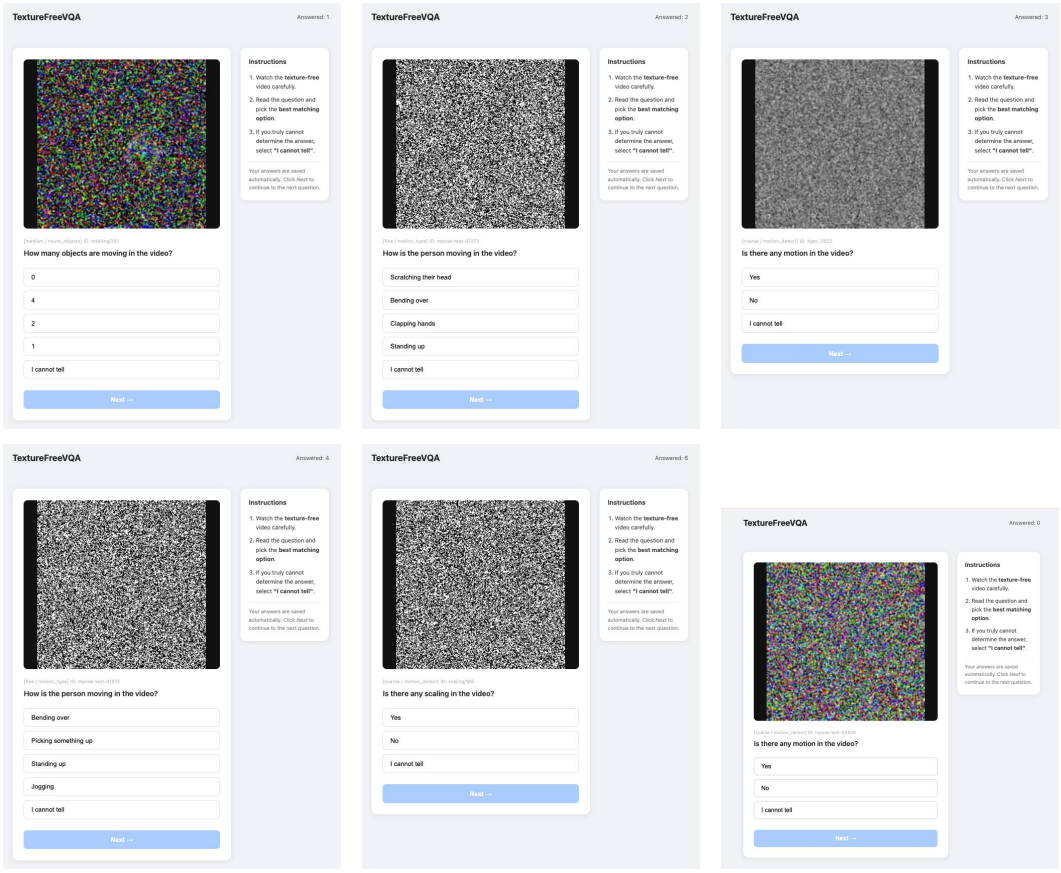


Figure A4: **Human evaluation interface.**

We optimize each frame $\boldsymbol{X}^{(n)}$ with the AdamW optimizer at learning rate $\eta = 0.1$ for $T = 200$ gradient steps, initializing from the `ForwardWarp` output. We set $\lambda = 0.9$ for all categories.

Appearance-free loss. At each gradient step, we sample $K_B = 512$ independent batches of $K = 32$ pixels, fit the empirical mean $\hat{\mu}_{\mathcal{S}'}$ and variance $\hat{\sigma}^2_{\mathcal{S}'}$ to each batch, and evaluate the closed-form KL divergence against $\mathcal{N}(0, 1)$:

$$\mathrm{KL}\big(\mathcal{N}(\hat{\mu}_{\mathcal{S}'}, \hat{\sigma}^2_{\mathcal{S}'}) \,\big\|\, \mathcal{N}(0,1)\big) = \frac{1}{2}\big(\hat{\sigma}^2_{\mathcal{S}'} + \hat{\mu}^2_{\mathcal{S}'} - 1 - \log \hat{\sigma}^2_{\mathcal{S}'}\big)\,. \tag{A5}$$

Averaging over the K_B batches yields $\mathcal{L}_{\text{appear}}$ per Eq. 2.

A3.2 Appearance Level Optimization

To control the degree of appearance information in the videos, we introduce a parameter $\gamma \geq 0$ that interpolates between fully appearance-free ($\gamma = 0$) and natural RGB ($\gamma \to \infty$). Let $\boldsymbol{X}^{(n)}$ denote the noise-warped frame and $\boldsymbol{I}^{(n)}$ the corresponding original RGB frame. The semi-appearance-free frame $\hat{\boldsymbol{X}}^{(n)}$ is obtained by optimizing,

$$\mathcal{L}'_{\text{appear}}(\boldsymbol{X}) = \left\|\boldsymbol{X} - \boldsymbol{X}^{(n)}\right\|^2 + \gamma \left\|\boldsymbol{X} - \boldsymbol{I}^{(n)}\right\|^2. \tag{A6}$$

At $\gamma = 0$ the frame is identical to the appearance-free original; as $\gamma \to \infty$ it converges to the RGB frame $\boldsymbol{I}^{(n)}$. We solve it via gradient descent (100 steps, $\eta = 0.05$) to maintain implementation consistency with `MotionOptim` and allow flexible control over the step budget.

A3.3 Optical Flow Visualization and Prompting

For the oracle experiments (Sec. 4.2), each optical flow field is rendered as an RGB frame using the standard Middlebury color wheel (Baker et al., 2011), which encodes flow direction as hue and magnitude as saturation. We additionally prompt the VLM with an explicit description of this encoding, informing the model that the input video is an optical flow visualization where colors represent motion direction and intensity.

A3.4 Computational Resources

Data generation. Synthetic **Symbol** videos are rendered on CPU and require no GPU acceleration. Target flow $\boldsymbol{F}$ for Symbol and Humanoid needs no estimator: it comes directly from the known warping/joint parameters used to generate the video. For Animal, where TigDog has no dense ground-truth flow, $\boldsymbol{F}$ is instead computed with FlowSeek (Poggi & Tosi, 2025) on the original real RGB video. **Scene** needs no estimator either: Virtual KITTI 2 (Cabon et al., 2020) ships its own dense per-pixel ground-truth flow, used directly as $\boldsymbol{F}$. We use FlowSeek (Poggi & Tosi, 2025) for all optical flow prediction, run on a single A100 GPU.

Model evaluation. All open-source VLMs are evaluated on a single node with 4 A100 GPUs. All proprietary models are called through their respective APIs.

Fine-tuning. Fine-tuning experiments are trained with DeepSpeed ZeRO-3 on 1 node with 4 A100 GPUs for 2 epochs. We set the global batch size to 1024 and gradient accumulation over 2 steps.